Sentinel: Embodied Cooperative Spatial Reasoning and Planning

Xiangye Lin[1⋆] Hongxin Zhang[1⋆], Ruxi Deng[2], Qinhong Zhou[1], Chuang Gan[1]
[1] University of Massachusetts Amherst
[2] University of Illinois Urbana-Champaign
{xiangyelin,hongxinzhang,qinhongzhou,chuangg}@umass.edu

Abstract. In this work, we study *Cooperative Spatial Intelligence*, the ability of decentralized embodied agents to coordinate effectively under dynamic environmental constraints across city-scale outdoor domains. We introduce *Sentinel Challenge*, a benchmark where multiple decentralized embodied agents must communicate in natural language to agree on a mutually safe and convenient meeting point within large, city-scale outdoor environments. Each agent must then navigate safely while avoiding dynamic sentinels patrolling the area, with a tool providing coarse spatial information. To address this, we propose *CoSaR* (Cooperative Spatial Reasoning and Planning), a framework that bridges the high-level communication and planning abilities of foundation models with the precision of classical spatial navigation algorithms. *CoSaR* enables agents to exchange situational updates, reason over evolving spatial constraints, and collaboratively replan trajectories. Evaluated across 14 city-level scenes with 3–5 agents, *CoSaR* consistently leads to faster gathering, shorter path lengths, and improved safety. Our results demonstrate that integrating dynamic communication with spatial reasoning is essential for robust multi-agent cooperation. By formalizing this new setting and providing a scalable benchmark, we aim to build a foundation for advancing cooperative spatial intelligence in embodied multi-agent systems. Code and challenge are available at `https://github.com/UMass-Embodied-AGI/Sentinel`.

Keywords: Embodied AI · Multi-Agent · Spatial Planning

1 Introduction

Humans excel at coordinating in complex spatial environments, determining safe and convenient meeting points, sharing updates about obstacles, and adapting plans as situations evolve. We refer to this capability as *Cooperative Spatial Intelligence*. Despite rapid progress in embodied AI, whether multi-agent systems can exhibit such competence in large, dynamic environments remains largely unexplored. In this work, we study embodied multi-agent cooperation under dynamic spatial constraints in large-scale outdoor environments. As shown in Fig. 1, multiple decentralized embodied agents must communicate in natural

⋆ denotes equal contribution.

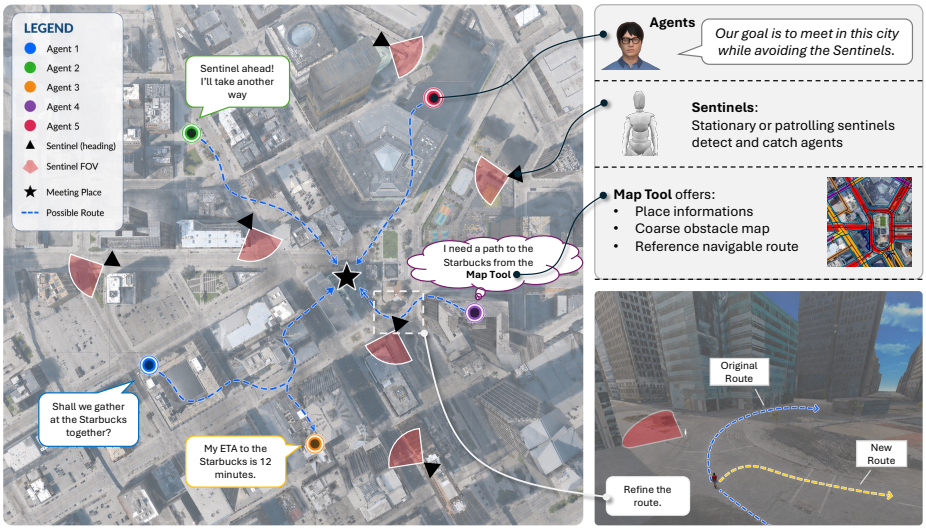

Fig. 1: Embodied agents need to have **Cooperative Spatial Intelligence** to cooperate efficiently under dynamic spatial constraints in large spatial extents.

language to agree on a mutually safe and convenient gathering point in large, city-scale outdoor scenes. After reaching consensus, each agent must navigate safely while avoiding sentinels patrolling the environment, relying on a map tool that provides only coarse point-to-point waypoints, an abstraction inspired by how humans coordinate while using a map to meet in a city.

Recent advances in foundation models [12, 33] have significantly strengthened the capabilities of embodied multi-agent systems. Prior work, such as *CoELA* [65] demonstrates that agents equipped with large language models can communicate and cooperatively solve indoor rearrangement tasks, and *CaPo* [24] further improves cooperation efficiency through meta-plan generation. However, these systems operate only in confined, mostly static indoor environments where simple obstacle-avoidance navigation is sufficient. Scaling cooperation to realistic outdoor settings introduces new challenges: vastly larger spatial extents, path uncertainties, dynamically changing constraints, and the need for continual multi-agent information exchange.

To address these challenges, we introduce **CoSaR**, a cooperative spatial reasoning and planning framework that combines the communication and planning strengths of foundation models with classical spatial navigation algorithms. CoSaR maintains a structured spatial memory consisting of a low-level occupancy map and high-level representations of each agent and patrolling sentinels. This memory is continually updated using visual observations and natural-language communications, enabling agents to infer missing information, query a map tool or peers to gather additional context, refine coarse waypoints, and collaboratively replan routes in real time as conditions evolve.

We instantiate this problem setting within Virtual Community [70] and develop the **Sentinel** Challenge featuring configurable difficulty based on sentinel

numbers and moving patterns. Experiments across 14 large-scale city environments with 3–5 agents demonstrate that CoSaR achieves consistently faster gathering, shorter path lengths, and substantially higher safety compared to strong baselines such as CoELA [65] and RoCo [29]. These improvements stem from CoSaR's ability to integrate dynamic communication, structured spatial reasoning, and continual adaptation to changing environmental constraints. By formalizing this new problem space and releasing a scalable benchmark, we aim to advance the study of cooperative spatial intelligence in embodied multi-agent systems through the Sentinel Challenge and CoSaR methods. Our contributions:

- We introduce the **Sentinel** Challenge to evaluate the embodied agent's *Cooperative Spatial Intelligence*, the capability to cooperate effectively under dynamic spatial constraints.
- We propose **CoSaR**, an embodied cooperative spatial reasoning and planning method featuring a dynamic spatial memory that enables agents to share information, reason about spatial dynamics, and replan collaboratively.
- Experiments across 14 city-scale scenes with 3–5 agents show that our method effectively leverages spatial information and dynamic communication, outperforming strong baselines across multiple metrics.

2 Related Work

2.1 Multi-Agent Cooperation

Multi-agent cooperation is a long-standing and fundamental research problem in artificial intelligence [2,4,5,8,15,16,27,40,49,50,52,58]. For embodied intelligence, existing works primarily focus on multi-agent cooperation in household environments [36,37], while recent studies further extend this setting to communicative scenarios [24, 65, 67, 72]. Different from the existing works, in this paper, we focus on the multi-agent cooperation problem in dynamic city-level environments, where agents are often spatially distant and rarely encounter each other directly, requiring effective communication strategies, complex spatial reasoning, and robust inference over continuously changing environments.

2.2 Foundation Models for Embodied AI

Foundation models have shown significant success in building autonomous agents in both general and embodied settings [21, 23, 30, 34, 37, 41, 51, 54, 60–62, 69]. Specifically, large language models are widely used for planning and decision-making with text input [1,35,45,53,56], while vision-language models are adopted to facilitate end-to-end planning from visual perceptions [17, 20, 31, 57, 66, 71]. In this paper, we use foundation models as the backbones of our agent framework for perception, communication, and decision-making.

2.3 Embodied Navigation and Spatial Intelligence

Embodied navigation [3, 11, 14, 19, 28, 42, 43, 46, 59, 68] focuses on how agents perceive, reason, and act within 3D environments to reach goals given visual and linguistic inputs. The problem studied in this paper is closely related to the subfield of embodied urban navigation [18, 22, 26, 32, 44, 48], which extends

navigation into city-scale environments, emphasizing long-horizon planning, multimodal perception, and real-world constraints. A key capability for tackling such problems is spatial reasoning [7, 10, 13, 55, 63, 64], which enables agents to infer geometric structures, spatial relations, and affordances from limited observations to support effective decision-making in complex environments. In this paper, we introduce a novel spatial reasoning framework to address these challenges.

3 Problem Statement

We study the problem of embodied multi-agent cooperation under dynamic spatial constraints. A group of decentralized embodied agents is randomly initialized across a city-scale outdoor environment. Their objective is to communicate in natural language and navigate safely to gather at a **Place** as soon as possible. The environment contains dynamically changing hazards—modeled as **Sentinels**—that can detect and "catch" agents within their ego-centric field of view. The task succeeds *only if all agents safely gather at a place* without being caught. To support navigation across vast spatial scales, agents may query a **Map Tool** that provides only coarse, high-level spatial information (e.g., waypoint routes), mirroring how humans coordinate using a city map when meeting in an unfamiliar area. More formally, our problem can be formalized as a decentralized partially observable Markov decision process with communication (DEC-POMDP-COM) [6, 47, 65], defined by $(n, S, A_i, O_i, T, G, h)$, where

- n is the number of agents.
- S is a finite set of world states.
- $A_i = A^W \cup A^C \cup A^M$ is the action space for agent i, consisting of
  - A^W actions for interacting with the world, such as navigation and task-completion signals,
  - A^C communication actions,
  - A^M map tool query actions.
- $O_i = O^V \times O^C \times O^M$ is the observation space for agent i, including
  - O^V visual observations,
  - O^C messages received from other agents,
  - O^M results returned by the map tool.
- T represents the joint transition model.
- G represents the task goal.
- h denotes the task horizon.

Places are predefined indoor or outdoor landmarks in the environment that may serve as candidate gathering points. Agents are also initialized at indoor places to prevent them from being immediately caught by sentinels. Initially, each agent knows a part of these places. To get information about unknown places, they must query the Map Tool.

Sentinels represent dynamic spatial hazards that constrain safe navigation. We include two types: *Stationary Sentinels*, which remain fixed and rotate continually; *Patrolling Sentinels*, which move along predefined patrol routes.

Both types of sentinels stay outdoors and cannot detect entities indoors. A sentinel initiates a warning countdown when an agent stays within its ego-view.

When starting the countdown, it will send a *warning signal* to the agent. The countdown decreases faster when the agent is closer. An agent is considered *caught* when the countdown reaches zero, and will be transported outside of the city immediately. More formally,

$$\text{If } \sum_{p \in G} [label_p = label_{agent}] > \frac{|G|}{1000}, \text{ countdown starts}$$

Where G is the RGB-D image, $label_p$ is the label of pixel p based on the segmentation result, and $label_{agent}$ is the label of this agent. Countdown time t is initially set to 15 when it starts. Then every second,

$$t = t - 1000 \times \frac{\sum_{p \in G} [label_p = label_{agent}]}{|G|}$$

If the agent exits the sentinel's view, the countdown resets, allowing escape but requiring coordinated and cautious path planning.

Map Tool provides coarse spatial information through five query types:

- *QueryRoute*(p_1, p_2): returns a waypoint-based route between p_1 and p_2.
- *QueryNearby*(p_1): returns a list of places near coordinate p_1.
- *QueryPlace*(*place*): returns metadata for a given place (*place*), including its location and bounding box.
- *QueryMap*(): returns an aerial-view binary image of the whole scene, where white denotes navigable space. It is derived from the Virtual Community dataset and serves only as a coarse approximation of the environment rather than an exact representation of the physical layout.
- *QueryRefinedRoute*(*list*[*p*]): return a waypoint-based route. The Map Tool first maps each input coordinate p to the nearest valid waypoint, and then connects the waypoints sequentially, using the same algorithm as *QueryRoute*.

The Map Tool is implemented based on the Virtual Community dataset. It generates waypoints in the city environment, constructs a connected waypoint graph, and performs *QueryRoute* operations using the Dijkstra algorithm [9]. Additional implementation details are provided in the Appendix B. The tool intentionally exposes only coarse, high-level spatial information, requiring agents to communicate, reason about incomplete spatial knowledge, and collaboratively refine plans under uncertainty—mirroring real-world human coordination in large-scale environments.

4 CoSaR: Embodied Cooperative Spatial Reasoning and Planning

CoSaR equips embodied agents with a spatially grounded reasoning architecture that integrates perception, memory, and planning, as shown in Figure 2. After receiving egocentric sensory input and natural-language messages from teammates, the agent first processes visual signals through a dedicated *Perception Module*,

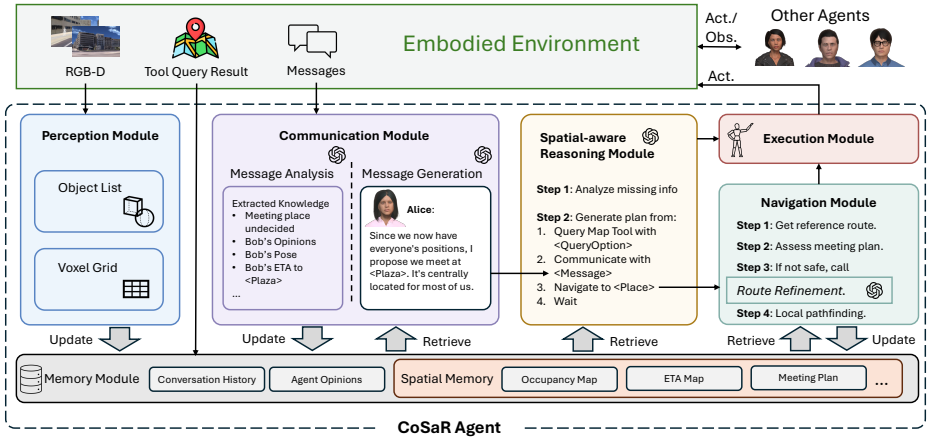

Fig. 2: Method overview. CoSaR maintains a Spatial Memory that integrates perceived visual observations, map-tool responses, and spatial information extracted from natural-language communication. A spatial-aware reasoning module then uses this memory to plan the actions.

which extracts objects, identifies sentinels, and produces semantic segmentation suitable for spatial reasoning (section 4.1). The natural-language messages are analyzed by the *Communication Module*, extracting useful spatial information and cooperation information (section 4.2). These heterogeneous inputs are converted into structured spatial information such as updated poses, travel-time estimates, occupancy patterns, and sentinel danger zones and consolidated into a persistent *Spatial Memory* (Section 4.3), which stores the agent's evolving understanding of the environment, including meeting plans, occpupancy maps encoding traversability and known poses of teammates and sentinels, estimated arrival-time relationships. The memory serves as a unified state representation supporting both communication and navigation. Building on this memory, the *Spatial-aware Reasoning Module* plans the next action by deciding when to communicate with others, query the map tool, or wait, or by guiding movement through the environment (section 4.4). When the agent identifies that its current route intersects a dangerous region, it invokes the *Route Refinement* module, which reconstructs a schematic map from spatial memory and solicits spatial reasoning from a VLM to propose a safer route before obtaining an actionable path from the map tool (section 4.5). Together, these components enable agents to perceive, communicate, and navigate in a spatially grounded manner, forming the core of CoSaR's embodied cooperative spatial reasoning and planning capabilities.

4.1 Perception Module

The agent first processes the observed raw RGB-D input with open-set object detection models [25] to identify all visible objects o, including other agents, and then uses Segment Anything [39] to predict segmentation masks m, and projects each object into a point cloud p with the depth input. To supplement

with semantic features, the agent calculates the CLIP feature [38] for each object with cropped RGB images with respective segmentation masks. A voxel grid g is also constructed with the entire RGB image and depth map for later use. Specifically, the agent matches each perceived object with the sentinel features to identify the existence of the sentinels, and marks the grids occupied by sentinels with warning labels. Therefore, the RGB-D observation is perceived into a list of objects $o(p, f)$ and a voxel grid g.

4.2 Communication Module

The agent uses the communication module to analyze messages received from other agents and generate potential messages to exchange based on current knowledge of the environment and other agents. If there are communication messages, the agent uses LLM to extract potentially useful information, including spatial information such as the current positions of other agents and sentinels, ETA information from an agent to a specific place, and cooperation information such as the opinions of each teammate. With this updated information, the agent determines the most reasonable target location for gathering. To facilitate planning, CoSaR generates potential communication messages to send at each step. To generate reasonable communication messages that facilitate the task, the communication module utilizes LLMs as a message generator and prompts it with key components of our spatial memory, including *Poses list*, *ETA map*, *available places*. We also equipped it with the stored conversation history and previously analyzed opinions of each teammate.

4.3 Spatial Memory

The spatial memory is the core component of our agent that stores the key knowledge about the environment and other teammates. The basic components of the spatial memory include:

Occupancy map and danger zones The agent constructs a local occupancy map that captures ground accessibility with the perceived voxel grids. The known locations of all agents and sentinels are also noted on the map. To support the meeting challenge, the agent further designates **danger zones** surrounding all known sentinel locations. This map is maintained through a combination of visual perception and information exchange from communication.

Estimated Time of Arrival (ETA) Map The ETA map records the estimated travel time of each agent to each candidate place. The agent infers others' ETAs through communication and obtains its own by querying the map tool.

Meeting plan The meeting plan consists of a selected meeting place and a reference route to it. While the meeting place is continuously negotiated and updated through communication, the reference route is initially obtained from the map tool via the *Query Route* action and subsequently maintained through iterative refinement during navigation.

4.4 Spatial-aware Reasoning Module

The Spatial-aware Reasoning Module serves as the central cognitive engine, synthesizing high-level coordination strategies from the agent's underlying spatial

memory. This module systematically interprets stored spatial trajectories and environmental constraints to orchestrate communication and navigation in a decentralized manner.

The spatial-aware reasoning module scans the spatial memory to determine what information remains missing, what should be queried, and what actions are most beneficial for progressing the discussion. We leverage the zero-shot reasoning capabilities of a Large Language Model (LLM), prompting it with a structured representation of the spatial memory. This includes the Pose Registry (agent coordinates), an Estimated Time of Arrival (ETA) Map, and a list of Candidate Meeting Places. By integrating these with the historical dialogue and the inferred preferences of teammates, the LLM selects from four plan types:

- **Query**: Invokes the map tool with specific parameters to reduce spatial uncertainty.
- **Communicate**: Broadcasts a natural language message synthesized by the communication module to align teammate beliefs.
- **Navigate**: Initiates a trajectory toward a prioritized spatial goal.
- **Wait**: Temporarily halts movement to synchronize with teammate arrivals or await critical information.

Each selected action is accompanied by a justification, providing a trace of the agent's reasoning process. Navigation-type plans are subsequently offloaded to the Navigation Module (Sec. 4.5) for low-level execution and obstacle avoidance.

To reduce inference cost, the spatial-aware reasoning module is activated only under the following conditions:

- A new event is observed (e.g., incoming messages or newly detected sentinels).
- The communication module detects disagreement on the meeting place.
- At regular intervals (every 120 seconds).

4.5 Navigation Module

The navigation module translates high-level navigation plans into executable dynamic hazard-aware trajectories. Given a target destination, the agent first queries the map tool for a baseline global route. This route is subjected to a Safety Assessment phase, where it is cross-referenced against known Danger Zones. If a safety violation is detected, the agent triggers a multi-stage Route Refinement process to synthesize a viable detour.

Route Refinement via Multimodal Grounding. When a route is flagged as unsafe, the agent invokes a *QueryMap* action to retrieve a coarse-scale, top-down aerial representation of the environment. As illustrated in Fig. 3, we construct a Schematic Spatial Map by augmenting this visual representation with metadata from the agent's spatial memory, including current ego-pose, detected sentinel coordinates, and the current reference path. This augmented map is then passed to a Vision-Language Model (VLM). We provide the VLM with explicit coordinates for the agent, the destination, and the intersection points of the danger zones and the reference route. The VLM is prompted to return a series of refined coordinates. To account for the inherent spatial stochasticity and potential hallucinations of the VLM, the model's proposed trajectory is passed through a *QueryRefinedRoute*

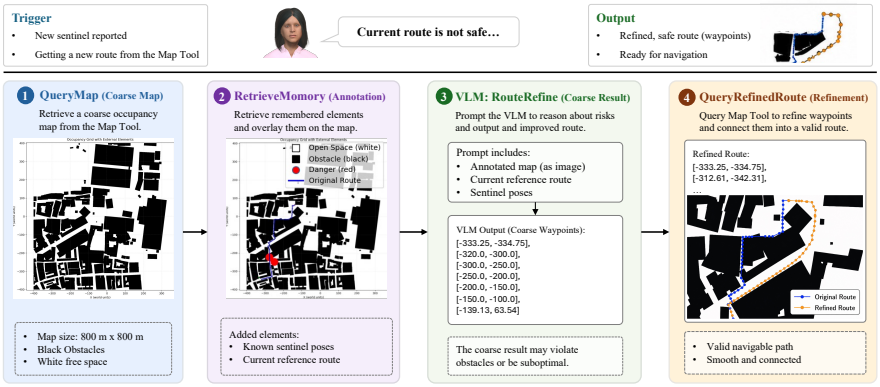

Fig. 3: Route Refinement. The agent will call *QueryMap* and get a coarse-level 2D-map image, which is dynamically augmented with visual prompts representing the current route (blue) and danger zones (red) corresponding to sentinel locations. A Vision-Language Model (VLM) performs zero-shot spatial reasoning over this annotated visual input to synthesize a new route that ensures safe navigation toward the meeting point while maintaining a safe buffer from dynamic threats.

verification step. This step maps the VLM's high-level route back to the valid navigation waypoints that the agent can directly follow. To ensure robustness against repeated VLM failures, we implement a retry-limit mechanism, falling back to a heuristically-modified map-tool route, where we simply remove the waypoints intersecting with any danger zone, if a valid path is not synthesized within N attempts.

Emergency Avoidance. During navigation, certain circumstances may rise, such as a sentinel emerging suddenly after a corner. If the agent decides it is in an immediately dangerous pose, it will perform an emergency avoidance. The further details are included in the **Appendix** C .

Low-level Control. Once the refined waypoints are finalized, the agent employs the A^* algorithm for local pathfinding based on the occupancy map, ensuring fine-grained obstacle avoidance and adherence to the global trajectory in a continuous action space.

5 Sentinel: An Embodied Cooperative Spatial Reasoning and Planning Benchmark

We instantiate the above problem in a city-scale environment using the *Virtual Community* platform [70].

Environment. The benchmark consists of 24 fully annotated scenes. Each scene spans approximately 800 m × 800 m and is constructed from real-world maps. Agents receive egocentric 512 × 512 RGB-D observations and interact with the environment through navigation, communication, and map queries.

Places. Each scene contains 50–150 predefined places, which serve as candidate gathering points. Agents initially possess only partial knowledge of these places

and must discover additional locations through interaction with the environment and the Map Tool.

Agents. Each scene includes up to 15 agents with predefined configurations, including their names, initial places, and knowledge. The number of participating agents in each run is controlled by configurable parameters.

Sentinels. Each scene includes up to 20 sentinels with predefined configurations (both stationary and patrolling). Sentinels are initialized at random locations within the scene. To prevent trivial solutions, a subset of sentinels (e.g., the first five) is placed near the geometric center of agents' initial places, increasing task difficulty. The number of active sentinels in each run is also controlled via configurable parameters.

Metrics. The benchmark reports the following metrics: **Success Rate**, fraction of episodes in which all agents successfully gather within a horizon of 1500 steps; **Caught Rate**: fraction of agents that are captured by sentinels; **Detected Rate**: proportion of time steps in which any agent is detected by a sentinel; **Time Cost**: number of steps taken for all agents to gather; **Distance Traveled**: aggregate distance traveled by all agents.

6 Experiment

6.1 Experimental Setup

Sentinel Challenge. We evaluate our method in the Sentinel Benchmark 5. Task difficulty is controlled by varying the number of sentinels and their motion patterns. We conducted experiments with 3 agents and 5 sentinels of two different moving patterns as introduced in section 3, over 14 diverse scenes. We also evaluate a setting with oracle perception, in which ground-truth segmentation is provided during perception. This significantly improves the probability of correctly detecting a sentinel.

Baselines We compare against six representative baselines:

- **Oracle Centered**: selects the meeting location as the geometric center of all agents, assuming oracle access to their ground-truth positions.
- **Oracle Centered w/ DZ**: extends the above with danger-zone–based sentinel avoidance (Section 4.3).
- **MCTS Planner**: selecting the meeting location using MCTS planners. Each agent uses an MCTS planner to minimize the distance to other agents, the total traveled distance, and the detection rate.
- **RoCo** [29]: a strong multi-agent cooperation method adapted to our setting. RoCo represents methods that plan first and execute later: agents communicate only at the beginning of the task to agree on a meeting place and the full navigation plan.
- **CoELA** [65]: a modular framework powered by foundation models with dynamic communication and planning, enhanced with danger-zone avoidance.
- **MAT** [58]: Multi-Agent Transformer adapted to our setting following the CoELA paper. A shared encoder–decoder Transformer consumes a top-down

Table 1: Results for 5 agents and 10 sentinels We report the average score over 14 scenes and 6 runs here. Best performance is shown in **bold**. The Oracle Centered baseline is excluded from the Oracle-perception scenario, since the method itself does not utilize Oracle perception.

Method	10 Stationary Sentinels					10 Patrolling Sentinels				
	Succ. Rate↑	Caught Rate↓	Detect. Rate↓	Time↓	Dist.↓	Succ. Rate↑	Caught Rate↓	Detect. Rate↓	Time↓	Dist.↓
Sentinel Challenge										
Oracle Centered	7.14	51.43	1.80	1429.86	1800.07	14.29	28.57	1.50	1367.07	1401.93
Oracle Centered w/ DZ	11.90	39.29	1.86	1401.40	1598.50	29.76	22.14	1.66	**1228.89**	1381.91
MCTS	10.71	47.38	2.21	1371.26	2206.20	15.48	40.48	2.02	1338.67	2226.39
RoCo	26.19	26.90	1.13	1255.81	**1333.27**	29.76	28.57	1.73	1306.00	1480.76
CoELA	16.67	25.24	**1.08**	1340.77	1522.53	26.19	**17.62**	0.94	1304.94	1436.53
MAT	0.00	35.71	1.13	1500.00	1453.82	7.14	20.24	**0.78**	1447.17	**1269.10**
CoSaR (Ours)	**32.14**	**21.43**	1.27	**1194.45**	1552.45	**32.14**	27.62	2.22	1246.86	1577.99
Sentinel Challenge w/ Oracle Perception										
Oracle Centered w/ DZ	28.57	24.29	2.09	1237.50	1459.10	35.71	19.76	0.95	**1230.06**	1393.05
MCTS	33.33	22.38	2.55	1196.80	2147.75	19.05	31.43	2.26	1309.76	2393.93
RoCo	46.43	15.00	1.44	1106.30	1332.00	**38.10**	17.14	1.41	1268.10	1472.20
CoELA	34.52	9.52	1.24	1188.02	**1237.82**	36.90	**10.95**	0.66	1248.24	1350.23
MAT	7.14	19.76	1.00	1455.43	1815.91	0.00	11.43	**0.49**	1500.00	**1315.29**
CoSaR (Ours)	**53.57**	**9.05**	**0.97**	**1080.46**	1387.31	**38.10**	20.95	1.72	1247.62	1708.91

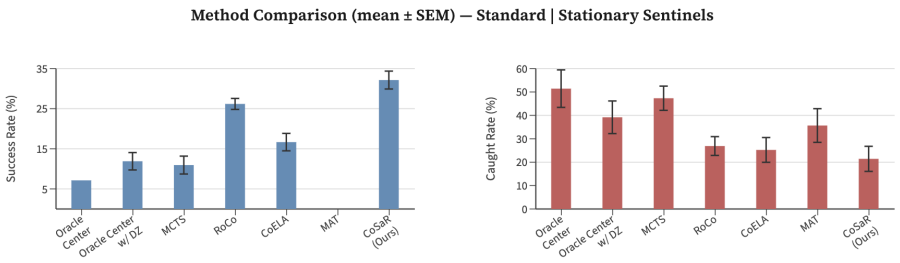

Fig. 4: Method comparison under the standard setting with stationary sentinels. Results are reported as mean ± SEM over 14 scenes with 5 agents and 10 sentinels (6 runs per scene). Left: task success rate (%). Right: caught rate (%).

semantic map plus per-agent features and autoregressively emits joint meeting-place decisions over the same top-K candidate set, trained online with PPO. MAT serves as a centralized end-to-end multi-agent RL baseline.

Implementation Details We use `gpt-4o` as the foundation model backbones for all methods unless otherwise specified. Experiments are conducted on A16 GPUs.

6.2 Results

Failure Reasons. We show the failure reasons breakdown in Figure 5. The reasons for task failure are threefold. First, some agents are captured by sentinels. Second, some agents take too long to reach the target location, exceeding the

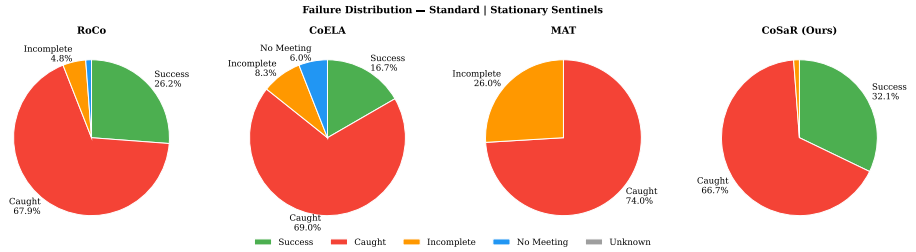

Fig. 5: Failure breakdown (%) under the setting of 5 agents and 10 stationary sentinels over 14 scenes and 6 runs. Most failures lie in unsafe decisions (**Caught**).

time limit. Third, agents may issue a task completion signal while not being located at a common meeting place.

CoSaR Agents generally achieve a higher success rate with lower cost. As shown in Table 1 and Figure 4, CoSaR agents consistently obtain the highest success rate across tasks. The high performance of RoCo is also predictable, as we incorporate part of the spatial memory into its implementation. CoSaR also has the overall lowest *Time* and *Distance* cost, indicating more efficient coordination and faster convergence.

Spatial memory is important. Without dynamic discussion, Oracle Centered Agents and RoCo agents may get stuck at unsafe meeting points. However, their performance remains considerably higher than that of CoELA agents, which support dynamic communication but lack any spatially aware module. This contrast highlights the essential role of spatial memory in effective coordination.

CoSaR Agents can communicate effectively to exchange spatial information. Figure 6 illustrates the qualitative communication process among the agents. The examples demonstrate that, supported by our framework's reasoning and planning capabilities, the agents can respond appropriately to communicative requirements and pose pertinent questions that help advance the discussion. During the meeting procedure, certain circumstances may arise, like a new sentinel or invalid reference route, which will result in an increase in the current estimated time to arrival. Through coordinated decision-making, the agents are also able to address unexpected difficulties that arise during the meeting process, adjusting their meeting location accordingly.

CoSaR performs best across different numbers of agents and sentinels. To evaluate whether our proposed method is effective across different numbers of agents and sentinels, we experimented with three additional settings:

- 3 agents, 10 sentinels.
- 5 agents, 5 sentinels.
- 5 agents, 20 sentinels.

The results are shown in Table 2. As shown in the results, increasing the agent number or the sentinel number makes the challenge overall harder to complete, since the possibility of an agent being caught gets higher, and it requires more agents to arrive at the same place. CoSaR maintains a consistently lower caught rate across varying numbers of agents, with its advantage increasing as the team

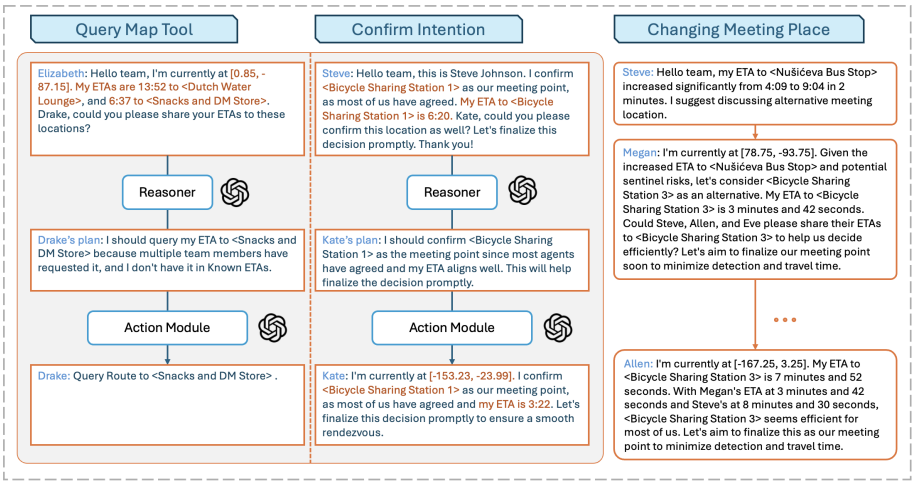

Fig. 6: Case Study. The left panels demonstrate how CoSaR agents decide their next action when actively communicating. The right panel showcases an example of how CoSaR agents successfully change their meeting place upon changing circumstances. More case studies are provided in the **Appendix** D.

size grows. Moreover, CoSaR prevents its success rate from deteriorating as sharply as the other methods, demonstrating stronger robustness under multi-agent settings.

Efficiency-Safety Tradeoff. As shown in Table 1 and Table 2, CoELA occasionally achieves a lower caught rate than CoSaR. However, this does not reflect safer navigation. Instead, CoELA tends to spend much more time in initial locations, where agents are not exposed to sentinels. While this reduces the caught rate, it also delays coordination and movement, leading to a lower overall success rate.

6.3 Ablation Study

To systematically evaluate the effectiveness of different components in our method, we conduct an ablation study by removing each of the following four components:

1. Route Refinement: The retry limit for route refinement attempts is set to zero.
2. Message Analyzer: The message analyzer is not invoked when a new message is received. Consequently, the ETA map and the list of known poses are not stored, and the summarized opinions of each agent are unavailable.
3. Spatial Memory: Agents are disabled from accessing spatial memory when making decisions. As a result, they cannot utilize the ETA map or the list of known poses.
4. Emergency Avoidance: Agents do not perform emergency avoidance when being too close to any sentinel. This leads to bolder navigation behavior.

The results are shown in Table 3. The success rate consistently decreases as these components are removed, with the largest drop occurring when the ETA

Table 2: CoSaR performs best across different numbers of agents and sentinels. We report the mean success rate and detected rate under the patrolling sentinel setting over 14 scenes and 2 runs.

Method	**3 Agents 10 sentinels**		**5 Agents 5 sentinels**		**5 Agents 10 sentinels**		**5 Agents 20 sentinels**	
	Success Rate↑	Caught Rate↓	Success Rate↑	Caught Rate↓	Success Rate↑	Caught Rate↓	Success Rate↑	Caught Rate↓
Oracle Center	21.43	40.48	7.14	51.43	7.14	51.43	7.14	57.14
Oracle Center w/ DZ	21.43	40.48	14.29	32.86	7.14	38.46	7.17	44.62
MCTS	25.00	33.33	17.86	40.00	7.14	50.71	7.14	40.77
RoCo [29]	28.57	27.38	35.71	25.71	28.57	24.29	17.86	36.43
CoELA [65]	39.29	**26.19**	21.43	22.14	21.43	**21.43**	7.14	43.08
CoSaR (Ours)	**42.86**	27.38	**39.29**	**20.00**	**35.71**	24.29	**21.43**	**32.14**

Table 3: Ablation Results. We report the results under the setting of 5 agents and 10 stationary sentinels over 14 scenes and 6 runs here.

Method	Success Rate↑	Caught Rate↓	Detection Rate↓	Time.↓	Dist.↓
CoSaR (Ours)	**32.14**	**21.43**	1.27	**1194.45**	1552.45
without Route Refinement	26.19	28.80	1.18	1243.31	**1446.25**
without Message Analyzer	19.26	31.38	1.29	1310.20	1488.76
without Spatial Memory	21.42	31.19	1.27	1283.70	1489.28
without Emergency Avoidance	24.05	31.95	**1.15**	1268.58	1513.40

map and the list of known poses are unavailable due to disabling the Analyzer and Spatial Memory. This mainly reduces the agents' ability to identify suitable meeting locations and increases their exposure to sentinel threats.

6.4 Backbone Comparison

To compare the performance of our method across different Foundation Model backbones, we additionally experimented using `Qwen3-VL-30B-A3B-Instruct` with our method. As shown in Table 4, our method with `Qwen` backbone also achieves competing performance, demonstrating the robustness and effectiveness of our methods across different backbones.

7 Discussion

Beyond robotics applications—such as search-and-rescue, human-aware navigation, and autonomous delivery under dynamic constraints—this benchmark is also motivated by large-scale multi-agent environments, including game-like settings where agents must coordinate while avoiding hostile or visibility-constrained entities. In this formulation, sentinels abstract dynamic hazards such as guards, adversaries, surveillance regions, or unsafe zones, requiring agents to jointly reason about risk, visibility, and coordination.

Table 4: Performance Comparison Across Foundation Model Backbones. We report the results under the setting of 5 agents and 5 stationary sentinels over 14 scenes and 2 runs here.

Method	Success Rate↑	Caught Rate↓	Detection Rate↓	Time.↓	Dist.↓
`gpt-4o`	67.85	5.00	0.57	950.5	1172.15
`Qwen3-VL-30B`	60.71	7.69	0.77	1013.38	1292.97

More broadly, the benchmark is designed to study cooperative spatial reasoning and communication under partial observability and evolving constraints. It provides a controlled setting for analyzing how agents share information, adapt plans, and maintain coordination in the presence of dynamic risks.

Dual-use Considerations. We acknowledge that capabilities for coordinated navigation and inference under partial observability are inherently dual-use. While such capabilities may benefit applications such as search-and-rescue, disaster response, and autonomous logistics, they could also be adapted for surveillance-oriented tasks, including non-consensual tracking or monitoring.

Potential Mitigation. Our benchmark mitigates several of these risks at the source: it is conducted entirely in simulation and involves only synthetic agents and environments. Consequently, no real-world human data, biometric information, or personally identifiable information is used. The primary risk therefore stems from the transfer of learned methods rather than from the benchmark artifact itself. To further reduce misuse risks, we will release the benchmark with explicit usage restrictions and document both intended and out-of-scope applications. Further discussions are included in the **Appendix** A.

8 Conclusion

In this work, we introduced the Sentinel Challenge as a benchmark for studying embodied multi-agent cooperation under dynamic and adversarial spatial constraints. To address this setting, we proposed CoSaR, a cooperative spatial reasoning and planning framework that equips agents with a dynamic spatial memory for integrating perception, communication, and map-based signals. This unified representation enables agents to share information effectively, reason about evolving spatial structure, and adapt their plans through coordinated replanning. Across 14 diverse city-scale environments involving 3–5 agents, CoSaR consistently outperforms strong baselines, demonstrating the importance of spatially grounded memory and communication in challenging multi-agent navigation tasks. We believe Sentinel Challenge and CoSaR provide a promising foundation for future research on large-scale embodied cooperation, resilient multi-agent planning, and spatially grounded communication.

Acknowledgement

This work was supported by NSF IIS-2441250, NSF IIS-2404386 and MURI.

References

1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)
2. Amato, C., Konidaris, G., Kaelbling, L.P., How, J.P.: Modeling and planning with macro-actions in decentralized pomdps. Journal of Artificial Intelligence Research **64**, 817–859 (2019)
3. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
4. Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., Mordatch, I.: Emergent tool use from multi-agent autocurricula. In: International Conference on Learning Representations (2020), `https://openreview.net/forum?id=SkxpxJBKwS`
5. Bard, N., Foerster, J.N., Chandar, S., Burch, N., Lanctot, M., Song, H.F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al.: The hanabi challenge: A new frontier for ai research. Artificial Intelligence **280**, 103216 (2020)
6. Bernstein, D.S., Givan, R., Immerman, N., Zilberstein, S.: The complexity of decentralized control of markov decision processes. Mathematics of operations research **27**(4), 819–840 (2002)
7. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
8. Carroll, M., Shah, R., Ho, M.K., Griffiths, T., Seshia, S., Abbeel, P., Dragan, A.: On the utility of learning about humans for human-ai coordination. Advances in neural information processing systems **32** (2019)
9. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik **1**(1), 269–271 (1959)
10. Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual understanding and reasoning. arXiv preprint arXiv:2403.11401 (2024)
11. Gadre, S.Y., Wortsman, M., Ilharco, G., Schmidt, L., Song, S.: Clip on wheels: Zero-shot object navigation as object localization and exploration. arXiv preprint arXiv:2203.10421 **3**(4), 7 (2022)
12. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36**, 20482–20494 (2023)
14. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. arXiv preprint arXiv:2210.05714 (2022)
15. Jaderberg, M., Czarnecki, W.M., Dunning, I., Marris, L., Lever, G., Castaneda, A.G., Beattie, C., Rabinowitz, N.C., Morcos, A.S., Ruderman, A., et al.: Human-level performance in 3d multiplayer games with population-based reinforcement learning. Science **364**(6443), 859–865 (2019)
16. Jain, U., Weihs, L., Kolve, E., Farhadi, A., Lazebnik, S., Kembhavi, A., Schwing, A.: A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 471–490. Springer (2020)

17. Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., Fan, L.: Vima: General robot manipulation with multimodal prompts. In: Fortieth International Conference on Machine Learning (2023)
18. Kahn, G., Abbeel, P., Levine, S.: Badgr: An autonomous self-supervised learning-based navigation system. IEEE Robotics and Automation Letters **6**(2), 1312–1319 (2021)
19. Khandelwal, A., Weihs, L., Mottaghi, R., Kembhavi, A.: Simple but effective: Clip embeddings for embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14829–14838 (2022)
20. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
21. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017)
22. Kümmerle, R., Ruhnke, M., Steder, B., Stachniss, C., Burgard, W.: A navigation system for robots operating in crowded urban environments. In: 2013 IEEE International Conference on Robotics and Automation. pp. 3225–3232. IEEE (2013)
23. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: Conference on Robot Learning. pp. 80–93. PMLR (2023)
24. Liu, J., Zhou, P., Du, Y., Tan, A.H., Snoek, C.G., Sonke, J.J., Gavves, E.: Capo: Cooperative plan optimization for efficient embodied multi-agent cooperation. arXiv preprint arXiv:2411.04679 (2024)
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
26. Liu, X., Li, J., Jiang, Y., Sujay, N., Yang, Z., Zhang, J., Abanes, J., Zhang, J., Feng, C.: Citywalker: Learning embodied urban navigation from web-scale videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6875–6885 (2025)
27. Lowe, R., Tamar, A., Harb, J., Pieter Abbeel, O., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems **30** (2017)
28. Majumdar, A., Aggarwal, G., Devnani, B., Hoffman, J., Batra, D.: Zson: Zero-shot object-goal navigation using multimodal goal embeddings. Advances in Neural Information Processing Systems **35**, 32340–32352 (2022)
29. Mandi, Z., Jain, S., Song, S.: Roco: Dialectic multi-robot collaboration with large language models. arXiv preprint arXiv:2307.04738 (2023)
30. Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., Artzi, Y.: Mapping instructions to actions in 3d environments with visual goal prediction. arXiv preprint arXiv:1809.00786 (2018)
31. Mon-Williams, R., Li, G., Long, R., Du, W., Lucas, C.G.: Embodied large language models enable robots to complete complex tasks in unpredictable environments. Nature Machine Intelligence pp. 1–10 (2025)
32. Morales, Y., Carballo, A., Takeuchi, E., Aburadani, A., Tsubouchi, T.: Autonomous robot navigation in outdoor cluttered pedestrian walkways. Journal of Field Robotics **26**(8), 609–635 (2009)
33. OpenAI: Gpt-4 technical report (2023)

34. Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., Hakkani-Tur, D.: Teach: Task-driven embodied agents that chat. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2017–2025 (2022)
35. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442 (2023)
36. Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.H., Tenenbaum, J.B., Fidler, S., Torralba, A.: Watch-and-help: A challenge for social perception and human-ai collaboration. In: International Conference on Learning Representations (2021)
37. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., et al.: Habitat 3.0: A co-habitat for humans, avatars and robots. arXiv preprint arXiv:2310.13724 (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), https://arxiv.org/abs/2408.00714
40. Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T.G., Hung, C.M., Torr, P.H., Foerster, J., Whiteson, S.: The starcraft multi-agent challenge. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 2186–2188 (2019)
41. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
42. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Ving: Learning open-world navigation with visual goals. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13215–13222. IEEE (2021)
43. Shah, D., Osiński, B., Levine, S., et al.: Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. pp. 492–504. PMLR (2023)
44. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation. arXiv preprint arXiv:2306.14846 (2023)
45. Sharma, P., Torralba, A., Andreas, J.: Skill induction and planning with latent language. arXiv preprint arXiv:2110.01517 (2021)
46. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2998–3009 (2023)
47. Spaan, M.T., Gordon, G.J., Vlassis, N.: Decentralized planning under uncertainty for teams of communicating agents. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. pp. 249–256 (2006)
48. Sridhar, A., Shah, D., Glossop, C., Levine, S.: Nomad: Goal masked diffusion policies for navigation and exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 63–70. IEEE (2024)

49. Stone, P., Veloso, M.: Multiagent systems: A survey from a machine learning perspective. Autonomous Robots **8**, 345–383 (2000)
50. Suarez, J., Du, Y., Isola, P., Mordatch, I.: Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents. arXiv preprint arXiv:1903.00784 (2019)
51. Sumers, T., Yao, S., Narasimhan, K., Griffiths, T.L.: Cognitive architectures for language agents. arXiv preprint arXiv:2309.02427 (2023)
52. Szot, A., Jain, U., Batra, D., Kira, Z., Desai, R., Rai, A.: Adaptive coordination in social embodied rearrangement. In: International Conference on Machine Learning. pp. 33365–33380. PMLR (2023)
53. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291 (2023)
54. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432 (2023)
55. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19757–19767 (2024)
56. Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., Liang, Y.: Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
57. Wang, Z., Cai, S., Liu, A., Ma, X., Liang, Y.: JARVIS-1: Open-world multi-task agents with memory-augmented multimodal language models. In: Second Agent Learning in Open-Endedness Workshop (2023), `https://openreview.net/forum?id=xzPkZyHlOW`
58. Wen, M., Kuba, J., Lin, R., Zhang, W., Wen, Y., Wang, J., Yang, Y.: Multi-agent reinforcement learning is a sequence modeling problem. Advances in Neural Information Processing Systems **35**, 16509–16521 (2022)
59. Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., Savva, M., Batra, D.: Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. arXiv preprint arXiv:1911.00357 (2019)
60. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023)
61. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)
62. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11097–11107 (2020)
63. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
64. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3d-mem: 3d scene memory for embodied exploration and reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17294–17303 (2025)

65. Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J.B., Shu, T., Gan, C.: Building cooperative embodied agents modularly with large language models (2023)
66. Zhang, H., Wang, Z., Lyu, Q., Zhang, Z., Chen, S., Shu, T., Dariush, B., Lee, K., Du, Y., Gan, C.: Combo: compositional world models for embodied multi-agent cooperation. arXiv preprint arXiv:2404.10775 (2024)
67. Zhang, H., Zhang, Z., Wang, Z., Zhang, Z., Fang, L., Zhou, Q., Gan, C.: Ella: Embodied social agents with lifelong memory. arXiv preprint arXiv:2506.24019 (2025)
68. Zheng, D., Huang, S., Zhao, L., Zhong, Y., Wang, L.: Towards learning a generalist model for embodied navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13624–13634 (2024)
69. Zhou, Q., Chen, S., Wang, Y., Xu, H., Du, W., Zhang, H., Du, Y., Tenenbaum, J.B., Gan, C.: Hazard challenge: Embodied decision making in dynamically changing environments (2024)
70. Zhou, Q., Zhang, H., Lin, X., Zhang, Z., Chen, Y., Liu, W., Zhang, Z., Chen, S., Fang, L., Lyu, Q., et al.: Virtual community: An open world for humans, robots, and society. arXiv preprint arXiv:2508.14893 (2025)
71. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)
72. Zu, L., Lin, L., Fu, S., Zhao, N., Zhou, P.: Collaborative tree search for enhancing embodied multi-agent collaboration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29513–29522 (2025)