DECEPTIVE PLANNING VIA LAST DECEPTIVE STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing deceptive planning methods face two key limitations. Graph-based deceptive path planning (DPP) splits deception into two phases via a Last Deceptive Point (LDP), but its switching criterion can trigger premature exposure when decoy goals lie near the optimal route to the real goal. The Deceptive Decision Making (DDM) framework handles MDPs via constrained linear programming, yet it optimizes deception uniformly across all states—failing to account that once the agent is sufficiently close to the real goal, continuing deceptive actions not only increases path length but also prolongs the exposure of its true intention. We introduce the *Last Deceptive State* (LDS) concept within deterministic MDPs and propose LDS-guided deceptive planning (LDS-DP), a framework that integrates the phase-switching intuition of DPP with DDM's LP formulation. LDS-DP identifies the value-maximal LDS candidate through V-differences, then decomposes planning into a deceptive linear programming phase and a shortest-path truthful phase, thereby addressing the premature-exposure issue of DPP while eliminating the post-LDS inefficiency of DDM. Experiments in deterministic and near-deterministic grid worlds show that LDS-DP preserves DDM's deception quality while shortening the post-LDS phase, with the two-phase advantage being robust to mild stochasticity.

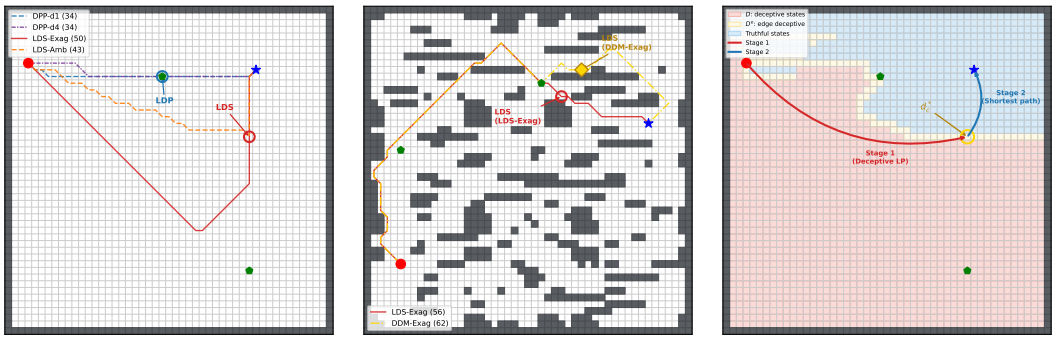
1 INTRODUCTION

Deception, defined as the cultivation and sustained maintenance of false beliefs in others' minds Carson (2010), is a crucial skill that enables agents to gain strategic advantages by concealing their true intentions in adversarial environments. Its applications are widespread, spanning multi-agent negotiation Greenberg (1982); Matsubara & Yokoo (1998), fugitive pursuit Shieh et al. (2012), network intrusion Geib & Goldman (2001), business domain Chelliah & Swamy (2018) and privacy preservation Keren et al. (2016). Non-player characters capable of deception are also known to create more credible and engaging interactions than those that behave honestly Dias et al. (2013).

The challenge of executing a task while deceiving an observer is exemplified by scenarios such as troop maneuver Tzu (2008). Here, a commander must move units to a target while avoiding ambush. Success hinges on balancing the efficiency of the movement with the need to conceal the final destination for as long as possible. Since moving directly toward the real target will inevitably reveal it, the troops must take deliberate, often circuitous, actions to mislead the adversary, thereby seizing the initiative. This work formulates and addresses this core problem within the framework of decision-making in complex environments.

In recent years, several deceptive planning methods have been proposed, including path-planning-based methods and MDP-based optimization methods. The graph-based *Deceptive Path Planning* (DPP) (Masters & Sardina, 2017b) introduced a two-phase framework: the agent maintains deception until a switching point (the Last Deceptive Point, or LDP), then proceeds directly toward the real goal. While intuitive, DPP exhibits a structural limitation. The LDP criterion, computed through graph-level cost differences, can trigger premature exposure when decoy goals lie near the optimal route to the real goal. Its formulation additionally prevents application to settings with even mild stochasticity. The *Deceptive Decision Making* (DDM) framework (Savas et al., 2022) addresses stochastic transitions by formulating deception as a constrained linear program over an MDP, producing globally optimal stationary policies with reachability guarantees. However, a structural limitation remains: DDM optimizes deception uniformly across all states. Consequently, DDM's single-phase LP continues to invest deception-aware actions long after the agent's true intention has become

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Premature exposure in DPP (b) Prolonged exposure in DDM (c) The LDS-DP concept

Figure 1: **Left**: DPP's premature exposure. The red circle (○) marks the start, green pentagons denote decoy goals, and the blue star (⋆) is the real goal g_r. A decoy goal lies near the direct route to g_r. DPP-d1 (blue dashed) and DPP-d4 (indigo dash-dot) trigger their LDP (blue ○) close to that decoy, leaving 14 truthful steps. LDS-Exag (red solid) and LDS-Amb (orange dashed) push their LDS (red ○) to $(36, 29)$, reducing the truthful phase to 10 steps. **Middle**: DDM's prolonged exposure. After both methods pass their last deceptive state, DDM-Exag (yellow dash-dot) continues taking deceptive detours, exposing the true intention over 16 additional steps. In contrast, LDS-Exag (red solid) switches to the shortest path at its LDS $(29, 35)$, completing the truthful phase in only 13 steps. **Right**: Illustration of the LDS-DP concept. The deceptive set D (light red) comprises all states where a decoy's posterior meets or exceeds that of g_r, with the edge deceptive set D^e (light yellow) forming its boundary. Truthful states (light blue) lie outside D. The value-maximal LDS candidate d_c^* (yellow ○) is the deceptive state maximizing V_{g_r}. Stage 1 (red arrow) solves a deceptive LP from the start to d_c^*; Stage 2 (blue arrow) follows the shortest path to g_r.

apparent, incurring unnecessary trajectory cost. Figure 1a and 1b illustrates these limitations with concrete examples.

To address the limitations of DPP and DDM, we propose *Last Deceptive State-guided Deceptive Planning* (LDS-DP). LDS-DP identifies a *Last Deceptive State* (LDS)—an intermediate state that partitions the agent's trajectory into a deceptive phase and a truthful phase. Grounded in MDP value functions, the LDS pushes the switching point deeper into the deceptive region than DPP's graph-based LDP. We characterize the value-maximal LDS candidate as $d_c^* = \arg\max_{s \in D} V_{g_r}(s)$ (formal treatment in Section 4), enabling efficient computation from precomputed value functions. Based on d_c^*, LDS-DP decomposes planning into two phases: Stage 1 solves a deceptive LP targeting d_c^*, inheriting DDM's optimality guarantees; Stage 2 follows the shortest path to g_r, eliminating the post-deception inefficiency of single-phase LP methods. This two-phase framework addresses both the premature-exposure problem of DPP and the prolonged truthful tail of DDM within a unified approach.

We conduct two experiments on 40 grid-world problems spanning five map types. The first is under deterministic transitions ($\alpha = 1.0$), comparing LDS-DP against DPP-d1 (Masters & Sardina, 2017b), DPP-d4 (Masters & Sardina, 2017b), AM (Liu et al., 2021), DDM-Exag (Savas et al., 2022), DDM-Amb (Savas et al., 2022), and an Honest baseline. The second is under near-deterministic transitions with 10% action slip ($\alpha = 0.9$), comparing LDS-DP against AM, DDM-Exag, DDM-Amb, and Honest. Across both settings, LDS-DP preserves DDM's deception quality while systematically shortening the post-LDS tail.

In summary, this work makes three contributions:

(1) We analyze the structural limitations of DPP and DDM in deceptive planning—premature exposure under LDP, and uniform optimization without phase distinction.

(2) We propose the concept of the Last Deceptive State (LDS) as a principled switching criterion within MDPs, and show that LDS-DP addresses these limitations.

(3) We provide a theoretical characterization connecting the value-maximal deceptive state to a shortest truthful tail, and empirically evaluate the decomposition under mild transition noise.

2 Related Works

Our work builds upon and connects ideas from several threads of research on computational deception, with a focus on planning in fully observable environments.

The concept of *Deceptive Path Planning* (DPP) was pioneered by Masters and Sardina (Masters & Sardina, 2017b), who introduced formal notions of *simulation* ("showing the false") and *dissimulation* ("hiding the real"). A key contribution is the Last Deceptive Point (LDP), a switching state computed by measuring deviations from optimal paths via path-generation heuristics. While intuitive, the LDP criterion suffers from one limitation—the path-generation heuristics (Hart et al., 1968) used by DPP-d1/d4 may yield trajectories whose LDP occurs early under particular decoy-goal geometries.

Research has also extended deception to settings that account for uncertain environments. Ornik and Topcu (Ornik & Topcu, 2018) incorporated the observer's belief into the agent's model, formalizing deceptive MDPs for stochastic environments. Savas et al. (Savas et al., 2022) later formulated deception as a constrained LP over an MDP, introducing the *Deceptive Decision Making* (DDM) framework. Employing a maximum-entropy observer model, DDM produces globally optimal stationary policies with reachability guarantees and supports both exaggeration and ambiguity through carefully designed cost functions on the state-action occupancy measure. Unlike our approach, DDM optimizes deception uniformly across all states without distinguishing between deceptive and truthful phases.

Liu et al. (Liu et al., 2021) proposed the *Ambiguity Model* (AM), which models observer beliefs via Q-difference-based plan recognition and selects actions maximizing posterior entropy at each step, pruning those with negative Q-gain toward the real goal. Lewis and Miller (Lewis & Miller, 2023) later introduced DEAM to address AM's exploration limitations in model-free and continuous domains. Based on Q-differences, their heuristic action-selection mechanism supports deceptive planning across general MDPs; however, they focus exclusively on "dissimulation" without explicitly modeling "simulation".

Deception strategies inherently counter an observer's plan recognition capability. Our work adopts the goal recognition framework where observers infer an agent's goal from observed behavior (Ramírez & Geffner, 2010). A widely adopted approach, and the one we assume our observer employs, is *cost-difference-based* recognition (Masters & Sardina, 2017a), which quantifies how much observed behavior deviates from the optimal path to each candidate goal. Masters and Sardina (Masters & Sardina, 2019) later extended this framework to accommodate both rational and irrational observers. The value-difference variant from (Savas et al., 2022) generalizes this principle to MDPs, using state value differences (V-differences).

Recent work has extended deceptive planning to broader settings. Fang et al. (Fang et al., 2026) propose a receding-horizon multi-agent planner with stochastic policies, while Cao et al. (Cao et al., 2026) study repeated deception against adaptive, learnable observers. Suttle et al. (Suttle et al., 2025) introduce a value-of-information-based formulation under adversarial interventions, where the observer can actively impede the agent. The work (Gutierrez et al., 2025) proposes polynomial path planning. These works, however, address substantially different settings—multi-agent coordination, repeated interactions, adversarial observers, or continuous motion planning, and are thus orthogonal to our single-agent framework with a passive, rational observer. They reflect a growing interest in multi-agent, dynamic, and learnable-observer deception scenarios.

3 Preliminaries

This section reviews the foundational concepts and models from prior work necessary to understand our proposed deceptive planning framework.

3.1 Markov Decision Processes

Definition 1 (Markov Decision Process)**.** *A Markov decision process (MDP) is a six-tuple* $\mathcal{M} = (S, s_1, A, P, r, \gamma)$*, where* $s_1 \in S$ *is the unique initial state;* A *denotes a finite set of actions;* $P : S \times A \times S \to [0, 1]$ *is the stochastic state transition function such that* $\sum_{s' \in S} P(s, a, s') = 1$ *for*

all $s \in S$ and $a \in A$, where $P(s, a, s')$ represents the probability of transitioning to state s' when executing action a in state s; $r : S \times A \times S \to \mathbb{R}$ is the reward received when executing action a in state s and transitioning to state s'; $\gamma \in (0, 1]$ is the discount factor.

Problems related to MDPs are solved by finding a policy $\pi(a \mid s)$, which prescribes the probability of taking action a in state s. The objective is to maximize the value function, defined as: $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t, s_{t+1})\right]$. The optimal value function can be computed via value iteration (Sutton et al., 1998; Baier & Katoen, 2008).

A trajectory ζ is a sequence of states and actions $(s_1, a_1, s_2, a_2, s_3, \dots)$ such that $P(s_t, a_t, s_{t+1}) > 0$ for all $t \in \mathbb{N}$. A partial trajectory $\zeta_{1:t}$ of length $t \in \mathbb{N}$ is the sequence $(s_1, a_1, s_2, \dots, s_t)$. Let $\zeta_{1:t}[n] = s_n$ with $1 \le n \le t$ denote the state visited at the n-th step along $\zeta_{1:t}$.

3.2 Observer Models for Goal Recognition

Prior work has proposed different approaches for modeling how observers infer an agent's goal from observed behavior. We review the approach that forms the foundation of our framework.

***V*-difference-based recognition.** The work Savas et al. (2022) proposed a predictive model that integrates maximum-entropy reinforcement learning with rationality assumptions, computing a probability $\Pr(g \mid \zeta_{1:t})$ for each potential goal $g \in G$. Prior beliefs form a probability distribution over the agent's potential goals, formalizing the observer's judgment of "the goal state the agent is expected to reach from the initial state s_1". Priors are typically uniform; in repeated or specialized interactions, they can be constructed from historical data via Bayesian methods (Ziebart et al., 2009). Given the prior, the observer's posterior prediction follows Bayes' theorem:

$$\Pr(g \mid \zeta_{1:t}) = \frac{\Pr(\zeta_{1:t} \mid g)\Pr(g)}{\sum_{g' \in G} \Pr(\zeta_{1:t} \mid g')\Pr(g')},$$

where $\Pr(\zeta_{1:t} \mid g)$ denotes the probability that the agent follows trajectory $\zeta_{1:t}$ to reach g.

The intent recognition model provides a formula for calculating probabilities based on state value differences (V-differences). Let $Q_g(s, a)$ denote the goal-conditioned action-value function, i.e., the expected return of taking action a in state s and then following the soft-optimal policy toward goal g, defined recursively by the soft Bellman equation

$$Q_g(s, a) = \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[r(s, a, s') + \gamma \, V_g(s')], \qquad V_g(s) = \mathrm{softmax}_{a \in A} Q_g(s, a).$$

When transition randomness has a limited impact on the agent's behavior and the discount factor γ is sufficiently large, for this softmax entropy iteration we have:

$$\Pr(g \mid \zeta_{1:t}) \approx \frac{\exp(\Delta V_g(\zeta_{1:t}))\Pr(g)}{\sum_{g' \in G} \exp(\Delta V_{g'}(\zeta_{1:t}))\Pr(g')},$$

where $\mathrm{softmax}_x f(x) = \tau \log \sum_x \exp(f(x)/\tau)$, and $\Delta V_g(\zeta_{1:t}) = \Delta V_g(s_1, s_t) = V_g(s_t) - V_g(s_1)$. Intuitively, $\Delta V_g(\zeta_{1:t})$ represents the state value growth for goal state g after the agent follows trajectory $\zeta_{1:t}$. The model predicts the observer's inference by controlling τ: as $\tau \to \infty$, the observer deems the trajectory inefficient and assumes random behavior (equal posterior probabilities for all goal states); as $\tau \to 0$, the agent is predicted to be fully efficient, following the least-cost trajectory.

3.3 DDM: Deceptive Decision Making via LP

DDM (Savas et al., 2022) supports two deception strategies parameterized by the function $f(s, a)$.

Exaggeration. The agent aims to convince the observer that it is heading toward a specific decoy:

$$f_{\mathrm{exag}}(s, a) = 1 + \Pr(g_r \mid s_1, s) - \max_{g \neq g_r} \Pr(g \mid s_1, s) \tag{1}$$

for $s \in S \setminus G$, and $f_{\mathrm{exag}}(s, a) = 0$ for goal states. Lower values indicate stronger deception.

Ambiguity. The agent aims to maximize the observer's uncertainty among all goals:

$$f_{\mathrm{amb}}(s, a) = \sum_{g, g' \in G} \left| \Pr(g \mid s_1, s) - \Pr(g' \mid s_1, s) \right| \tag{2}$$

for $s \in S \setminus G$, zero otherwise. Minimizing this cost keeps all goal posteriors close.

DDM synthesizes a deceptive policy by solving a constrained LP over the state-action occupancy measure $x(s, a)$. Let S_0 be states from which no goal is reachable, and $S_r = S \setminus (\{g_r\} \cup S_0)$. Since decoy goals are ordinary non-absorbing states, they remain in S_r and are traversable. The agent's objective is:

$$\min_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} g(s_t, a_t) \right] \quad \text{s.t.} \quad \Pr_\pi(\text{Reach}[g_r]) = R_{\max}(g_r) \tag{3}$$

where $g(s, a) = \gamma_a^{T_{\min}(s)} \cdot f(s, a)$ discounts the deception cost f by the shortest distance $T_{\min}(s)$ from s_1 to s, using $\gamma_a \in (0, 1]$.

This is solved via two successive LPs. The first LP minimizes the expected deception cost:

$$\min_{x \geq 0} \sum_{s \in S_r} \sum_a g(s, a)\, x(s, a) \tag{4}$$

$$\text{s.t.} \sum_a x(s, a) - \sum_{s', a} P(s', a, s)\, x(s', a) = \alpha_s, \ \forall s \in S_r \tag{5}$$

$$\sum_{s \in S_r} \sum_a x(s, a)\, r(s, a) = R_{\max}(g_r) \tag{6}$$

where $\alpha_s = 1$ if $s = s_1$ and 0 otherwise, and $r(s, a) = P(s, a, g_r)$. The second LP minimizes total steps subject to achieving the optimal deception cost. The resulting stationary policy $\pi_{\text{DDM}}(a \mid s) = x^*(s, a) / \sum_{a'} x^*(s, a')$ is globally optimal.

4 Deceptive Planning via Last Deceptive State

The theoretical development in this section assumes deterministic transitions, consistent with the path-planning setting of DPP (Masters & Sardina, 2017b): the agent incurs per-step costs of 1 (cardinal) or $\sqrt{2}$ (diagonal), all non-terminal transitions have strictly positive cost, and only g_r is absorbing with terminal reward 0. Under this standard model: *(i)* $V_{g_r}(s) = -J^*(s, g_r)$ where J^* is the shortest-path distance from s to g_r on the grid; *(ii)* along any shortest path from s to g_r, each successive state has strictly larger V_{g_r}; and *(iii)* all non-obstacle states are mutually reachable. All results below are proved under this setting.

4.1 Components of a Deceptive Trajectory

To structure our planning framework, we first introduce several key concepts that characterize the agent's state from the perspective of the observer's beliefs. These components allow us to formally define the critical juncture—the Last Deceptive State—where the agent's strategy should transition from deception to truthfulness.

Definition 2 (Truthful State)**.** *A **truthful state** is a state $s \in S$ at which the probability of the real goal g_r is greater than the probability of any other goal in G. Formally, for all $g \in G \setminus \{g_r\}$, $\Pr(g_r \mid s_1, s) > \Pr(g \mid s_1, s)$. Otherwise, the state is a **deceptive state**.*

This binary classification is fundamental: it tells the agent whether it is currently succeeding at deception (s is deceptive) or whether its true intent has been uncovered (s is truthful).

Definition 3 (Last Deceptive State (LDS))**.** *For a trajectory $\zeta_{1:T}$ from s_1 to g_r, its **Last Deceptive State** is a state $s_k \in \zeta_{1:T}$ satisfying:*

- $\zeta_{1:T}[k] = s_k$ is a deceptive state;
- For all $j \in \{k+1, \ldots, T\}$ and all $g \neq g_r$, $\Pr(g_r \mid \zeta_{1:j}) > \Pr(g \mid \zeta_{1:j})$.

The LDS defines the precise switching point in our two-phase planning framework: after this state, all subsequent states are truthful (by Definition 2), meaning the observer is no longer deceived, and thus the agent should simply follow the shortest path to g_r. Identifying this state is therefore the key to balancing deception and efficiency.

Notably, we assume that termination at g_r is observed, making g_r truthful by convention.

Corollary 1. *The LDS always exists, because the initial state s_1 has uniform prior probabilities over all goals, making s_1 inherently a deceptive state.*

Note: This corollary ensures our search for an LDS is well-founded: every trajectory begins under ambiguity. A fully observable deceptive trajectory eventually becomes truthful as its final action necessarily leads to g_r. Thus, for an agent with multiple goal states, every trajectory contains a deceptive state such that all subsequent states are truthful, guaranteeing an LDS exists.

4.2 Maximizing the Value of the Last Deceptive State

Knowing that an LDS exists is not sufficient for planning; we need to find the *best* one. This section formalizes the properties of deceptive states to ultimately derive a method for computing the value-maximal LDS candidate, which serves as the anchor point for our two-phase planning framework.

Definition 4 (Deceptive States Set D)**.** *For an agent with initial state s_1, the **deceptive states set** D contains all states $s \in S$ where there exists $g \in G \setminus \{g_r\}$ such that $\Pr(g_r \mid s_1, s) \leq \Pr(g \mid s_1, s)$ (equivalently, $\Delta V_g(s_1, s) \geq \Delta V_{g_r}(s_1, s)$).*

The set D is our search space for LDS candidates. However, not all states in D are viable. We now identify progressively stricter subsets. In deterministic settings, we denote by $s' = T(s, a)$ the successor state of (s, a), uniquely determined.

Definition 5 (Edge Deceptive States Set D^e)**.** $D^e = \{s \in D \mid \exists\, a \in A(s) : V_{g_r}(T(s, a)) > V_{g_r}(s)$ *and* $T(s, a) \notin D\}$.

Intuition: A state is on the "edge" of the deceptive set if there *exists* an action that improves the agent's value with respect to g_r and *immediately forces it to exit* the deceptive set (i.e., the next state becomes truthful). This is a critical property for the LDS.

Definition 6 (Critical Deceptive States Set D^c)**.** $D^c = \{s \in D \mid \forall\, a \in A(s) : V_{g_r}(T(s, a)) > V_{g_r}(s) \Rightarrow T(s, a) \notin D\}$.

Intuition: For a critical state, *it is guaranteed* that *all* actions improving the agent's position toward g_r will make the next state truthful. This makes critical states strong candidates for the LDS.

We now progressively narrow the search for the value-maximal LDS candidate through the following results.

Theorem 1. *For any $s \in D$ that can reach g_r, there exists an action $a \in A(s)$ such that $V_{g_r}(s') > V_{g_r}(s)$ (with $s' = T(s, a)$).*

Proof. Since $s \in D$, $s \neq g_r$ (otherwise s would be truthful). Because the agent can always reach g_r from s, there must exist an action $a \in A(s)$ leading to some s' where $V_{g_r}(s') > V_{g_r}(s)$ (otherwise $V_{g_r}(g_r)$ could never be achieved). Thus, $V_{g_r}(g_r) \geq V_{g_r}(s') > V_{g_r}(s)$. □

Theorem 2. *For any $s \in D \setminus D^e$ and all $a \in A(s)$: if $V_{g_r}(s') > V_{g_r}(s)$ (with $s' = T(s, a)$), then $s' \in D$.*

Proof. Assume for contradiction that $s' \notin D$. By Definition 5, if all actions $a \in A(s)$ leading to some s' where $V_{g_r}(s') > V_{g_r}(s)$ and $s' \notin D$, then $s \in D^e$. This contradicts $s \in D \setminus D^e$, so s' must belong to D. □

Corollary 2. *For any state $s \in D \setminus D^e$, there exists an action $a \in A(s)$ such that $V_{g_r}(s') > V_{g_r}(s)$ and $s' \in D$ (with $s' = T(s, a)$).*

Theorem 3 (Value-Maximal LDS Candidate)**.** *Given a state $s^* \in D$ with $V_{g_r}(s^*) \geq V_{g_r}(s)$ for all $s \in D$, then $s^* \in D^c$. Hence, the value-maximal LDS candidate is:*

$$d_c^* = \arg\max_{s \in D} V_{g_r}(s) \tag{7}$$

Proof. Take any action $a \in A(s^*)$ with $V_{g_r}(T(s^*, a)) > V_{g_r}(s^*)$. If $T(s^*, a) \in D$, then $V_{g_r}(T(s^*, a)) > V_{g_r}(s^*)$ contradicts s^* being the maximizer of V_{g_r} over D. Hence $T(s^*, a) \notin D$. Since every value-improving action from s^* leaves D, by Definition 6, $s^* \in D^c$. □

Implication: Theorem 3 is the key result. It states that the deceptive state with the highest value is necessarily a critical state. This provides a straightforward, computable selection rule: d_c^* **is the state in** D **that maximizes** $V_{g_r}(s)$. Conceptually, d_c^* minimizes the remaining truthful-path cost among deceptive states: it is the farthest into D the agent can reach while keeping the subsequent shortest path to g_r as short as possible. This critical deceptive state, $d_c^* = \arg\max_{s \in D} V_{g_r}(s)$, is the **value-maximal LDS candidate**—the linchpin of the LDS-DP framework, defining the transition point between the deceptive and truthful phases.

If multiple states attain the maximum, any one can be chosen; the proof of Theorem 3 holds for every maximizer.

4.3 LDS-DP: Two-Phase Decomposition

LDS-DP exhibits three characteristics: (1) it introduces a two-phase structure that separates deceptive LP optimization in Stage 1 from efficient shortest-path execution in Stage 2; (2) it supports both exaggeration and ambiguity cost functions from DDM, enabling the agent to choose between showing the false and hiding the real; (3) its computational overhead is minimal: computing d_c^* requires only the value functions V_g that are already available from the preprocessing step. Given d_c^*, LDS-DP decomposes trajectory planning into two phases:

$$\zeta = \underbrace{s_1 \xrightarrow{\text{deceptive}} d_c^*}_{\text{Stage 1}} \oplus \underbrace{d_c^* \xrightarrow{\text{optimal}} g_r}_{\text{Stage 2}}.$$

Stage 1 (Deceptive LP Phase): We adopt DDM's LP framework (Eqs. 3–6) with one key modification: the target absorbing state is d_c^* instead of g_r. Concretely, the state space S_r is redefined as $S \setminus (\{d_c^*\} \cup S_0)$, and the reachability constraint in Eq. 6 uses $r(s,a) = P(s,a,d_c^*)$ with maximum reachability $R_{\max}(d_c^*)$. The LP minimizes the expected deception cost while guaranteeing that the agent reaches d_c^* with maximum probability. The resulting stationary policy $\pi_1(a \mid s) = x_1^*(s,a) / \sum_{a'} x_1^*(s,a')$ inherits DDM's optimality guarantees over the reduced state space.

Stage 2 (Truthful Phase): From d_c^*, the agent follows the shortest path to g_r. Following a shortest path strictly increases V_{g_r} (Assumption (ii)), and since $d_c^* = \arg\max_{s \in D} V_{g_r}(s)$, no successor state can belong to D; deception is no longer possible, and the agent simply completes its task as efficiently as possible.

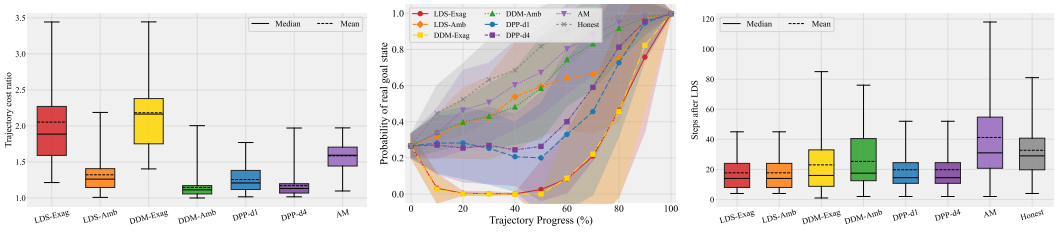
The separation ensures that LP-driven deception is confined to the deceptive phase, eliminating the wasteful post-LDS behavior observed in DDM, while the truthful phase is optimally efficient. Figure 1b illustrates this contrast, and Figure 1c summarizes the key constructs introduced.

5 Experiments

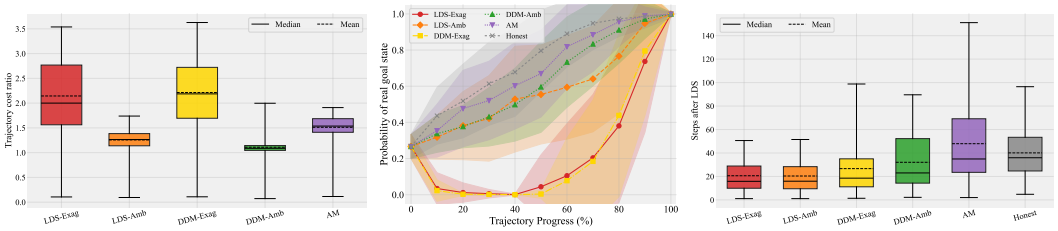
We evaluate LDS-DP against several baselines spanning three families of deceptive planning methods: graph-based (DPP-d1, DPP-d4), online heuristic (AM), and LP-based (DDM-Exag, DDM-Amb). We compare the following methods: (1) **DPP-d1** (simulation) and **DPP-d4** (full pruning) from (Masters & Sardina, 2017b); (2) **AM** from (Liu et al., 2021); (3) **DDM-Exag** and **DDM-Amb** from (Savas et al., 2022); (4) **LDS-Exag** and **LDS-Amb** (our method with Exaggeration and Ambiguity cost functions). We also include **Honest** as a non-deceptive baseline. DPP variants are evaluated only under deterministic transitions.

5.1 Experiment Design

Environment. We evaluate agents in discrete path planning environments, where the agent can move laterally or diagonally to reach the real goal state. We construct 40 grid-world problems across five map types, with 8 problems per map: (1) 49×49 with no obstacles; (2) 49×49 with several large obstacles; (3) 49×49 with many small obstacles; (4) 100×100 with large island obstacles; (5) 100×100 maze-type maps. Each problem specifies a start state, a real goal g_r, and 2–4 decoy goals. The agent moves in 8 directions with cardinal cost 1 and diagonal cost $\sqrt{2}$. To evaluate robustness



(a) Trajectory cost ($\alpha = 1.0$) (b) Deceptiveness ($\alpha = 1.0$) (c) Steps after LDS ($\alpha = 1.0$)



(d) Trajectory cost ($\alpha = 0.9$) (e) Deceptiveness ($\alpha = 0.9$) (f) Steps after LDS ($\alpha = 0.9$)

Figure 2: Shaded regions represent the standard deviation (SD) around the mean. In the box plots, solid lines represent medians, dashed lines represent means, and shaded areas represent interquartile ranges. We apply the two-sided Wilcoxon signed-rank test with per-problem averaged metrics as the sample unit. For $\alpha = 0.9$, each problem's metric is averaged across 10 rollouts.

to transition stochasticity, experiments are conducted under two settings: deterministic ($\alpha = 1.0$) and near-deterministic ($\alpha = 0.9$, 10% slip uniformly distributed among the other reachable adjacent cells). Value iteration $\gamma = 1.0$; the LP discount factor $\gamma_a = 0.95$; $\tau = 1.0$; for $\alpha = 0.9$, value functions are computed using the same transition model, so that d_c^* is re-derived from stochastic V_{g_r} rather than its deterministic counterpart. All experiments were conducted on a single machine with an AMD Ryzen 7 7840H CPU (8 cores, 16 threads), 32 GB RAM, running Windows 11. The LP solver is Gurobi 13.0.

Metrics. The Honest baseline follows the shortest path to g_r without any deception. We use the following metrics: (1) *Trajectory Cost*: the ratio of the trajectory cost generated by different models to that of the Honest model, measuring cost efficiency; (2) *Deceptiveness*: deceptiveness is measured by the probability of the real goal state at different percentages of trajectory completion; (3) *Steps after LDS*: the number of steps taken after the Last Deceptive State (LDS), which is the final state where the observer still assigns higher probability to a decoy goal than the real goal. A rational observer will not be deceived after LDS. Thus, fewer steps after LDS indicate better deception.

5.2 Experiment Results

Trajectory Cost. Figures 2(a) and 2(d) show the full cost ratio distributions. Under deterministic transitions, exaggeration methods are consistently more expensive than their ambiguity counterparts, with LDS-Exag (2.053) lower than DDM-Exag (2.183, $p < 0.01$) and LDS-Amb (1.321) higher than DDM-Amb (1.146). DDM-Amb (1.146) achieves the lowest cost among all deceptive methods, followed by DPP-d4 (1.171), DPP-d1 (1.255), LDS-Amb (1.321), and AM (1.585). Under near-deterministic transitions ($\alpha = 0.9$), LDS-Exag (2.144) is significantly lower than DDM-Exag (2.214, Wilcoxon signed-rank, $p < 0.01$). LDS-Amb (1.257) has somewhat higher cost than DDM-Amb (1.109). AM reaches 1.512. DPP variants are not evaluated under near-deterministic transitions. These results demonstrate that for the exaggeration strategy, LDS-DP effectively reduces DDM's trajectory length. But LDS-Amb shortens the truthful tail but introduces additional pre-LDS cost, leading to a higher overall trajectory cost than DDM-Amb.

Deceptiveness. Figures 2(b) and 2(e) visualize how $\Pr(g_r)$ evolves along trajectory progress. Under deterministic transitions, LDS-Exag achieves the lowest average real-goal posterior (0.229), with *no statistically significant difference* from DDM-Exag (0.237, $p > 0.05$). This is consistent with the LDS-DP design goal of preserving deception quality. LDS-Amb (0.635) and DDM-Amb (0.654) are

comparable. Both exaggeration methods significantly outperform each non-exaggeration baseline ($p < 0.001$, Wilcoxon signed-rank). DPP-d1 (0.430) and DPP-d4 (0.469), which rely on heuristic or pruning-based action selection, are notably weaker than LP-based exaggeration methods on this metric. AM (0.699), as a dissimulation-based approach, achieves the weakest deceptiveness among all methods. The Honest baseline, which follows the shortest path without deception, attains the highest true-goal posterior (0.752), confirming that all deceptive methods meaningfully reduce the observer's certainty. Under near-deterministic transitions ($\alpha = 0.9$), DDM-Exag achieves the lowest posterior (0.214) closely followed by LDS-Exag (0.215), with no statistically significant difference (Wilcoxon signed-rank, $p > 0.05$). Among ambiguity methods, LDS-Amb (0.582) improves over DDM-Amb (0.634), while AM's posterior reaches 0.692, and Honest reaches 0.750. For both exaggeration and ambiguity strategies, we detect no statistically significant difference between LDS-DP and DDM in deceptiveness.

Steps After LDS. Figures 2(c) and 2(f) show post-LDS step distributions. Under deterministic transitions, LDS-Exag achieves the fewest post-LDS steps (17.7), significantly fewer than both DDM-Exag (23.0, $p < 0.01$) and DPP-d1/DPP-d4 (19.8, $p < 0.001$). This result directly confirms the premature-exposure limitation of DPP identified in Section 1: while DPP triggers its switching point too early, leaving a longer truthful phase, LDS-Exag's value-function-based LDS pushes the switching point deeper into the deceptive region, yielding a 10.6% reduction in post-switching steps over DPP and a 23% reduction over DDM-Exag. LDS-Amb also achieves 17.7 steps, shorter than DDM-Amb (25.4). AM has the longest post-LDS phase (41.2 steps), while the Honest baseline reaches 32.7 steps—indicating that even the shortest path can traverse deceptive regions depending on map geometry. Under near-deterministic transitions ($\alpha = 0.9$), LDS-Exag (20.6 steps) is substantially shorter than DDM-Exag (26.7 steps)—a 23% reduction ($p < 0.001$). LDS-Amb (20.4) similarly outperforms DDM-Amb (32.1), while Honest reaches 40.1 steps, showing that the two-phase structure benefits ambiguity-based methods as well.

Robustness analysis. Under $\alpha = 0.9$, Stage 1 reaches d_c^* in 100% of trajectories and Stage 2 always terminates at g_r. In 85% of cases d_c^* coincides with the empirical LDS; the remaining 15% reflect mild slip causing the actual last deceptive state to shift beyond the planned switching point. Re-entry into the deceptive region after d_c^* occurs in approximately 20% of trajectories, consistent with the 10% slip rate. These results confirm that LDS-DP's two-phase decomposition is robust under near-deterministic transitions, with the planned d_c^* serving as a reliable switching point. A sensitivity analysis with respect to γ_a is provided in Section A.1.

Across deterministic and 10%-slip grid worlds, LDS-DP preserves DDM's deception quality while shortening the post-LDS tail. The contrast with DPP is equally instructive. Among the two variants, Exaggeration provides stronger deception at higher cost, while Ambiguity offers a cheaper alternative with moderate deception, giving practitioners a tunable choice between deception strength and trajectory efficiency. Notably, the consistent advantages of LDS-DP across both deterministic and 10%-slip settings demonstrate its robustness to mild stochasticity, confirming that the two-phase decomposition based on LDS remains effective even when transition uncertainty is present.

6 Conclusions and Limitations

We introduced Last Deceptive State (LDS), characterized the value-maximal LDS candidate, and built LDS-DP, a two-phase framework that decomposes deceptive planning into an LP-driven deceptive phase and a shortest-path truthful phase. Experiments show that LDS-DP effectively resolves DPP's premature-exposure problem, while also addressing DDM's wasteful post-LDS detours by switching to the shortest path at d_c^*. LDS-DP also exhibits robustness to mild stochasticity.

Several limitations point to promising directions for future work. This work assumes deterministic transitions, and our $\alpha = 0.9$ results are treated as empirical evidence. Both the LDS concept and DDM rely on the assumption that transition randomness has a limited impact and the discount factor γ is sufficiently large, which underpins the V-difference-based recognition approximation. Characterizing the precise boundary conditions under which this approximation remains valid, and quantifying how deception changes beyond them, is a theoretically challenging direction. Moreover, extending the LDS concept to MDPs with continuous state or time spaces is an intriguing direction. The LDS-DP framework can be used to deceive a rational observer; benign use cases include privacy preservation, while misuse involves coordinated deception.

Reproducibility Statement To facilitate the reproducibility of our work, we have taken the following measures: (1) For the theoretical contributions in Sections 3 and 4, including the definition of the Last Deceptive State (LDS), the value-maximal LDS candidate, and the two-phase LDS-DP framework, we provide complete theorem proofs and detailed algorithmic descriptions in the main text. (2) For the experimental results in Section 5, we include a comprehensive description of the experimental setup, including grid-world environment specifications, MDP transition configurations, observer belief model, baseline methods, hardware environment, solver parameters, and evaluation metrics.

REFERENCES

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT press, 2008.

Shiyue Cao, Pei Xu, Likun Yang, Lei Cui, Shizhao Yu, Shiyu Zhang, Yongjian Ren, Xiaotang Chen, and Kaiqi Huang. Repeated deceptive path planning against learnable observer. *arXiv preprint arXiv:2605.07174*, 2026.

Thomas L Carson. *Lying and deception: Theory and practice*. Oxford University Press, 2010.

John Chelliah and Yogita Swamy. Deception and lies in business strategy. *Journal of Business Strategy*, 39(6):36–42, 2018.

Joao Dias, Ruth Aylett, Ana Paiva, and Henrique Reis. The great deceivers: Virtual agents and believable lies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

Xubin Fang, Brian M Sadler, and Rick S Blum. Receding horizon multi-agent deceptive path planner. *arXiv preprint arXiv:2605.14085*, 2026.

Christopher W Geib and Robert P Goldman. Plan recognition in intrusion detection systems. In *Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01*, volume 1, pp. 46–55. IEEE, 2001.

Irwin Greenberg. The effect of deception on optimal decisions. *Operations Research Letters*, 1(4): 144–147, 1982.

Nolan B Gutierrez, Brian M Sadler, and William J Beksi. Agent deception via polynomial path planning. *Engineering Applications of Artificial Intelligence*, 159:111205, 2025.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Sarah Keren, Avigdor Gal, and Erez Karpas. Privacy preserving plans in partially observable environments. In *IJCAI*, pp. 3170–3176, 2016.

Alan Lewis and Tim Miller. Deceptive reinforcement learning in model-free domains. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 33, pp. 587–595, 2023.

Zhengshang Liu, Yue Yang, Tim Miller, and Peta Masters. Deceptive reinforcement learning for privacy-preserving planning. *arXiv preprint arXiv:2102.03022*, 2021.

Peta Masters and Sebastian Sardina. Cost-based goal recognition for path-planning. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, pp. 750–758, 2017a.

Peta Masters and Sebastian Sardina. Deceptive path-planning. In *IJCAI*, pp. 4368–4375, 2017b.

Peta Masters and Sebastian Sardina. Goal recognition for rational and irrational agents. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, pp. 440–448, 2019.

Shigeo Matsubara and Makoto Yokoo. Negotiations with inaccurate payoff values. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*, pp. 449–450. IEEE, 1998.

Melkior Ornik and Ufuk Topcu. Deception in optimal control. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 821–828. IEEE, 2018.

Miguel Ramírez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, pp. 1121–1126, 2010.

Yagiz Savas, Christos K Verginis, and Ufuk Topcu. Deceptive decision-making under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5332–5340, 2022.

Eric Shieh, Bo An, Rong Yang, Milind Tambe, Craig Baldwin, Joseph DiRenzo, Ben Maule, and Garrett Meyer. Protect: A deployed game theoretic system to protect the ports of the united states. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 1*, pp. 13–20, 2012.

Wesley A Suttle, Jesse Milzman, Mustafa O Karabag, Brian M Sadler, and Ufuk Topcu. Value of information-based deceptive path planning under adversarial interventions. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 3728–3735. IEEE, 2025.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Sun Tzu. The art of war. In *Strategic Studies*, pp. 63–91. Routledge, 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 *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3931–3936. IEEE, 2009.

A Appendix

A.1 Sensitivity to the LP Discount Factor γ_a

We sweep the LP discount factor γ_a over four values $(0.7, 0.8, 0.9, 1.0)$ and compare LDS-DP with DDM under matched values of γ_a, using $\alpha = 1.0$ and $\gamma = 1.0$ together with the coordinate-based Dijkstra cost metric (diagonal cost $\sqrt{2}$) employed in the main text. Considering that the 100×100 maps are very time-consuming, we conduct the sensitivity analysis on Problems 1–24 (49×49).

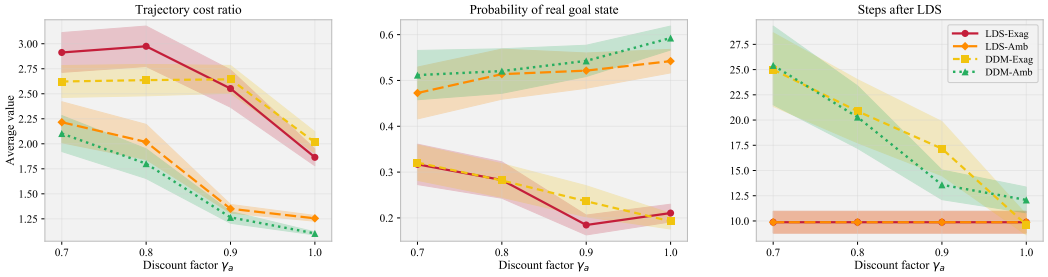


Figure 3: Sensitivity of LDS-DP and DDM to the LP discount factor γ_a on Problems 1–24 ($\alpha = 1.0$). Solid lines: DDM; dashed lines: LDS-DP. Shaded regions show mean $\pm$ one standard error over 24 problems.

Trajectory Cost. The leftmost panel of Figure 3 shows that DDM-Exag is most expensive in the middle of the scanned range, with its cost ratio rising from 2.62 at $\gamma_a = 0.7$ to a broad maximum near 2.64 before falling to 2.01 at $\gamma_a = 1.0$. Although LDS-Exag is somewhat more expensive at small γ_a, its cost decreases as γ_a grows and becomes cheaper than DDM-Exag (1.87 at $\gamma_a = 1.0$). Ambiguity methods are consistently cheaper, and LDS-Amb keeps a lower cost than DDM-Amb at every scanned value.

Deceptiveness. The middle panel shows that both exaggeration methods steadily reduce the real-goal posterior $\Pr(g_r)$ as γ_a increases, and the two curves largely overlap throughout the range, confirming that LDS-DP preserves DDM's deception strength under matched discount factors. For the ambiguity strategy, LDS-Amb keeps the posterior at or below that of DDM-Amb at every scanned value.

Steps After LDS. The rightmost panel reveals the most striking difference. For LDS-Exag, the post-LDS tail stays flat at about 10 steps across the entire scan. By contrast, DDM-Exag starts at roughly 25 steps at $\gamma_a = 0.7$ and decreases steadily, dropping back to about 10 steps only when γ_a approaches 1.0; DDM-Amb shows the same sensitivity. A paired Wilcoxon signed-rank test over the 24 problems confirms that LDS-Exag has a significantly shorter post-LDS tail than DDM-Exag at $\gamma_a \in \{0.7, 0.8, 0.9\}$ (all $p < 0.001$), while the two show no significant difference at $\gamma_a = 1.0$ ($p > 0.05$).

Overall, DDM is highly sensitive to the choice of γ_a, whereas the post-LDS behavior of LDS-DP is stable across the same range. DDM-Exag can match LDS-Exag on post-LDS steps only at the upper extreme ($\gamma_a = 1.0$), and even then it pays a higher trajectory cost (2.01 vs. 1.87, $p = 0.019$) and a slightly lower real-goal posterior (0.193 vs. 0.211, $p = 0.002$).

A.2 Large Language Model Usage Statement

In accordance with the ICLR 2027 policy on the use of large language models (LLMs), we provide the following transparent disclosure regarding the role of LLMs in the preparation of this manuscript.

The authors utilized Deepseek-V4-pro, a large language model developed by DeepSeek, as an assistive tool for improving the quality of this paper and its accompanying code. Specifically, the model was employed to:

1. Enhance the clarity, coherence, and grammatical accuracy of certain paragraphs;
2. Refine technical descriptions for LDS, the LDS-DP two-phase planning framework, and MDP formulations while maintaining precise scientific meaning;
3. Improve the flow and readability of the writing;
4. Assist with formatting and stylistic consistency for the manuscript;
5. Perform code review, syntax checking, and code-style refinement for the supplementary source code, without altering core algorithmic logic.

We explicitly confirm that the model was not involved in any aspects of research ideation, technical innovation, mathematical formulation, algorithm design, core experimental implementation, or result analysis. All scientific contributions, including the proposed Last Deceptive State concept, the LDS-DP two-phase deceptive planning framework, theoretical foundations, experimental design, and conclusions are solely the intellectual product of the human authors. For code artifacts, all core algorithm logic, experiment workflows, and hyper-parameter choices were fully designed and implemented by human authors. The LLM only supported readability improvements and syntax-level checking.

The human authors maintained full oversight and responsibility throughout the writing and code-preparation process. All content generated with model assistance, both manuscript text and code snippets, was rigorously reviewed, validated, and edited to ensure technical accuracy and alignment with the authors' original intent.

We affirm that the authors take complete responsibility for all content of this paper and supplementary code, including any portions that underwent linguistic or stylistic refinement with model assistance. The model was not used to generate any core scientific or algorithmic content, and all key insights and contributions originate from the human authors.

LLM Used: Deepseek-V4-pro
Scope of Use: Manuscript writing assistance and language polishing; code review, syntax checking, and code-style refinement only.