

# TAB-Fields: A Maximum Entropy Framework for Mission-Aware Adversarial Planning

**Gokul Puthumanailam\***  
**Jae Hyuk Song\***

GOKULP2@ILLINOIS.EDU  
JHSONG2@ILLINOIS.EDU

*University of Illinois Urbana-Champaign, USA*

**Nurzhan Yesmagambet**

NURZHAN.YESMAGAMBET@KAUST.EDU.SA

**Shinkyu Park**

SHINKYU.PARK@KAUST.EDU.SA

*King Abdullah University of Science and Technology, Saudi Arabia*

**Melkior Ornik**

MORNIK@ILLINOIS.EDU

*University of Illinois Urbana-Champaign, USA*

**Editors:** N. Ozay, L. Balzano, D. Panagou, A. Abate

## Abstract

Autonomous agents operating in adversarial scenarios face a fundamental challenge: while they may know their adversaries’ high-level objectives, such as reaching specific destinations within time constraints, the exact policies these adversaries will employ remain unknown. Traditional approaches address this challenge by treating the adversary’s state as a partially observable element, leading to a formulation as a Partially Observable Markov Decision Process (POMDP). However, the induced belief-space dynamics in a POMDP require knowledge of the system’s transition dynamics, which, in this case, depend on the adversary’s unknown policy. Our key observation is that while an adversary’s exact policy is unknown, their behavior is necessarily constrained by their mission objectives and the physical environment, allowing us to characterize the space of possible behaviors without assuming specific policies. In this paper, we develop Task-Aware Behavior Fields (TAB-Fields), a representation that captures adversary state distributions over time by computing the most unbiased probability distribution consistent with known constraints. We construct TAB-Fields by solving a constrained optimization problem that minimizes additional assumptions about adversary behavior beyond mission and environmental requirements. We integrate TAB-Fields with standard planning algorithms by introducing TAB-conditioned POMCP, an adaptation of Partially Observable Monte Carlo Planning. Through experiments in simulation with underwater robots and hardware implementations with ground robots, we demonstrate that our approach achieves superior performance compared to baselines that either assume specific adversary policies or neglect mission constraints altogether. Evaluation videos and code: <https://tab-fields.github.io>.

**Keywords:** Adversarial planning, Mission-constrained planning, Planning under uncertainty

## 1. Introduction

Effective autonomy in adversarial settings remains a fundamental problem in autonomous systems (Gu et al., 2014; Agmon, 2017; Huang et al., 2019). A core challenge in such settings lies in reasoning about the adversary’s state and its future trajectories, especially when critical aspects of their behavior—such as decision-making policies—are unknown (Paruchuri, 2007). This lack of knowledge is further complicated by environmental factors like obstacles, terrain constraints, and dynamic operational constraints. One way to address this challenge is to treat the adversary’s state

---

\* Equal contribution

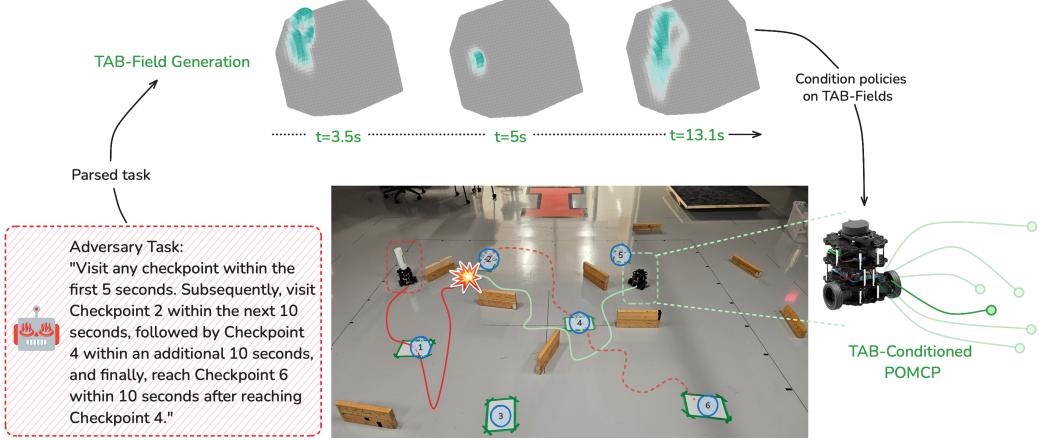


Figure 1: Overview of the proposed approach applied to an interception task. The adversary’s task is defined by mission objectives and environmental constraints (left). TAB-Fields are generated over time (top) to represent adversary state distributions and integrated into the planning process via TAB-conditioned POMCP (right). The resulting trajectories show the adversary’s path (red line), the agent’s response (green line), and the interception area (💥).

as a partially observable element within a broader system (Gronauer and Diepold, 2022; Zhang et al., 2020). In this extended state space, the problem can be described as a Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998). POMDPs enable reasoning about uncertainty through belief dynamics—probability distributions over possible states—thereby facilitating structured decision-making. However, a fundamental obstacle arises in this context: the transition dynamics of the system depend on the adversary’s unknown policy, making them inherently indeterminate. Traditional POMDP planning methods rely on a priori knowledge of transition dynamics (Castellini et al., 2021), which is unavailable here.

Our key observation is that while an adversary’s exact policy is unknown, their behavior is necessarily constrained by their mission objectives and the physical environment. Building on this observation, we propose an alternative approach: instead of assuming a specific adversary policy, we characterize the entire space of possible adversary behaviors that satisfy known mission objectives and environmental constraints. The key idea behind our approach is grounded in the principle of maximum entropy (Jaynes, 1982)—*among all probability distributions consistent with the given constraints, the one with the highest entropy offers the most unbiased and comprehensive representation of the current state of knowledge*. Leveraging this principle, we construct a distribution over adversary states that encapsulates the uncertainty in their decision-making while remaining consistent with all available information.

This perspective shifts the focus from predicting specific adversarial behavior to reasoning about them in a mission-aware manner. Through this lens, we introduce Task-Aware Behavior Fields (TAB-Fields), a novel representation that encodes adversary state distributions over time using constrained entropy maximization. As shown in Figure 1, TAB-Fields capture the evolution of belief states, demonstrating their ability to focus the distribution on regions consistent with mission constraints. TAB-Fields enable us to directly integrate adversary behavior into the belief update and planning process without relying on explicit policy assumptions or extensive training data.

*Statement of Contributions.* The primary contribution of this work is TAB-Fields, a novel representation that captures the distribution of possible adversary states through principled entropy maximization subject to mission and environmental constraints. We show how this representation can be effectively integrated with existing planners through TAB-conditioned POMCP, an adaptation that maintains computational tractability while leveraging our structured representation. Through comprehensive evaluation across diverse scenarios in both simulation and hardware experiments, we demonstrate that our approach significantly improves mission-constrained adversarial planning compared to existing methods.

## 2. Related Work

The presented work intersects several research areas in adversarial planning, behavior prediction, and planning under uncertainty. We discuss and highlight how our method differs from prior work.

*Planning Under Uncertainty.* Planning in environments with uncertainty has been extensively studied within the framework of Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998). Traditional POMDP solvers (Silver and Veness, 2010; Somanı et al., 2013) rely on known transition and observation models (Shani et al., 2013; Lauri et al., 2022) to perform belief updates and compute optimal policies. However, when the environment includes other agents with unknown policies—such as adversaries—the transition dynamics become partially unknown, complicating standard POMDP approaches (Ng et al., 2010; Egorov et al., 2016). Several works have extended POMDP frameworks to handle interactions with other agents. Interactive POMDPs (Han and Gmytrasiewicz, 2018, 2019) model other agents by maintaining beliefs over their beliefs and policies, but this quickly becomes intractable due to the curse of dimensionality. Decentralized POMDPs (Czechowski and Oliehoek, 2021) consider multiple cooperative agents, but are less suited for adversarial settings.

*Modeling Adversary Behavior.* In surveillance and security domains, predicting adversary behavior is critical. Traditional methods (Zhou and Tokekar, 2021; Santos Jr and Zhao, 2006; Santos Jr et al., 2008) often assume specific models of adversary policies, such as rational decision-makers optimizing a known utility function (Zuckerman et al., 2012). However, these assumptions may not hold in practice, leading to ineffective strategies. To mitigate this, some approaches use learning-based methods to model adversary behavior from observed data (Abouelyazid, 2023; Huang et al., 2019). While effective when ample data is available, these methods struggle when observations are sparse. Robust planning methods consider worst-case scenarios without relying on specific adversary models (Nilim and El Ghaoui, 2005; Iyengar, 2005). However, these can be overly conservative.

*Maximum Entropy Methods for Behavior Prediction.* The principle of maximum entropy (Chen and Han, 2014; Savas et al., 2019) has been employed to model behavior under uncertainty with known constraints (Jaynes, 1982). In the context of prediction, maximum entropy methods have been used to model motion (Pfeiffer et al., 2016; Ziebart et al., 2009; Korbacher and Tordeux, 2022), where the goal is to predict likely paths based on environmental features and goal destinations. Savas et al. (2018, 2019) applies the idea to design policies for agents under temporal logic constraints by maximizing entropy in constrained MDPs. Maximum entropy inverse reinforcement learning (IRL) (Ziebart et al., 2008; Aghasadeghi and Bretl, 2011) tackles this problem from a different perspective by recovering reward functions that explain observed behavior, without assuming specific policies. However, IRL requires observed trajectories for learning (Arora and Doshi, 2021; Adams et al., 2022), which may not be available in adversarial settings.

*Belief Planning with Unknown Dynamics* When transition models are partially unknown, belief planning becomes challenging. Methods like Robust MDPs (Wiesemann et al., 2013; Yang et al., 2023) and exploration-exploitation algorithms (Auer et al., 2008; Cheung, 2019) address uncertainty by optimizing for the worst-case scenario or learning the dynamics online. In the context of POMDPs, Thrun (1999) propose Monte Carlo POMDPs, where transition probabilities are sampled from a distribution to account for uncertainty. Abbeel and Ng (2005) address model uncertainty by learning models during planning. Puthumanailam et al. (2024a) extends this work to learn the transition probabilities in dynamic, time-varying POMDPs. Our approach avoids the need to learn the adversary’s transition dynamics by directly computing the distribution over possible states using the maximum entropy principle and known mission constraints. Some works consider planning under model uncertainty using robust or risk-sensitive approaches (García and Fernández, 2015). However, these methods typically do not leverage known constraints or objectives of other agents.

Our work differs from these approaches by avoiding assumptions about adversary policies or the need for behavior data. Instead of learning from demonstrations like maximum entropy IRL or using nested belief hierarchies as in I-POMDPs, we leverage mission specifications and environmental constraints to compute adversary state distributions through maximum entropy principles. This enables efficient real-time planning without requiring extensive adversary modeling or becoming overly conservative like robust planning methods. By integrating these distributions directly into the POMDP framework, we maintain computational tractability while making informed predictions about adversary behavior based on known constraints.

### 3. Preliminaries

We consider an ego agent operating in a shared environment with an adversary. The adversary’s mission objectives are known, but their exact policy and decision-making processes remain unknown. The environment contains obstacles and operational constraints that affect all agents’ feasible actions. Additionally, certain areas provide full observability of the adversary, while in other areas, the adversary is unobservable—a common scenario in surveillance missions where checkpoints or security cameras offer intermittent visibility.

*Objective.* Our primary problem is to enable the ego agent to plan effectively in this environment without knowledge of the adversary’s decision-making process, while maximizing its objectives encoded in the reward function. Given that the adversary’s state is partially observable, we can formulate this as a POMDP. A POMDP typically enables planning through belief space dynamics. However, the transition dynamics of the adversary depend on its unknown policy, making the transition probabilities involving the adversary’s state indeterminate—complicating the application of traditional POMDP methods, which typically require known transition models for belief updates and planning. Instead of assuming a specific adversary policy—which could lead to brittle or exploitable behaviors—we seek an approach to reason about the space of possible adversary behaviors.

*POMDP formulation.* Formally, we define our problem as a POMDP tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, \gamma \rangle$ . The joint state space  $\mathcal{S}$  encompasses our autonomous agent, adversary, and the environment, with states defined as  $s_t = (s_t^a, s_t^{adv}, s^e)$ , where  $s_t^a \in \mathcal{S}^a$  represents our agent’s state,  $s_t^{adv} \in \mathcal{S}^{adv}$  represents the adversary’s state, and  $s^e \in \mathcal{S}^e$  represents the static environment state. The action space  $\mathcal{A}$  comprises all available actions for our autonomous agent. The observation space  $\mathcal{O}$  is defined as  $o_t = (o_t^a, o_t^{adv}, o^e)$ , where  $o_t^a$  is our agent’s fully observable state,  $o_t^{adv}$  represents potentially partial observations of the adversary, and  $o^e$  represents environmental observations. The transition function

$T(s_{t+1} \mid s_t, a_t)$ , observation function  $O(o_{t+1} \mid s_{t+1}, a_t)$ , reward function  $R(s_t, a_t)$  capturing the agent’s objectives [where the reward depends on the both the state of the agent and the adversary], and discount factor  $\gamma$  follow standard POMDP definitions. Note that since the adversary’s policy is unknown, the component of  $T$  involving  $s_{t+1}^{adv}$  is indeterminate.

*Adversary missions.* This paper’s scope considers the adversary’s tasks to be specified in natural language defining the mission objectives and environment constraints. These specifications are further processed into an ordered sequence of constraint tuples  $\mathcal{M} = \{(s_{g_i}, t_{c_i}, \text{type}_i, \theta_i)\}_{i=1}^n$ , where each tuple specifies a goal state  $s_{g_i}$ , temporal constraints  $t_{c_i}$ , constraint type (exact time, deadline, until, eventually, or always), and additional constraints  $\theta_i$  such as speed limits or restricted zones. Since many prior works (Puthumanailam et al., 2024b; Jie et al., 2017) have focused on this conversion process, we do not explicitly address it here.

The objective is to compute an optimal policy  $\pi^*$  for the ego agent that maximizes the expected cumulative reward  $\mathbb{E}[\sum_{t=0}^T \gamma^t R(s_t, a_t) \mid \pi, b_0]$  while maintaining the belief state  $b_t(s)$  over possible states. This belief is updated recursively based on observations through the standard Bayesian update. The core challenge lies in performing effective belief updates and planning to maximize the agent’s reward function, despite not knowing how the adversary’s state evolves over time.

#### 4. Mission-Aware Adversary Behavior Representation

To enable effective belief updates and planning, we need a principled way to reason about the adversary’s possible states and transitions that is consistent with their known mission objectives and environmental constraints, without assuming knowledge of their specific policies. This problem is closely related to the Schrödinger bridge problem in stochastic processes (Marino and Gerolin, 2020), which seeks the most probable evolution of a system between two end-point distributions while minimizing deviation from a reference process (Léonard, 2013).

We adopt the principle of maximum entropy (Jaynes, 1957), which states that among all probability distributions satisfying given constraints, the one with the highest entropy is the most unbiased representation of the current state of knowledge. In our context, this means we seek the distribution that satisfies all known mission and environmental constraints while making the minimum number of additional assumptions about the adversary’s behavior.

A trajectory of the adversary through the environment can be represented as a sequence of states  $s_{0:T}^{adv} = (s_0^{adv}, \dots, s_T^{adv})$ , where  $s_t^{adv}$  represents the adversary’s state at time  $t$ . Let  $Q(s_{0:T}^{adv})$  denote a reference probability distribution representing physically feasible transitions based on environmental constraints and dynamics. This reference process, similar to uncontrolled dynamics in KL control (Todorov, 2009), assigns zero probability to infeasible paths (e.g., through obstacles) and encodes basic motion constraints. We seek a distribution  $P(s_{0:T}^{adv})$  that incorporates mission constraints while remaining as close as possible to  $Q$ , thereby providing a prediction of the adversary’s state evolution for use in belief updates. We formulate this as a constrained optimization problem:

$$\begin{aligned} \min_P D_{\text{KL}}(P \parallel Q) &= \sum_{s_{0:T}^{adv}} P(s_{0:T}^{adv}) \log \left( \frac{P(s_{0:T}^{adv})}{Q(s_{0:T}^{adv})} \right) \\ \text{subject to: } P(s_0^{adv}) &= \mu_0(s_0^{adv}), \quad (\text{initial state}) \\ \mathbb{E}_P[f_{\mathcal{M}}(s_{0:T}^{adv})] &= c_{\mathcal{M}}, \quad (\text{mission constraints}) \\ P(s_t^{adv} \in \mathcal{C}) &= 0, \quad \forall t \in [0, T], \quad (\text{environment constraints}). \end{aligned} \tag{1}$$

The optimization in equation (1) extends techniques from maximum entropy IRL (Ziebart et al., 2008), where similar techniques are used to model expert behavior without assuming specific reward functions. In equation (1),  $D_{KL}(P \parallel Q)$  measures how much additional information P contains beyond what is implied by the reference process Q.  $f_{\mathcal{M}}$  represents a vector of constraint functions derived from the mission specification tuples in  $\mathcal{M}$ . Each function maps trajectories to binary values indicating whether they satisfy the corresponding requirement. For example, given a mission tuple  $(s_g, t_c, \text{exact}, \theta)$ , the corresponding constraint function evaluates to 1 if and only if the trajectory reaches state  $s_g$  at time  $t_c$  while satisfying additional requirements  $\theta$ . The environment constraints ensure that at each timestep, trajectories through prohibited states ( $\mathcal{C}$ ) have zero probability.

The solution to the optimization problem (1) takes a form characteristic of the exponential family of probability distributions (Thomas and Joy, 2006), commonly arising in maximum entropy problems:

$$P^*(s_{0:T}^{adv}) = \frac{1}{Z} Q(s_{0:T}^{adv}) \exp(-\lambda^T f_{\mathcal{M}}(s_{0:T}^{adv}))$$

where Z is the normalization constant and  $\lambda_i$  are Lagrange multipliers corresponding to each constraint  $f_i$ . This solution modifies the reference distribution Q through exponential terms that enforce mission constraints, similar to how the Schrödinger bridge problem modifies a prior process to satisfy endpoint constraints (Léonard, 2013).

While computing this distribution exactly is intractable due to the high-dimensional state space, we can efficiently compute the marginal distributions  $P^*(s_t^{adv})$  using iterative algorithms from probabilistic graphical models (Koller, 2009). These marginal distributions over time form our Task-Aware Behavior Fields (TAB-Fields).

#### 4.1. TAB-Conditioned Planning

Building on TAB-Fields, we now address how to effectively integrate them into the planning process. In our setting, the adversary's transition dynamics depend on their unknown policy, making standard POMDP planning approaches inapplicable. Instead of assuming a specific adversary policy, we use TAB-Fields as a surrogate for the unknown transition dynamics. The intuitive idea is to perform belief updates using TAB-Fields to predict the adversary's state evolution. Specifically, when a new observation  $o_{t+1}^{adv}$  is received, we update our belief over the adversary's state as:

$$b_{t+1}(s_{t+1}^{adv}) = \eta \cdot O(o_{t+1}^{adv} | s_{t+1}^{adv}) \cdot P(s_{t+1}^{adv})$$

where  $P(s_{t+1}^{adv})$  is the probability distribution provided by TAB-Fields and  $\eta$  is a normalization constant. When no observations are available, the belief evolves according to TAB-Fields distribution.

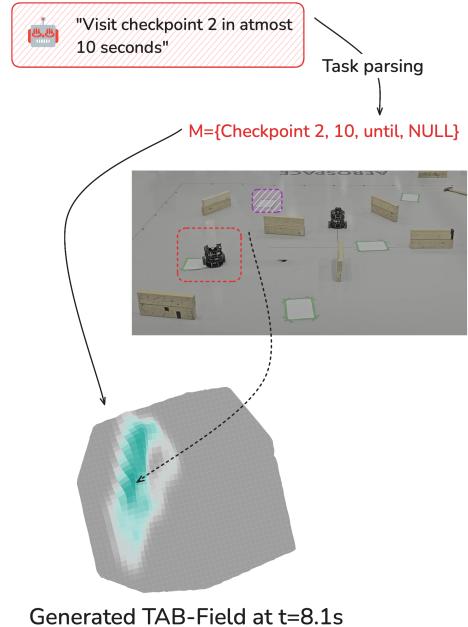


Figure 2: Example mission and its TAB-Field, where darker areas indicate higher probability of adversary presence. Red area denotes adversary start position and purple area indicates the goal checkpoint.

*TAB-POMCP.* While any POMDP solver could potentially be conditioned on TAB-Fields, we demonstrate our approach on POMCP (Silver and Veness, 2010) due to its ability to handle large state spaces efficiently and its natural integration with particle-based belief representations. In standard POMCP, particles representing possible states are propagated using known transition dynamics. Our TAB-conditioned variant instead uses TAB-Fields to guide particle propagation—during each simulation step, the next adversary state is sampled from the TAB-Field distribution. This ensures that simulated trajectories remain consistent with mission constraints and environmental limitations. The action selection process in TAB-POMCP remains unchanged, using Upper Confidence Tree (UCT) to balance exploration and exploitation. However, the value estimates now account for uncertainty in adversary behavior through the TAB-Field distributions rather than assumed transition models. When observations become available, particles are reweighted according to the observation likelihood, but unobserved adversary states continue to evolve according to the TAB-Fields. This approach maintains POMCP’s computational efficiency while enabling planning without explicit adversary policy assumptions.

## 5. Experiments and Results

We evaluate our approach through a series of experiments: hardware implementations with ground robots followed by ablation studies in simulation to evaluate performance across larger state spaces. *Motivating Scenario.* Consider a mission where an autonomous ego vehicle must intercept an adversarial agent targeting critical infrastructure. Through intelligence, our agent knows the adversary’s task which is defined by mission objectives and environmental constraints. However, the exact policy the adversary will use to execute this mission remains unknown. The agent can only observe the adversary’s position when it passes through monitored checkpoints, similar to security cameras providing visibility at key locations.

Note that this interception mission for the ego agent represents one instance of our framework. As described in Section 3, our approach maximizes a reward function capturing the ego agent’s objectives. While we focus on interception throughout our experiments as a concrete example, other missions like adversary avoidance or surveillance are equally applicable.

### 5.1. Experimental Setup

*Hardware platform.* We implement both the autonomous agent and the adversary using TurtleBot3 Burger platforms, each equipped with an onboard computer and a LDS-01 Lidar. The platforms run ROS2 with a custom navigation package (Puthumanailam et al., 2024c). Our experimental area includes markers providing precise localization at designated checkpoints. The environment includes obstacles creating restricted zones, while checkpoints are positioned to simulate critical areas which are monitored.

*Ego agent and adversary dynamics.* The agents operate under differential drive dynamics with state vector  $(x, y, \theta)$  representing position coordinates and heading angle. Control inputs are linear velocity  $v \in [0, 0.22 \text{ m/s}]$  and angular velocity  $\omega \in [0, 1.82 \text{ rad/s}]$ . A checkpoint-based observation model provides complete adversary state information only at designated locations, simulating security camera coverage at critical points.

*Adversary missions.* Following the formulation in Section 3, missions are specified in natural language and are encoded into constraint tuples  $\mathcal{M}$  defining goal states, temporal constraints, and additional requirements.

*Ego agent mission.* The ego agent aims to intercept the adversary before it reaches critical infrastructure. A reward of +50 is given for successful interception within 0.3m, while collisions incur a -30 penalty. A time step penalty of -1 encourages prompt action, and a control penalty of  $-0.1(v^2 + \omega^2)$  discourages abrupt movements.

## 5.2. Baselines

We evaluate TAB-conditioned POMCP against three baseline approaches representing different methods of handling adversary behavior uncertainty. (i) Standard POMCP (S-POMCP) ([Silver and Veness, 2010](#)) represents the original algorithm without mission awareness, where adversary transitions are modeled as uniform random movements within physical constraints – a common baseline that makes no assumptions about adversary behavior. (ii) Fixed-Policy POMCP (FP-POMCP) assumes the adversary follows a deterministic shortest-path policy to mission objectives, representing commonly used simplified models of goal-directed behavior. (iii) MLE-POMCP uses Maximum Likelihood Estimation to derive adversary transition probabilities from mission constraints and observed data, providing a data-driven comparison that attempts to learn adversary behavior patterns.

## 5.3. Results and Analysis

The performance of TAB-POMCP compared to the baseline methods is summarized in Table 1. TAB-POMCP consistently outperforms all baselines across all adversary mission types.

Adversary Mission Type	Metric	S-POMCP	FP-POMCP	MLE-POMCP	TAB-POMCP
<b>M1: Basic Reachability</b> Reach Checkpoint A within 5s	ATCR (%) (↓) StI (avg) (↓)	85.3% 1490	78.0% 1103	64.7% 852	13.3% 316
<b>M2: Sequential Objectives</b> Reach Checkpoint A and then Checkpoint B in exactly 5s	ATCR (%) (↓) StI (avg) (↓)	91.2% 1882	84.7% 1312	79.4% 1013	18.3% 380
<b>M3: Recurrent Objectives</b> Reach Checkpoint A every 5s	ATCR (%) (↓) StI (avg) (↓)	88.1% 1631	80.6% 1274	45.3% 953	19.8% 412
<b>M4: Restricted Operation Missions</b> Reach Checkpoint A while avoiding the central zone	ATCR (%) (↓) StI (avg) (↓)	82.0% 1445	74.6% 1102	59.8% 883	15.8% 297
<b>M5: Multi-Objective Missions</b> Mission Combination (Figure 3)	ATCR (%) (↓) StI (avg) (↓)	95.6% 2312	88.9% 1871	80.2% 1533	30.9% 545

Table 1: Performance comparison between TAB-POMCP and baseline methods on Adversary Task Completion Rate (ATCR) and Average Steps to Interception (StI). Results are averaged over 150 experiments per mission type. Example missions are provided below each category.

*Impact of conditioning policies on TAB-Fields.* The comparison between TAB-POMCP and S-POMCP (refer Table 1) clearly demonstrates the advantages of incorporating mission constraints into the planning process. S-POMCP, which does not utilize TAB-Fields, exhibits inefficient belief updates, particularly in periods of no observation. This often leads to overly dispersed belief distributions, resulting in ineffective tracking and search patterns. This inefficiency is reflected in consistently higher StI across missions, highlighting the method’s inability to effectively narrow down possible adversary states. In contrast, TAB-POMCP leverages mission constraints to focus

belief distributions on regions that align with the adversary’s objectives, even in the absence of observations. This enables more informed and targeted decision-making, leading to significantly higher interception rates. Figure 3 illustrates this behavior through representative trajectories: while S-POMCP exhibits aimless or overly cautious search patterns, TAB-POMCP efficiently prioritizes high-likelihood regions, demonstrating the impact of mission-aware reasoning.

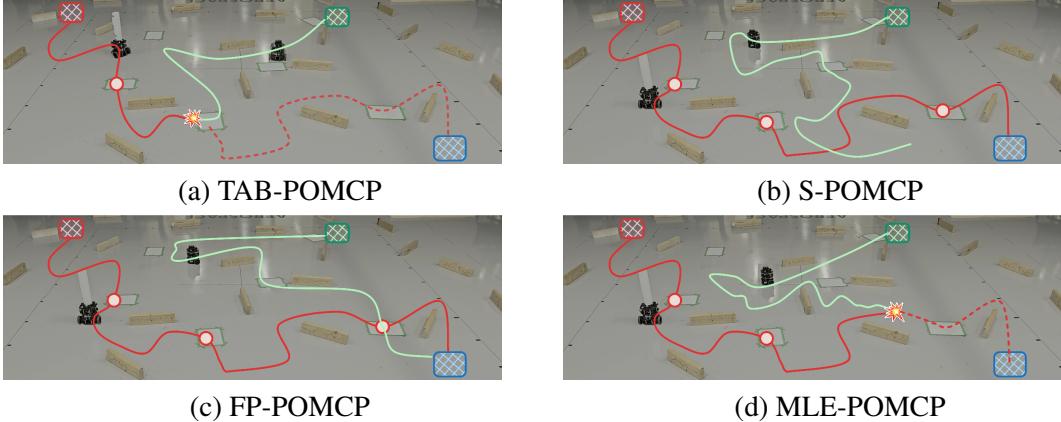


Figure 3: Comparison of agent (green) and adversary (red) trajectories followed by different approaches. Light red circles indicate full observability points at checkpoints, and  $\text{💥}$  marks the interception area. Adversary mission: Reach target [x,y] after visiting any three different checkpoints, taking no more than 10s between checkpoints, while avoiding the center of the environment.

*Comparison with alternative mission-aware approaches.* The results in Table 1 provide key insights into mission-aware planning. As expected, both FP-POMCP and MLE-POMCP outperform S-POMCP, highlighting the value of incorporating mission specifications into the planning process. However, their limitations are evident when examined closely. FP-POMCP assumes deterministic, shortest-path behavior for the adversary, which makes it highly brittle in scenarios where the adversary deviates from such paths. This limitation is clearly illustrated in Figure 3, where FP-POMCP struggles to adapt to behaviors that do not follow shortest path, leading to significant tracking inefficiencies. MLE-POMCP, on the other hand, demonstrates better flexibility by learning adversary behavior patterns from data. However, as shown in Figure 3, its reliance on sufficient past observations results in poor early-mission performance. The method only improves as it gathers enough data to refine its belief, leaving a critical gap during the initial stages of the mission. In contrast, TAB-POMCP enables robust performance across all phases of the mission. Unlike FP-POMCP, TAB-POMCP does not assume specific behavior patterns and is bounded by known constraints. Similarly, it avoids MLE-POMCP’s reliance on extensive behavioral data, allowing it to excel even in sparse-data scenarios.

*Scalability of TAB-conditioned planners.* We evaluate the scalability of TAB-conditioned planners through high-fidelity underwater vehicle simulations using the BlueROV2 model<sup>1</sup>. The environment simulates a subsea inspection scenario. The BlueROVs operate in three-dimensional space with state vector  $(x, y, z, \phi, \theta, \psi)$  and corresponding velocities. Similar to the ground robot experiments,

1. The simulation environment and vehicle dynamics are based on experimental data, available in our open-source repository [https://github.com/gokulp01/bluerov2\\_gym/](https://github.com/gokulp01/bluerov2_gym/)

the adversary is fully observable only when passing near underwater sensor networks (checkpoint), simulating acoustic or sonar detection zones.

Missions	S-POMCP	FP-POMCP	MLE-POMCP	TAB-POMCP
M1	90.1%	83.5%	70.2%	19.1%
M2	94.8%	88.3%	82.7%	24.4%
M3	91.7%	85.2%	50.8%	27.1%
M4	87.5%	80.1%	65.3%	21.7%
M5	97.2%	91.5%	85.6%	43.4%

Table 2: Performance comparison between different methods on ATCR across different mission categories in an underwater setting. Mission types are the same as that in Table 1 and are abbreviated as M1 through M5.

As shown in Table 2, TAB-POMCP maintains its performance advantage over baselines across all the five mission types (M1-M5). The simulation results reveal critical insights about scaling TAB-conditioned policies to higher-dimensional spaces. First, the performance gap between TAB-POMCP and baselines widens as mission complexity increases, particularly in missions with complex temporal dependencies like M5. This suggests that the maximum entropy formulation becomes more valuable precisely when the search space expands. Second, even in the most complex scenarios with multiple interacting constraints (M3), TAB-POMCP maintains a 3-4x improvement in interception efficiency over methods that make explicit policy assumptions. The key driver behind this scalability is TAB-Fields’ ability to automatically identify and exploit mission-constrained regions of the state space. Rather than maintaining beliefs over the full 6-DOF state space, TAB-POMCP effectively “collapses” the belief to high-probability regions defined by mission constraints. This implicit dimensionality reduction enables efficient planning even as the raw state space grows.

*Limitations.* Despite the performance benefits, TAB-Field generation incurs additional computational overhead. With efficient parallelized implementation, TAB-POMCP requires approximately 1.4x more computation time compared to standard POMCP. Additionally, while our current formulation handles static obstacles, it does not yet account for dynamic obstacles.

## 6. Conclusion

We presented Task-Aware Behavior Fields (TAB-Fields), a novel approach to reason about adversary behavior in scenarios where mission objectives are known but specific policies remain unknown. Our key contribution lies in recognizing that the maximum entropy principle can characterize the full space of possible adversary behaviors using just mission specifications and environmental constraints, eliminating the need for policy assumptions or hand-crafted rewards. By solving a constrained optimization problem that minimizes bias beyond known constraints, TAB-Fields provide a distribution over adversary states that captures all feasible behaviors consistent with mission objectives. When integrated with standard planning algorithms through TAB-conditioned POMCP, this representation enables effective decision-making in complex adversarial scenarios. Our experimental results demonstrate significant performance improvements over methods that either make specific policy assumptions or ignore mission constraints.

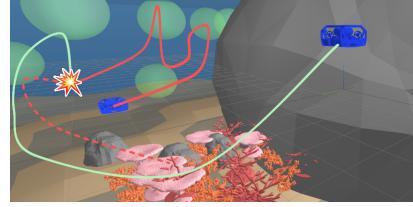


Figure 4: Agent (green) and adversary (red) trajectories using TAB-POMCP. Teal bubbles indicate checkpoints. Adversary task: Reach corals after visiting checkpoints 1, 2, 3 in order, taking no more than 30s between checkpoints.

## Acknowledgments

The work of Ornik, Puthumaillam and Song was supported by the Office of Naval Research under grants N00014-23-1-2651 and N00014-23-1-2505. The work of Park and Yesmagambet was supported by funding from King Abdullah University of Science and Technology (KAUST).

## References

- Pieter Abbeel and Andrew Y Ng. Exploration and apprenticeship learning in reinforcement learning. In *International Conference on Machine Learning*, 2005.
- Mahmoud Abouelyazid. Adversarial deep reinforcement learning to mitigate sensor and communication attacks for secure swarm robotics. *Journal of Intelligent Connectivity and Emerging Technologies*, 2023.
- Stephen Adams, Tyler Cody, and Peter A Beling. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, pages 4307–4346, 2022.
- Navid Aghasadeghi and Timothy Bretl. Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.
- Noa Agmon. Robotic strategic behavior in adversarial environments. In *International Joint Conference on Artificial Intelligence*, 2017.
- Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 2021.
- Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Neural Information Processing Systems*, 2008.
- Alberto Castellini, Enrico Marchesini, and Alessandro Farinelli. Partially observable Monte Carlo planning with state variable constraints for mobile robot navigation. *Engineering Applications of Artificial Intelligence*, 2021.
- Taolue Chen and Tingting Han. On the complexity of computing maximum entropy for Markovian models. *Leibniz International Proceedings in Informatics*, 2014.
- Wang Chi Cheung. Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *arXiv preprint arXiv:1905.06466*, 2019.
- Aleksander Czechowski and Frans A Oliehoek. Decentralized mcts via learned teammate models. In *International Conference on International Joint Conferences on Artificial Intelligence*, 2021.
- Maxim Egorov, Mykel Kochenderfer, and Jaak Uudmae. Target surveillance in adversarial environments using pomdps. In *AAAI Conference on Artificial Intelligence*, 2016.
- Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

S Gronauer and K Diepold. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 2022.

Weiqing Gu, Ranjeev Mittu, Julie Marble, Gavin Taylor, Ciara Sibley, Joseph Coyne, and William F Lawless. Towards modeling the behavior of autonomous systems and humans for trusted operations. In *AAAI Spring Symposium Series*, 2014.

Yanlin Han and Piotr Gmytrasiewicz. Learning others' intentional models in multi-agent settings using interactive POMDPs. In *Neural Information Processing Systems*, 2018.

Yanlin Han and Piotr Gmytrasiewicz. IPOMDP-net: A deep neural network for partially observable multi-agent planning using interactive POMDPs. In *AAAI Conference on Artificial Intelligence*, 2019.

Li Huang, MengChu Zhou, Kuangrong Hao, and Edwin Hou. A survey of multi-robot regular and adversarial patrolling. *IEEE/CAA Journal of Automatica Sinica*, 2019.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 2005.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 1957.

Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 1982.

Zhanming Jie, Aldrian Muis, and Wei Lu. Efficient dependency-guided named entity recognition. In *AAAI Conference on Artificial Intelligence*, 2017.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998.

Daphane Koller. Probabilistic Graphical Models: Principles and Techniques, 2009.

Raphael Korbacher and Antoine Tordeux. Review of pedestrian trajectory prediction methods: Comparing deep learning and knowledge-based approaches. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

Mikko Lauri, David Hsu, and Joni Pajarinen. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 2022.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Simone Di Marino and Augusto Gerolin. An optimal transport approach for the Schrödinger" odinger bridge problem and convergence of Sinkhorn algorithm. *Journal of Scientific Computing*, 2020.

Brenda Ng, Carol Meyers, Kofi Boakye, and John Nitao. Towards applying interactive pomdps to real-world adversary modeling. In *AAAI Conference on Artificial Intelligence*, 2010.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 2005.

- Praveen Paruchuri. *Reasoning in Uncertain Adversarial Environments in Agent/Multiagent Systems*. PhD thesis, Ph.D. Dissertation Proposal, 2007.
- Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.
- Gokul Puthumanailam, Xiangyu Liu, Negar Mehr, and Melkior Ornik. Weathering ongoing uncertainty: Learning and planning in a time-varying partially observable environment. In *IEEE International Conference on Robotics and Automation*, 2024a.
- Gokul Puthumanailam, Paulo Padrao, Jose Fuentes, Leonardo Bobadilla, and Melkior Ornik. Enhancing robot navigation policies with task-specific uncertainty management. *arXiv preprint arXiv:2410.15178*, 2024b.
- Gokul Puthumanailam, Manav Vora, and Melkior Ornik. ComTraQ-MPC: Meta-trained DQN-MPC integration for trajectory tracking with limited active localization updates. *arXiv preprint arXiv:2403.01564*, 2024c.
- Eugene Santos Jr and Qunhua Zhao. Adversarial models for opponent intent inferencing. *Adversarial Reasoning: Computational Approaches to Reading the Opponents Mind*, 2006.
- Eugene Santos Jr, Bruce McQueary, and Lee Krause. Modeling adversarial intent for interactive simulation and gaming: the fused intent system. In *Modeling and Simulation for Military Operations*, 2008.
- Yagiz Savas, Melkior Ornik, Murat Cubuktepe, and Ufuk Topcu. Entropy maximization for constrained Markov decision processes. In *Allerton Conference on Communication, Control, and Computing*, 2018.
- Yagiz Savas, Melkior Ornik, Murat Cubuktepe, Mustafa O Karabag, and Ufuk Topcu. Entropy maximization for Markov decision processes under temporal logic constraints. *IEEE Transactions on Automatic Control*, 2019.
- Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 2013.
- David Silver and Joel Veness. Monte-carlo planning in large POMDPs. In *Neural Information Processing Systems*, 2010.
- Adhiraj Soman, Nan Ye, David Hsu, and Wee Sun Lee. Despot: Online POMDP planning with regularization. In *Neural Information Processing Systems*, 2013.
- MTCAJ Thomas and A Thomas Joy. *Elements of Information Theory*. Wiley-Interscience, 2006.
- Sebastian Thrun. Monte carlo pomdps. In *Neural Information Processing Systems*, 1999.
- Emanuel Todorov. Efficient computation of optimal actions. *National Academy of Sciences*, 2009.
- Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 2013.

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Robust Markov Decision Processes without model estimation. *arXiv preprint arXiv:2302.01248*, 2023.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Neural Information Processing Systems*, 2020.

Lifeng Zhou and Pratap Tokekar. Multi-robot coordination and planning in uncertain and adversarial environments. *Current Robotics Reports*, 2021.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.

Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

Inon Zuckerman, Sarit Kraus, and Jeffrey S Rosenschein. The adversarial activity model for bounded rational agents. *Autonomous Agents and Multi-Agent Systems*, 2012.