Robust Permissive Controller Synthesis for Interval MDPs

Khang Vo Huynh[1], David Parker[2], and Lu Feng[1]

***Abstract*— We address the problem of robust permissive controller synthesis for robots operating under uncertain dynamics, modeled as Interval Markov Decision Processes (IMDPs). IMDPs generalize standard MDPs by allowing transition probabilities to vary within intervals, capturing epistemic uncertainty from sensing noise, actuation imprecision, and coarse system abstractions—common in robotics. Traditional controller synthesis typically yields a single deterministic strategy, limiting adaptability. In contrast, permissive controllers (multi-strategies) allow multiple actions per state, enabling runtime flexibility and resilience. However, prior work on permissive controller synthesis generally assumes exact transition probabilities, which is unrealistic in many robotic applications. We present the first framework for robust permissive controller synthesis on IMDPs, guaranteeing that all strategies compliant with the synthesized multi-strategy satisfy reachability or reward-based specifications under all admissible transitions. We formulate the problem as mixed-integer linear programs (MILPs) and propose two encodings: a baseline vertex-enumeration method and a scalable dualization-based method that avoids explicit enumeration. Experiments on four benchmark domains show that both methods synthesize robust, maximally permissive controllers and scale to large IMDPs with up to hundreds of thousands of states.**

I. Introduction

Decision-making under uncertainty is a central challenge in robotics [1], [2]. Robots operate in environments where sensing is noisy, actuation is imprecise, and system abstractions are inevitably coarse. Standard Markov decision processes (MDPs) are widely used to model sequential decision-making, but they assume exact transition probabilities, which are rarely available in practice. Recent surveys highlight the broad landscape of uncertainty models [3], [4].

Among these models, *Interval MDPs (IMDPs)* play a critical role. By allowing transition probabilities to vary within intervals, IMDPs capture epistemic uncertainty while retaining a tractable structure for analysis and synthesis. Robust controller synthesis for IMDPs has been studied extensively, with applications ranging from robotic planning and verification to hybrid and neural network dynamics [5], [6], [7], [8]. However, existing methods typically return a single policy—often deterministic or randomized—but not permissive, which can be overly restrictive in practice.

In contrast, *permissive controllers*, also known as multi-strategies, allow multiple actions to remain enabled in each state. Such controllers preserve flexibility at runtime: a robot may adaptively select among admissible actions based on external conditions, unforeseen disturbances, or higher-level objectives, while still maintaining correctness. Permissive synthesis has been studied for MDPs and probabilistic games with exact transitions, often via mixed-integer linear programming (MILP) encodings that maximize permissiveness subject to property satisfaction [9], [10], [11], [12]. Yet, existing permissive methods assume precise probability models and do not account for epistemic uncertainty as captured by IMDPs.

The lack of robust permissive synthesis methods creates a significant gap for robotics applications. Consider a mobile robot navigating in an uncertain environment. A deterministic controller may commit to a single action sequence that is safe under all transition realizations, but in doing so, it may forgo many other safe alternatives. A permissive controller, in contrast, enables all actions consistent with safety, allowing adaptation at runtime. To be useful in robotics, such permissive strategies must be robust to transition uncertainty, guaranteeing correctness under all admissible distributions of the IMDP.

Another way to interpret permissive controllers is as *shields* for safe reinforcement learning (RL) [12], [13], which filter unsafe actions online while leaving other choices unrestricted. A permissive multi-strategy serves precisely this role, specifying admissible actions in each state and blocking the rest. Prior permissive synthesis methods thus align with shields in known MDPs, while robust IMDP synthesis aligns with worst-case reasoning under uncertainty. Bridging these ideas, robust permissive synthesis over IMDPs naturally yields robust shields: multi-strategies that guarantee safety against all admissible transitions while minimally restricting learning or heuristic policies. This view is particularly relevant for emerging paradigms such as meta-RL [14] and in-context RL [15], where agents must adapt across distributions of MDPs. Both are attracting interest in robotics, but ensuring safety under distributional variability remains challenging; robust permissive shields offer a principled way to address this while preserving flexibility for adaptation.

In this paper, we present the *first framework for robust permissive controller synthesis on IMDPs*. We focus on reachability and reward-based specifications, which are widely used in robotics. Our approach guarantees that every strategy compliant with the synthesized multi-strategy satisfies the specification under all admissible transitions. We formulate the synthesis problem as mixed-integer linear programs and propose two alternative encodings:

1) **Vertex enumeration encoding**, a direct formulation that is conceptually simple but introduces exponentially

[1]Khang Vo Huynh and Lu Feng are with the Department of Computer Science, University of Virginia, Charlottesville, VA 22904, USA {szq2sj, lu.feng}@virginia.edu

[2]David Parker is with the Department of Computer Science, University of Oxford, Parks Road, Oxford OX1 3QD, United Kingdom david.parker@cs.ox.ac.uk

many constraints in the number of successors; and
2) **Dualization-based encoding**, which avoids explicit enumeration by leveraging robust optimization, yielding a compact formulation that scales linearly with the number of successors.

We implement both encodings in a prototype tool and evaluate them on four representative benchmark domains: aircraft collision avoidance, semi-autonomous vehicle control, obstacle navigation, and warehouse navigation. Our experiments show that both methods synthesize robust, maximally permissive controllers, scaling to IMDPs with hundreds of thousands of states.

By bridging permissive multi-strategy synthesis with robust IMDP reasoning, our work advances the state of the art in uncertainty-aware decision-making for robotics. It provides both theoretical foundations and practical algorithms for synthesizing controllers that are robust to model uncertainty while retaining runtime flexibility.

II. Related Work

IMDPs in Robotics. Interval MDPs have been widely applied as abstractions in robotic planning and control under uncertainty. Recent work has used IMDPs to model dynamics with learned uncertainty for navigation and control tasks, including legged locomotion [5], safe motion planning [6], and switched stochastic systems [16]. IMDP abstractions have also been constructed from perception pipelines, yielding conservative bounds with probabilistic guarantees [17]. Other works demonstrate the use of IMDPs in stochastic hybrid systems [7], neural-network dynamics [8], and compositional abstractions for scalability [18]. Data-driven approaches extend this further, with IMDP abstractions learned from samples and equipped with PAC guarantees [19], [20]. These studies highlight the versatility of IMDP abstractions for robotics, though they primarily focus on model construction and verification rather than permissive controller synthesis.

IMDP Controller Synthesis. Robust controller synthesis for IMDPs has been extensively studied, with algorithms addressing temporal-logic objectives, reachability, and multi-objective properties. Classical results established robust dynamic programming under interval uncertainty [21], [22], while subsequent work developed robust synthesis for probabilistic temporal logics [23], [24]. Recent research has focused on scalability, richer objectives and learning, including multi-objective synthesis [25], PAC-based learning of IMDPs [26] and efficient PAC-bounded abstractions for dynamical systems [27]. Compositional and structural approaches enable scalable synthesis for high-dimensional systems [18], and tool support continues to advance [28], [29], [30]. While these works solve robust synthesis for IMDPs under various objectives, they all yield deterministic or randomized strategies, limiting runtime flexibility.

Permissive Controller Synthesis. Permissive controller synthesis was introduced to preserve flexibility by allowing multiple actions per state while still satisfying specifications. Initial work encoded permissive multi-strategy synthesis as a MILP with penalties for disabled actions [9]. Subsequent studies extended this idea to multi-objective settings, synthesizing permissive controllers that accommodate uncertain human preferences [11], and to robust planning in models with set-valued transitions under temporal-logic objectives [31]. Beyond formal synthesis, permissive controllers are closely related to *shields* in safe reinforcement learning, which block unsafe actions online while leaving other choices unrestricted; in this setting, a permissive multi-strategy naturally serves as a shield that enforces safety with minimal interference [12], [13]. While these approaches demonstrate the utility of permissive strategies, they assume exact or set-valued probabilities. To the best of our knowledge, no prior work addresses robust permissive synthesis for IMDPs under interval uncertainty, which is the gap filled by this paper.

III. Problem Formulation

Interval MDPs (IMDPs). Formally, an IMDP is a tuple $\mathcal{M} = (S, s_\iota, A, \check{P}, \hat{P}, R)$, where S is a finite set of states with initial state $s_\iota \in S$, A is a finite set of actions, $\check{P}, \hat{P} : S \times A \times S \to [0,1]$ assign lower and upper bounds to each transition, and $R : S \times A \to \mathbb{R}_{\geq 0}$ is a reward function. Intuitively, an IMDP represents a family of MDPs. For each state-action pair (s,a), the admissible successor distributions are $\mathcal{P}(s,a) = \{\, P \in \Delta(S) \mid \check{P}(s,a,s') \leq P(s,a,s') \leq \hat{P}(s,a,s') \;\; \forall s' \in S \,\}$. We define $\mathcal{P}$ as the set of all transition functions $P : S \times A \to \Delta(S)$ with $P(s,a,\cdot) \in \mathcal{P}(s,a)$ for every (s,a).

Permissive Controller. An IMDP is controlled by a *strategy* (also called a *policy*), which in this work is assumed to be deterministic and memoryless. Formally, a strategy is a function $\sigma : S \to A$ such that $\sigma(s) \in \alpha(s)$, where $\alpha(s) \subseteq A$ denotes the set of *enabled actions* in state s. A *permissive controller*, or *multi-strategy*, generalizes this notion by allowing a set of actions at each state; formally, it is a mapping $\theta : S \to 2^A$ such that $\theta(s) \subseteq \alpha(s)$ and $\theta(s) \neq \emptyset$ for all $s \in S$. For a multi-strategy θ, define its *permissiveness* as $\beta(\theta) \coloneqq \sum_{s \in S} \sum_{a \in \alpha(s)} \mathbf{1}[a \in \theta(s)]$. A strategy σ is said to be *compliant* with a multi-strategy θ, denoted $\sigma \triangleleft \theta$, if $\sigma(s) \in \theta(s)$ for every $s \in S$. For two multi-strategies θ and θ', we write $\theta \sqsubset \theta'$ if $\theta(s) \subset \theta'(s)$ for all $s \in S$.

Specifications. We consider two common types of properties. A *probabilistic reachability* property has the form $\mathsf{P}_{\bowtie p}(\Diamond T)$, where $T \subseteq S$ is a target set, $\Diamond$ denotes eventual reachability, $p \in [0,1]$, and $\bowtie \in \{\leq, \geq\}$; it requires that the probability of eventually reaching T satisfies the bound $\bowtie p$. An *expected total reward* property has the form $\mathsf{R}_{\bowtie b}(\Diamond T)$, where $b \in \mathbb{R}_{\geq 0}$, meaning the expected total reward accumulated until reaching T satisfies the bound $\bowtie b$. More complex temporal specifications, such as those in PCTL or LTL, can often be reduced to these reachability properties.

Satisfaction Semantics. Given an IMDP $\mathcal{M}$, a property φ, a strategy σ, and a transition function $P \in \mathcal{P}$, we write $(\mathcal{M}, \sigma, P) \models \varphi$ if φ holds in the MDP obtained by resolving

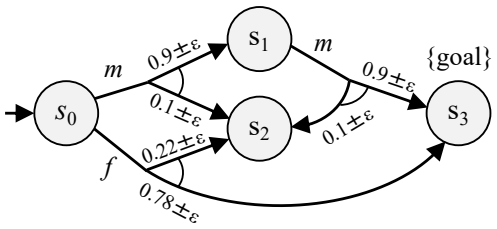

Fig. 1. A simple robotic navigation task modeled as an IMDP.

all choices in $\mathcal{M}$ using strategy σ and transition function P. For a multi-strategy θ, $(\mathcal{M}, \theta, P) \models \varphi$ if this holds for all $\sigma \triangleleft \theta$. Robust satisfaction then quantifies over all admissible transitions: $(\mathcal{M}, \sigma) \models_{\text{rob}} \varphi$ if $(\mathcal{M}, \sigma, P) \models \varphi$ for all $P \in \mathcal{P}$, and $(\mathcal{M}, \theta) \models_{\text{rob}} \varphi$ if $(\mathcal{M}, \theta, P) \models \varphi$ for all $P \in \mathcal{P}$.

Problem Statement. We aim to synthesize a *robust permissive controller* that guarantees satisfaction of the specification under all admissible uncertainties, while retaining as many action choices as possible.

Problem: Robust Permissive Controller Synthesis

Given an IMDP $\mathcal{M} = (S, s_\iota, A, \check{P}, \hat{P}, R)$ and a specification φ of the form $\mathsf{P}_{\bowtie p}(\Diamond T)$ or $\mathsf{R}_{\bowtie b}(\Diamond T)$, synthesize a multi-strategy θ such that:

1. (**Robust satisfaction**) $(\mathcal{M}, \theta) \models_{\text{rob}} \varphi$.
2. (**Maximal permissiveness**) There exists no multi-strategy θ' with $\theta \sqsubset \theta'$ such that $(\mathcal{M}, \theta') \models_{\text{rob}} \varphi$.

Example 1. Consider a robotic navigation task modeled as the IMDP $\mathcal{M}$ in Figure 1. A robot starts in s_0 and must reach the goal state s_3. From s_0, it may choose a *fast* action (f), which can reach s_3 directly but with risk of failure into s_2, or a *medium* action (m), which is safer but must be applied twice through s_1. Transition probabilities are specified as intervals of radius $\varepsilon \in [0, 1]$, where ε is a fixed constant capturing the level of epistemic uncertainty. For example, $P(s_0, f, s_3) \in [0.78 - \varepsilon, 0.78 + \varepsilon]$, clipped to $[0, 1]$. A property such as $\varphi = \mathsf{P}_{\geq 0.65}(\Diamond s_3)$ asks whether the probability of eventually reaching the goal is at least 0.65. The synthesis problem then asks for a multi-strategy θ such that $(\mathcal{M}, \theta) \models_{\text{rob}} \varphi$ while being maximally permissive. ■

IV. Methods

We now present methods for solving the IMDP robust permissive controller synthesis problem. A straightforward approach is to explicitly enumerate all vertices of the uncertainty polytopes in the IMDP and adopt existing permissive controller synthesis techniques for MDPs [9]. This yields a mixed-integer linear program (MILP) encoding, but the number of vertices grows exponentially with the number of successor states per state–action pair. To address this scalability issue, we develop a dualization-based encoding that captures the adversarial choice of transition probabilities without explicit enumeration, yielding a compact MILP. We present the vertex enumeration encoding in Section IV-A and the dualization-based encoding in Section IV-B.

Encoding 1: Vertex Enumeration for $\mathsf{P}_{\geq p}(\Diamond T)$

$$\text{maximize} \sum_{s \in S} \sum_{a \in \alpha(s)} y_{s,a} \tag{1a}$$

subject to:

$$\forall s \in S: \sum_{a \in \alpha(s)} y_{s,a} \geq 1 \tag{1b}$$

$$x_{s_\iota} \geq p \tag{1c}$$

$$\forall s \in T: \; x_s = 1 \tag{1d}$$

$$\forall s \in S \setminus T, \; \forall a \in \alpha(s), \; \forall v \in \mathcal{V}(s, a): \; x_s \leq \sum_{s' \in S} P^{(v)}(s, a, s') \cdot x_{s'} + M(1 - y_{s,a}) \tag{1e}$$

A. Vertex Enumeration Encoding

We present an MILP for the probabilistic reachability property $\mathsf{P}{\geq}\, p(\Diamond T)$, given in Encoding 1. The formulation introduces real-valued variables $x_s \in [0, 1]$, denoting the minimal probability of reaching T from state s, and binary variables $y_{s,a}$, indicating whether the multi-strategy θ admits action $a \in \alpha(s)$.

The objective function (1a) maximizes the number of admitted actions, thereby ensuring maximal permissiveness. Constraint (1b) ensures that all states with enabled actions admit at least one of them. Constraint (1c) enforces that the initial state achieves the threshold p. Constraint (1d) fixes the value of all target states to 1.

Constraint (1e) enforces robustness by requiring x_s to hold for all feasible successor distributions of each non-target pair (s, a). The term $M(1-y_{s,a})$ deactivates the constraint when action a is not allowed, where M is a sufficiently large constant (the standard big-M method) used for linearization and constraint deactivation. The vertex set $\mathcal{V}(s, a)$ consists of the extreme points of $\mathcal{P}(s, a)$, obtained by pushing transition probabilities to their interval bounds while maintaining that they remain nonnegative and sum to one. Each $v \in \mathcal{V}(s, a)$ defines a concrete distribution $P^{(v)}(s, a, \cdot)$, and enforcing the inequality for all v ensures that x_s is robust.

Example 2. We apply Encoding 1 to the IMDP in Figure 1 with respect to the property $\varphi = \mathsf{P}_{\geq 0.65}(\Diamond s_3)$. Variables x_s record the minimal reachability probability from each state, while binary variables $y_{s_0,f}$, $y_{s_0,m}$, and $y_{s_1,m}$ specify which actions (fast or medium) are allowed by the multi-strategy in states s_0 and s_1. The objective is to maximize the sum $y_{s_0,f} + y_{s_0,m} + y_{s_1,m}$. Constraints (1b)–(1d) enforce that at least one action is chosen in s_0 and s_1, that $x_{s_0} \geq 0.65$, and that $x_{s_3} = 1$. Constraint (1e) is instantiated for each vertex of the uncertainty polytope. For example, for the fast action at s_0, $\mathcal{P}(s_0, f)$ reduces to an interval over successors

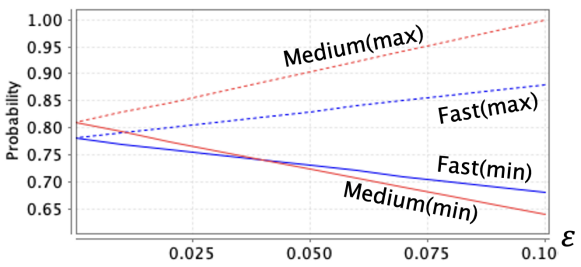

Fig. 2. Robust reachability probability as the uncertainty radius ε varies.

$\{s_2, s_3\}$ and yields two inequalities:

$$x_{s_0} \le (0.22+\varepsilon)\, x_{s_2} + (0.78-\varepsilon)\, x_{s_3} + M(1-y_{s_0,f}),$$
$$x_{s_0} \le (0.22-\varepsilon)\, x_{s_2} + (0.78+\varepsilon)\, x_{s_3} + M(1-y_{s_0,f}).$$

Solving the MILP shows how the resulting multi-strategy depends on the uncertainty radius ε and the threshold p. For $\varepsilon = 0.1$, requiring $\mathsf{P}_{\ge 0.65}(\lozenge s_3)$ admits only the fast action. Lowering the threshold to $p = 0.6$ or reducing uncertainty to $\varepsilon = 0.05$ admits both fast and medium actions. These cases are illustrated in Figure 2, which plots the minimal (dashed) and maximal (solid) reachability probabilities for each action as the uncertainty radius ε varies. ■

Correctness. The correctness of the MILP in Encoding 1 is stated in the following theorem, with the proof provided in the appendix.

Theorem 1: Let $\mathcal{M}$ be an IMDP and $\varphi = \mathsf{P}_{\ge p}(\lozenge T)$ a property. There exists a robust, maximally permissive multi-strategy θ in $\mathcal{M}$ for φ if and only if there exists an optimal assignment to the MILP in Encoding 1 whose objective value equals the permissiveness of θ, i.e., $\beta(\theta) = \sum_{s\in S}\sum_{a\in\alpha(s)} y_{s,a}$.

Adapting Encoding 1 for Other Specifications. For properties of the form $\mathsf{P}_{\le p}(\lozenge T)$, constraint (1c) is replaced with $x_{s_\iota} \le p$, and constraint (1e) becomes

$$x_s \ge \sum_{s'\in S} P^{(v)}(s,a,s')\cdot x_{s'} - M(1-y_{s,a}).$$

For expected total reward properties $\mathsf{R}_{\ge b}(\lozenge T)$, we instead use $x_s \in \mathbb{R}_{\ge 0}$ to denote reward-to-go variables, modify constraint (1c) to $x_{s_\iota} \ge b$, set constraint (1d) to $x_s = 0$, and replace constraint (1e) with

$$x_s \le \sum_{s'\in S} P^{(v)}(s,a,s')\cdot x_{s'} + r(s,a) + M(1-y_{s,a}),$$

where $r(s,a)$ is the immediate reward of action a in state s. Analogous changes apply to properties of the form $\mathsf{R}_{\le b}(\lozenge T)$ by reversing the inequalities accordingly. We also impose constraints on set T being reached with probability 1.

Scalability. Although conceptually straightforward, the vertex enumeration encoding can scale poorly. The number of vertices $|\mathcal{V}(s,a)|$ of each polytope $\mathcal{P}(s,a)$ can be exponential in the number of successor states, and constraint (1e) must be instantiated for every vertex. As a result, the MILP may contain exponentially many constraints in the size of the IMDP, making this encoding impractical for larger models and motivating the dualization-based approach as follows.

Encoding 2: Dualization for $\mathsf{P}_{\ge p}(\lozenge T)$

$$\text{maximize} \sum_{s\in S}\sum_{a\in\alpha(s)} y_{s,a} \tag{2a}$$
subject to:
$$\forall s\in S: \sum_{a\in\alpha(s)} y_{s,a} \ge 1 \tag{2b}$$
$$x_{s_\iota} \ge p \tag{2c}$$
$$\forall s\in T:\ x_s = 1 \tag{2d}$$
$$\forall (s,a,s'):\ \lambda_{s,a} - \hat{u}_{s,a,s'} + \check{u}_{s,a,s'} \le x_{s'} \tag{2e}$$
$$\forall (s,a):\ x_s \le M(1-y_{s,a}) + \eta_{s,a}$$
$$+ \sum_{s'\in S}\left(\check{P}(s,a,s')\,\check{u}_{s,a,s'} - \hat{P}(s,a,s')\,\hat{u}_{s,a,s'}\right) \tag{2f}$$
$$\forall (s,a):\ -M\, y_{s,a} \le \eta_{s,a} \le M\, y_{s,a} \tag{2g}$$
$$\forall (s,a):\ \lambda_{s,a} - M(1-y_{s,a}) \le \eta_{s,a}$$
$$\le \lambda_{s,a} + M(1-y_{s,a}) \tag{2h}$$

B. Dualization-Based Encoding

We now present a compact MILP encoding based on linear programming duality as shown in Encoding 2. The objective (2a) and constraints (2b)–(2d) are identical to those in Encoding 1.

Instead of enumerating all vertices of the uncertainty polytopes, we encode the worst-case choice of transition probabilities directly using dual variables, a standard approach in robust optimization [32]. Here, $\hat{u}_{s,a,s'}, \check{u}_{s,a,s'} \ge 0$ are dual variables associated with the upper and lower bounds $\hat{P}(s,a,s')$ and $\check{P}(s,a,s')$, $\lambda_{s,a}$ is a free dual variable, and $\eta_{s,a}$ is an auxiliary variable used to linearize the product $\lambda_{s,a} y_{s,a}$ with the big-M method [33].

Constraint (2e) enforces *dual feasibility*: for each state–action pair (s,a) and successor s', the dual variables satisfy $\lambda_{s,a} - \hat{u}_{s,a,s'} + \check{u}_{s,a,s'} \le x_{s'}$. Constraint (2f) is the core robust inequality, bounding the value x_s by the dual objective. Intuitively, it ensures that the worst-case distribution consistent with the intervals is captured without explicit vertex enumeration. Finally, constraints (2g)–(2h) implement the big-M linearization, guaranteeing $\eta_{s,a} = 0$ when $y_{s,a} = 0$ and $\eta_{s,a} = \lambda_{s,a}$ when $y_{s,a} = 1$.

Example 3. We revisit the IMDP and property from Example 2, now using Encoding 2. For the fast action at s_0, constraint (2e) introduces dual feasibility inequalities linking $\lambda_{s_0,f}, \hat{u}_{s_0,f,s'}, \check{u}_{s_0,f,s'}$ to the successor values $x_{s'}$. Constraint (2f) then yields the robust bound

$$x_{s_0} \le M(1-y_{s_0,f}) + \eta_{s_0,f}$$
$$+ (0.22-\varepsilon)\,\check{u}_{s_0,f,s_2} - (0.22+\varepsilon)\,\hat{u}_{s_0,f,s_2}$$
$$+ (0.78-\varepsilon)\,\check{u}_{s_0,f,s_3} - (0.78+\varepsilon)\,\hat{u}_{s_0,f,s_3}.$$

Here the auxiliary variable $\eta_{s_0,f}$ linearizes the product $\lambda_{s_0,f} y_{s_0,f}$, ensuring that $\lambda_{s_0,f}$ contributes only when $y_{s_0,f} = 1$. Solving this encoding yields the same multi-strategy outcomes as shown in Figure 2. ■

Correctness. The correctness of the MILP in Encoding 2 is stated in the following theorem, with the proof provided in the appendix.

Theorem 2: Let $\mathcal{M}$ be an IMDP and $\varphi = \mathsf{P}_{\geq p}(\lozenge T)$ a property. There exists a robust, maximally permissive multi-strategy θ in $\mathcal{M}$ for φ if and only if there exists an optimal assignment to the MILP in Encoding 2 whose objective value equals the permissiveness of θ, i.e., $\beta(\theta) = \sum_{s \in S} \sum_{a \in \alpha(s)} y_{s,a}$.

Adapting Encoding 2 for Other Specifications. For $\mathsf{P}_{\leq p}(\lozenge T)$, constraint (2c) is replaced by $x_{s_\iota} \leq p$, and constraint (2f) becomes

$$x_s \geq -M(1 - y_{s,a}) + \eta_{s,a} + \sum_{s' \in S} \big(\check{P}(s,a,s')\, \check{u}_{s,a,s'} - \hat{P}(s,a,s')\, \hat{u}_{s,a,s'} \big).$$

For reward properties $\mathsf{R}_{\bowtie b}(\lozenge T)$, the variables x_s represent reward-to-go, constraint (2c) is replaced by $x_{s_\iota} \bowtie b$, constraint (2d) is set to $x_s = 0$ for all $s \in T$, and the immediate reward $r(s,a)$ is added in constraint (2f). In both cases, all other constraints remain unchanged.

Scalability. The dualization-based encoding may introduce more constraints in small instances such as the IMDP in Figure 1. However, while vertex enumeration grows exponentially with the number of successor states, dualization requires only a linear number. This advantage becomes evident in larger IMDPs, as demonstrated in Section V.

V. Experiments

We have developed a prototype implementation of the proposed synthesis methods. The implementation leverages the PRISM model checker [28] to parse IMDP models and specifications, and the Gurobi Optimizer [34] to solve the resulting MILP encodings. All experiments were conducted on a laptop equipped with a 2.8 GHz Quad-Core Intel Core i7 processor and 16 GB of RAM.

A. Benchmark Domains

We evaluated our methods on four representative IMDP benchmark domains, each capturing a different robotic decision-making scenario under uncertainty. In the first three domains, the specification is a probabilistic reachability property of the form $P_{\geq p}(\lozenge T)$, requiring that the target set T is reached with probability at least p under all admissible transitions. The final domain considers a reward-based specification of the form $R_{\leq b}(\lozenge T)$.

Aircraft Collision Avoidance (ACA). Adapted from [35], this domain models a controlled aircraft navigating a discretized airspace with an adversarial aircraft. The objective is to reach a target region while avoiding collisions with probability at least p under all admissible uncertainties. We consider three IMDP instances, where the parameter sets the maximum number of successor states per state–action pair, controlling the branching factor.

Semi-Autonomous Vehicle (SAV). Adapted from [10], this domain models a ground vehicle navigating a grid world while maintaining intermittent communication with a controller. Communication occurs via two lossy channels, with message-loss probabilities depending on the relay location. At each step, the vehicle may either attempt communication (with limited retries) or move to an adjacent cell. The objective is to reach the opposite corner with probability at least p; failure occurs if the vehicle moves too far without successful intermediate communication. The model parameter controls the size of the grid world.

Obstacle Navigation (OBS). Inspired by the FrozenLake problem [36], this domain models an agent moving through a stochastic grid world containing frozen cells that permanently trap the agent. The agent must reach a goal region with probability at least p. The model parameters control both the grid dimension and the number of steps required for a move, thereby varying the granularity of the state space.

Warehouse Navigation (WH). Adapted from [11], this domain models a robot traversing a warehouse partitioned into zones connected by checkpoints. Unlike the previous domains, the specification is reward-based, requiring that the expected number of steps to reach the target zone is at most b under all admissible uncertainties. The parameter specifies the number of steps needed to move between checkpoints.

B. Experimental Setup

We compare the vertex enumeration (Encoding 1) and dualization-based (Encoding 2) MILP formulations across the benchmark domains. Table I reports the IMDP model size (number of states and transitions), MILP encoding size (number of binary variables, real-valued variables, and constraints), and solution time for each instance.

Recall that permissiveness $\beta(\theta)$ counts the total number of state–action pairs allowed by a multi-strategy. For reporting, we normalize this measure to obtain values in $[0,1]$. We use two variants: (i) *normalized permissiveness*, given by $\frac{\sum_{s \in S} |\theta(s)|}{\sum_{s \in S} |\alpha(s)|}$ over all states S, and (ii) *normalized choice-state permissiveness*, the same ratio restricted to non-target states that are reachable from the initial state and have at least two available actions ($|\alpha(s)| \geq 2$). The latter focuses on states where the controller faces genuine non-deterministic choices, offering a clearer view of the flexibility preserved by the synthesized multi-strategy.

C. Results Analysis

The results in Table I show that both encodings are capable of synthesizing maximally permissive robust strategies. In all benchmark instances, the normalized permissiveness is close to 1, confirming that the synthesized strategies preserve nearly all available actions.

Both approaches also scale to very large IMDPs with up to hundreds of thousands of states and transitions, solving

Domain		IMDP Size		Vertex Enumeration (Encoding 1)					Dualization (Encoding 2)				
Name	Parameters	#States	#Trans	#Binary	#Real	#Constraints	Time(s)	Permissive(%)	#Binary	#Real	#Constraints	Time(s)	Permissive(%)
ACA	10	328	4,739	984	390	110,398	4.3	98.6 [90.1]	984	11,774	31,010	7.7	98.6 [87.0]
	12	453	6,498	1,203	529	585,749	22.7	98.9 [92.5]	1,203	15,855	41,507	5.6	98.9 [88.9]
	14	640	9,583	1,632	733	2,814,507	145.4	99.4 [95.8]	1,632	23,070	60,163	4.3	99.4 [93.9]
SAV	5	279	999	727	363	1,370	0.6	97.0 [33.3]	727	3,731	8,274	5.3	97.0 [35.3]
	6	411	1,503	1,099	539	2,050	46.3	97.6 [35.0]	1,099	5,615	15,755	218.6	97.6 [35.0]
	7	567	2,103	1,543	747	2,858	2,564.7	97.6 [31.5]	1,543	7,859	22,071	4,030.2	97.6 [35.1]
OBS	5; 1000	73,025	73,299	73,083	73,030	146,471	2.1	99.9 [63.2]	73,083	365,789	951,108	4.8	99.9 [51.9]
	6; 100	11,336	1,760	11,424	11,342	23,324	1.4	99.7 [42.3]	11,424	57,704	115,840	3.7	99.7 [31.8]
	6; 1000	113,036	113,460	113,124	113,042	226,724	3.2	99.9 [50.0]	113,124	566,204	1,132,840	14.8	99.9 [67.4]
WH	250	2,762	3,785	3,768	2,762	6,549	0.2	99.9 [99.9]	3,768	17,868	36,861	1.1	99.9 [99.9]
	500	5,512	7,535	7,518	5,512	13,049	0.4	99.9 [99.9]	7,518	35,618	63,261	1.8	99.9 [99.9]
	1000	11,012	15,035	15,018	11,012	26,049	0.7	99.9 [99.9]	15,018	71,118	146,261	3.0	99.9 [99.9]

TABLE I

EXPERIMENTAL RESULTS COMPARING VERTEX ENUMERATION (ENCODING 1) AND DUALIZATION (ENCODING 2). PERMISSIVE (%) VALUES ARE REPORTED AS NORMALIZED PERMISSIVENESS, WITH THE CHOICE-STATE PERMISSIVENESS SHOWN IN BRACKETS.

all but some SAV instances within seconds. SAV-7 instances are notably slower, due to their more complex structure and communication-related modeling overhead.

Regarding MILP size, both encodings introduce the same number of binary variables, but differ in continuous variables and constraints. The vertex enumeration method is conceptually straightforward and yields relatively compact encodings for small models. However, its number of constraints grows exponentially with the number of successors, resulting in very large MILPs and long runtimes on larger instances (e.g., ACA models). In contrast, the dualization approach scales linearly with the number of successors, leading to more balanced MILPs and shorter runtimes in such cases.

Finally, while both methods achieve the same normalized permissiveness across all benchmarks, their choice-state permissiveness often differs. This indicates that the two encodings can produce different multi-strategies, even though both satisfy the same robust permissiveness objective.

VI. CONCLUSION

We presented the first framework for robust permissive controller synthesis on IMDPs, enabling controllers that guarantee specification satisfaction under all admissible transitions while preserving multiple action choices for runtime flexibility. The problem was formulated as mixed-integer linear programs with two encodings: a conceptually simple vertex-enumeration method and a dualization-based method that scales linearly with the number of successors. Experiments on four benchmark domains showed that both approaches synthesize maximally permissive robust strategies, with the dualization method offering superior scalability on large models.

Future work includes applying our methods to a wide range of robotic domains, and integrating permissive synthesis with learning-based control. In particular, robust permissive strategies align naturally with the notion of shielding in safe reinforcement learning, suggesting opportunities for safety-preserving integration with meta- and in-context RL, where agents must adapt across distributions of tasks.

APPENDIX

Correctness of Encoding 1. We first establish two auxiliary lemmas and then present the proof of Theorem 1.

Lemma 1: For any (s,a) and vector $x \in \mathbb{R}^S$, the extremum of $\sum_{s'} P(s,a,s')\, x_{s'}$ over $P \in \mathcal{P}(s,a)$ is attained at some $P^{(v)}(s,a,\cdot) \in \mathcal{V}(s,a)$.

Proof: For fixed (s,a) and x, the expression $\sum_{s'} P(s,a,s')\, x_{s'}$ is linear, and $\mathcal{P}(s,a)$ is a convex polytope. Hence any extremum over $\mathcal{P}(s,a)$ is attained at a vertex $P^{(v)}(s,a,\cdot) \in \mathcal{V}(s,a)$. ∎

Lemma 2: Let $\mathcal{M}$ be an IMDP, $\varphi = \mathsf{P}_{\geq p}(\Diamond T)$ a property, and θ a multi-strategy. Consider the inequalities for $s \in S$: $x_s \leq \min_{a \in \theta(s)} \sum_{s' \in S} P^{(v)}(s,a,s') \cdot x_{s'}$. Then values $\bar{x}_s \in \mathbb{R}$ for $s \in S$ are a solution to the above inequalities iff $\bar{x}_s = \inf_{\sigma \lhd \theta} Pr^{\sigma,v}_{\mathcal{M},s}(\Diamond T)$, where $Pr^{\sigma,v}_{\mathcal{M},s}(\Diamond T)$ denotes the probability of reaching the target set T under strategy σ and transition distribution $P^{(v)}$ of $\mathcal{M}$, starting from state s.

Proof: Fixing a vertex v yields a standard MDP with transitions $P^{(v)}$. The inequalities correspond to the Bellman equations for minimum reachability; by monotonicity and standard MDP results [37], their solution $\bar{x}_s$ coincides with the value function. ∎

Proof of Theorem 1: (Soundness) Let (x,y) be feasible for Encoding 1 and define $\theta(s) = \{a \mid y_{s,a} = 1\}$. Constraints (1b)–(1d) enforce non-deadlock, $x_{s_\iota} \geq p$, and $x_s = 1$ for $s \in T$. Constraint (1e), together with Lemmas 1 and 2, ensures that x is a sub-fixed-point of the robust Bellman operator under θ. Since $\bar{x}$ is the least fixed point of this operator, we obtain $x \leq \bar{x}$, hence $\bar{x}_{s_\iota} \geq x_{s_\iota} \geq p$. Optimality of the MILP objective yields maximal permissiveness.

(Completeness) Conversely, given a robust maximally permissive θ, let y encode θ and set $x = \bar{x}$, the robust reachability values from Lemma 2. Then (x,y) satisfies all constraints, the objective equals $\beta(\theta)$, and optimality follows from maximal permissiveness. ∎

Correctness of Encoding 2. We first establish an auxiliary lemma and then present the proof of Theorem 2.

Lemma 3: Fix (s,a) and $x \in \mathbb{R}^S$. The robust inequality

$x_s \leq \min_{P(\cdot|s,a)\in\mathcal{P}(s,a)} \sum_{s'\in S} P(s,a,s')\, x_{s'}$ holds iff there exist dual variables $\hat{u}_{s,a,s'}, \check{u}_{s,a,s'} \geq 0$ and $\lambda_{s,a} \in \mathbb{R}$ such that $\lambda_{s,a} - \hat{u}_{s,a,s'} + \check{u}_{s,a,s'} \leq x_{s'}$ $(\forall s' \in S)$ and $x_s \leq \lambda_{s,a} + \sum_{s'\in S}\big(\check{P}(s,a,s')\,\check{u}_{s,a,s'} - \hat{P}(s,a,s')\,\hat{u}_{s,a,s'}\big)$.

Proof: The right-hand side of the robust inequality is the optimum of an LP over $P(s,a,\cdot)$ subject to interval and simplex constraints. Its dual has variables $\hat{u}, \check{u} \geq 0$ and λ, yielding exactly the displayed conditions. By strong LP duality [32], the primal minimum equals the dual maximum, establishing equivalence. ∎

Proof of Theorem 2: (Soundness) Let $(x, y, \hat{u}, \check{u}, \lambda, \eta)$ be feasible for Encoding 2 and set $\theta(s) = \{a \mid y_{s,a} = 1\}$. Constraints (2b)–(2d) enforce non-deadlock, $x_{s_\iota} \geq p$, and $x_s = 1$ for $s \in T$. By Lemma 3, (2e)–(2f) are equivalent to the robust Bellman inequality $x_s \leq \min_{P\in\mathcal{P}(s,a)} \sum_{s'} P(s,a,s') x_{s'}$ for each allowed (s,a); (2g)–(2h) linearize $\eta_{s,a} = \lambda_{s,a} y_{s,a}$ (big-M), so the dual bound is active only if $y_{s,a} = 1$. Hence x is a sub-fixed-point of the robust Bellman operator under θ. Let $\bar{x}$ be the robust reachability values; then $x \leq \bar{x}$ by Lemma 2, so $\bar{x}_{s_\iota} \geq p$. Optimality of the objective (2a) yields maximal permissiveness.

(Completeness) Let θ be a robust maximally permissive multi-strategy. Set y to encode θ and $x = \bar{x}$ (the robust values from Lemma 2). For each allowed (s,a), Lemma 3 provides dual variables $\hat{u}, \check{u}, \lambda$ making (2e)–(2f) hold at x; take $\eta_{s,a} = \lambda_{s,a}$ if $y_{s,a} = 1$ and $\eta_{s,a} = 0$ otherwise, so (2g)–(2h) hold. Thus all constraints are satisfied and the objective equals $\beta(\theta)$; optimality follows from maximal permissiveness. ∎

REFERENCES

[1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. The MIT Press, 2005.

[2] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray, *Algorithms for decision making*. MIT press, 2022.

[3] T. Badings, T. D. Simao, M. Suilen, and N. Jansen, "Decision-making under uncertainty: beyond probabilities: Challenges and perspectives," *International Journal on Software Tools for Technology Transfer*, vol. 25, no. 3, pp. 375–391, 2023.

[4] M. Suilen, T. Badings, E. M. Bovy, D. Parker, and N. Jansen, "Robust Markov decision processes: A place where AI and formal methods meet," *arXiv e-prints*, pp. arXiv–2411, 2024.

[5] J. Jiang, S. Coogan, and Y. Zhao, "Abstraction-based planning for uncertainty-aware legged navigation," *IEEE Open Journal of Control Systems*, vol. 2, pp. 221–234, 2023.

[6] J. Jiang, Y. Zhao, and S. Coogan, "Safe learning for uncertainty-aware planning via interval MDP abstraction," *IEEE Control Systems Letters*, vol. 6, pp. 2641–2646, 2022.

[7] N. Cauchi, L. Laurenti, M. Lahijanian, A. Abate, M. Kwiatkowska, and L. Cardelli, "Efficiency through uncertainty: Scalable formal synthesis for stochastic hybrid systems," in *Proceedings of the 22nd ACM international conference on hybrid systems: computation and control*, 2019, pp. 240–251.

[8] S. Adams, M. Lahijanian, and L. Laurenti, "Formal control synthesis for stochastic neural network dynamic models," *IEEE Control Systems Letters*, vol. 6, pp. 2858–2863, 2022.

[9] K. Drager, V. Forejt, M. Kwiatkowska, D. Parker, and M. Ujma, "Permissive controller synthesis for probabilistic systems," *Logical Methods in Computer Science*, vol. 11, 2015.

[10] S. Junges, N. Jansen, C. Dehnert, U. Topcu, and J.-P. Katoen, "Safety-constrained reinforcement learning for MDPs," in *International conference on tools and algorithms for the construction and analysis of systems*. Springer, 2016, pp. 130–146.

[11] S. Chen, K. Boggess, D. Parker, and L. Feng, "Multi-objective controller synthesis with uncertain human preferences," in *ACM/IEEE 13th International Conference on Cyber-Physical Systems*, 2022.

[12] N. Jansen, B. Könighofer, S. Junges, A. Serban, and R. Bloem, "Safe reinforcement learning using probabilistic shields," in *31st International Conference on Concurrency Theory*, 2020.

[13] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] J. Beck, R. Vuorio, E. Z. Liu, Z. Xiong, L. Zintgraf, C. Finn, and S. Whiteson, "A survey of meta-reinforcement learning," *arXiv preprint arXiv:2301.08028*, 2023.

[15] A. Moeini, J. Wang, J. Beck, E. Blaser, S. Whiteson, R. Chandra, and S. Zhang, "A survey of in-context reinforcement learning," *arXiv preprint arXiv:2502.07978*, 2025.

[16] J. Jackson, L. Laurenti, E. Frew, and M. Lahijanian, "Strategy synthesis for partially-known switched stochastic systems," in *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, 2021, pp. 1–11.

[17] M. Cleaveland, P. Lu, O. Sokolsky, I. Lee, and I. Ruchkin, "Conservative perception models for probabilistic verification," *arXiv preprint arXiv:2503.18077*, 2025.

[18] F. B. Mathiesen, S. Haesaert, and L. Laurenti, "Scalable control synthesis for stochastic systems via structural IMDP abstractions," in *Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control*, 2025, pp. 1–12.

[19] M. Nazeri, T. Badings, S. Soudjani, and A. Abate, "Data-driven yet formal policy synthesis for stochastic nonlinear dynamical systems," *arXiv preprint arXiv:2501.01191*, 2025.

[20] J. Skovbekk, L. Laurenti, E. Frew, and M. Lahijanian, "Formal verification of unknown dynamical systems via Gaussian process regression," *IEEE Transactions on Automatic Control*, 2025.

[21] A. Nilim and L. E. Ghaoui, "Robust control of Markov decision processes with uncertain transition matrices," in *Proceedings of the 19th Conference on Neural Information Processing Systems*, 2005, pp. 839–846.

[22] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[23] E. M. Wolff, U. Topcu, and R. M. Murray, "Robust control of uncertain Markov decision processes with temporal logic specifications," in *Proceedings of the 51st IEEE Conference on Decision and Control*, 2012, pp. 3372–3379.

[24] A. Puggelli, W. Li, A. L. Sangiovanni-Vincentelli, and S. A. Seshia, "Polynomial-time verification of PCTL properties of MDPs with convex uncertainties," in *Proceedings of the 25th International Conference on Computer Aided Verification*, ser. Lecture Notes in Computer Science, vol. 8044. Springer, 2013, pp. 527–542.

[25] E. M. Hahn, V. Hashemi, H. Hermanns, M. Lahijanian, and A. Turrini, "Multi-objective robust strategy synthesis for interval Markov decision processes," in *International Conference on Quantitative Evaluation of Systems*. Springer, 2017, pp. 207–223.

[26] M. Suilen, T. D. Simão, D. Parker, and N. Jansen, "Robust anytime learning of markov decision processes," in *NeurIPS 2022*, 2022.

[27] T. Badings, L. Romao, A. Abate, D. Parker, H. A. Poonawala, M. Stoelinga, and N. Jansen, "Robust control for dynamical systems with non-Gaussian noise via formal abstractions," *Journal of Artificial Intelligence Research*, vol. 76, pp. 341–391, 2023.

[28] M. Kwiatkowska, G. Norman, and D. Parker, "Prism 4.0: Verification of probabilistic real-time systems," in *International conference on computer aided verification*. Springer, 2011, pp. 585–591.

[29] T. Wooding and J. Lavaei, "Impact: An efficient framework for interval Markov decision processes," in *Proceedings of the 27th International Conference on Hybrid Systems: Computation and Control*. ACM, 2024.

[30] F. B. Mathiesen, M. Lahijanian, and L. Laurenti, "Intervalmdp.jl: Accelerated value iteration for interval markov decision processes," in *ADHS*, ser. IFAC-PapersOnLine, vol. 58, no. 11. Elsevier, 2024, pp. 1–6.

[31] P. Yu, Y. Li, D. Parker, and M. Kwiatkowska, "Planning with linear temporal logic specifications: Handling quantifiable and unquantifiable uncertainty," *arXiv preprint arXiv:2502.19603*, 2025.

[32] A. Ben-Tal, A. Nemirovski, and L. El Ghaoui, *Robust optimization*. Princeton university press, 2009.

[33] G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems," *Mathematical programming*, vol. 10, no. 1, pp. 147–175, 1976.

[34] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2025.

[35] Y. Schnitzer, A. Abate, and D. Parker, "Certifiably robust policies for uncertain parametric environments," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2025, pp. 63–83.

[36] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[37] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.