Strategic Communication under Threat: Learning Information Trade-offs in Pursuit-Evasion Games

Valerio La Gatta
Northwestern University
Evanston, United States
valerio.lagatta@northwestern.edu

Dolev Mutzari
Bar Ilan University
Ramat Gan, Israel
dolevmu@gmail.com

Sarit Kraus
Bar Ilan University
Ramat Gan, Israel
sarit@cs.biu.ac.il

V.S. Subrahmanian
Northwestern University
Evanston, United States
vss@northwestern.edu

ABSTRACT

Adversarial environments require agents to navigate a key strategic trade-off: acquiring information enhances situational awareness, but may simultaneously expose them to threats. To investigate this tension, we formulate a Pursuit-Evasion-Exposure-Concealment Game (PEEC) in which a pursuer agent must decide when to communicate in order to obtain the evader's position. Each communication reveals the pursuer's location, increasing the risk of being targeted. Both agents learn their movement policies via reinforcement learning, while the pursuer additionally learns a communication policy that balances observability and risk. We propose SHADOW (Strategic-communication Hybrid Action Decision-making under partial Observation for Wargaming), a multi-headed sequential reinforcement learning framework that integrates continuous navigation control, discrete communication actions, and opponent modeling for behavior prediction. Empirical evaluations show that SHADOW pursuers achieve higher success rates than six competitive baselines. Our ablation study confirms that temporal sequence modeling and opponent modeling are critical for effective decision-making. Finally, our sensitivity analysis reveals that the learned policies generalize well across varying communication risks and physical asymmetries between agents.

KEYWORDS

Pursuit-Evasion Games, Reinforcement Learning, Strategic Communication, Opponent Modeling, Partial Observability

ACM Reference Format:
Valerio La Gatta, Dolev Mutzari, Sarit Kraus, and V.S. Subrahmanian. 2026. Strategic Communication under Threat: Learning Information Trade-offs in Pursuit-Evasion Games. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 16 pages.

1 INTRODUCTION

Intelligent agents operating in adversarial or high-stakes environments such as surveillance, search-and-rescue, or contested terrain, must often manage a fundamental strategic tension: the need to gather information for situational awareness versus the risk of being exposed [10, 26]. This dilemma arises in many real-world scenarios, where communication and sensing actions not only provide critical data about an adversary's position or intent, but also reveal the agent's own presence or location to hostile observers. Decision-making systems that can reason about this trade-off are essential for enabling safe and effective autonomous behavior.

We address this challenge by extending the Pursuit-Evasion-Exposure-Concealment Game (PEEC) [19] where a pursuer seeks to intercept an evader under partial observability. The pursuer can choose to obtain the evader's position, but doing so reveals its own location, potentially aiding the evader's escape or increasing the risk of being eliminated. This PEEC formalizes the dilemma of acting to reduce uncertainty versus remaining covert to reduce risk. The game ends under one of the following conditions: (i) the pursuer captures the evader, i.e., their distance falls below a fixed capture radius, (ii) the pursuer is eliminated, i.e., it is shot with a certain probability when it chooses to query the evader's position, or (iii) the evader escapes, i.e., a fixed time horizon is reached.

Prior work in traditional Pursuit-Evasion Games (PEGs) [25, 53] has addressed partial observability [5, 36–38] and cost-sensitive communication [2, 3, 15, 27, 32], but has rarely considered implicit exposure costs where the act of gathering information can itself be exploited. Our PEEC setup targets this trade-off: silence acts as a protective measure, while communication carries the risk of revealing the pursuer's position. To our knowledge, the only prior work that explicitly models a PEEC setting is [19], which offers a closed-form solution but makes strong simplifying assumptions: First, the game is treated as zero-sum, implying that the pursuer and evader have perfectly symmetric goals—an idealization that rarely holds in realistic settings where the two agents face fundamentally different risks, constraints, and operational objectives. For example, the pursuer may seek to minimize exposure, while the evader aims only to delay capture. Second, the pursuer is assumed to be guaranteed survivability, meaning that it cannot be eliminated even when its position is revealed. Third, the model presumes favorable dynamics for the pursuer, such as higher maneuverability than the evader or noiseless communication channels.

We relax these assumptions and propose **SHADOW** (Strategic-communication Hybrid Action Decision-making under partial Observation for Wargaming), a reinforcement learning (RL) framework

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.

for solving PEECs under realistic asymmetries and nonlinear dynamics[1]. SHADOW learns both a continuous navigation policy and a discrete communication policy, jointly optimized to balance the benefit of acquiring information with the risk of exposure. Crucially, SHADOW agents also include an RL-based opponent modeling predictor to estimate the position of the adversary when the pursuer is not querying the evader's state.

We instantiate a SHADOW-controlled pursuer and evader in a PEEC game. The pursuer strategically queries the evader's position at the cost of being revealed, while the evader learns to evade under uncertainty (without initiating communication). To capture imperfect information exchange, we additionally model noise in the communication channel: when a query occurs, both agents receive the opponent's position corrupted by stochastic noise rather than its exact location.

Our results demonstrate that SHADOW pursuers significantly outperform both static baselines (Random Communication, Periodic Communication [19]) and adaptive RL methods (MultiHead PPO [12], P-DQN [51], HyAR [23], and LIAM [35]), achieving higher success rates with reduced exposure risk. The cross-strategy evaluation across 20 pursuer-evader combinations confirms this advantage holds against diverse evaders. SHADOW pursuers adapt their strategies to varying threat levels and speed disadvantages, and reduce unnecessary communication over time through opponent modeling. Notably, SHADOW and all RL baselines maintain consistent performance even under imperfect communication channels with stochastic observation noise. See Appendix A for illustrative examples of learned strategies.

Our Contribution

(1) **Generalization of PEECs**: We extend PEECs to accommodate non-holonomic and nonlinear dynamics, as well as asymmetric, non-quadratic payoffs.

(2) **SHADOW, an RL Framework for PEECs**: We develop a corresponding RL model, designed to address this expanded class of PEECs. SHADOW employs dynamic opponent modeling to balance information acquisition with the risk of adversarial exposure.

(3) **Cost of Information Acquisition**: We provide the first formal *quantitative definition* of the cost of information acquisition in PEECs. It captures how much the pursuer is willing to pay per query under equilibrium behavior. Assuming zero-sum, we prove a non-negative lower bound, that is used in the experimental section.

(4) **Extensive Experimental Evaluation**: We systematically evaluate SHADOW across several pursuer-evader configurations, analyze learning dynamics, communication strategies, and performance under varying threat levels, agent speeds and communication noise[2]. Our results show that SHADOW pursuers learn to adapt their communication frequency, balance risk and reward, and outperform both periodic and RL-based baselines in pursuit success and efficiency.

[1] RL-based solutions exist for partially observable PEGs with multi-agent coordination [11] and delayed communication [18, 45, 47], but ignore the strategic cost of exposure. To our knowledge, no prior RL method explicitly targets a PEEC game.

[2] Code is available at https://github.com/nsail-lab/SHADOW/

2 RELATED WORK

Pursuit-Evasion Differential Games (PEGs) have long explored how agents operate under uncertainty, particularly in adversarial settings. Prior work models limited observability through three main approaches: (i) *exogenous visibility limits* due to environmental constraints, (ii) *internal sensing costs* that penalize information queries; and (iii) *implicit exposure costs*, where observing reveals the agent's own state to the opponent. While the first two have been extensively studied, implicit exposure remains underexplored. A full survey and comparison of these models is in Appendix I, where we review both classical and modern RL-based PEG formulations. Here, we focus on the third setting as it is both underexplored and central to our work. To our knowledge, the only prior work that explicitly models *implicit exposure cost* is the PEEC framework in [19], where information acquisition comes at the strategic cost of revealing one's own state.[3] Specifically, [19] studies a two-player PEG, where each observation incurs both an explicit sensing cost and an implicit exposure cost, as querying the opponent's state simultaneously discloses the querying agent's position. The study decouples control and sensing decisions, proves the existence of Nash equilibria, and derives a closed-form solution characterized by a periodic "sense–then–hide" policy. While this framework provides a foundational treatment of the exposure–information dilemma, it comes with notable limitations: (i) it is restricted to the LQG setting: while this structure offers analytical tractability, it precludes modeling nonlinear or nonholonomic dynamics that commonly arise in real pursuit–evasion domains, (ii) the framework assumes a strictly zero-sum objective, whereas practical scenarios often involve asymmetric or regularizing costs (e.g., energy use, collision avoidance), and the agents' goals may not be perfectly opposed (e.g., delaying versus capturing), (iii) exposure is treated as a purely strategic cost and does not incorporate the physical risk of elimination that occurs when an agent reveals its position, (iv) the environmental conditions are ideal with higher manueverability for the pursuer and noiseless communication channels.

3 METHODOLOGY

We outline the key modeling challenges (CGs) of PEEC games and present SHADOW, our RL framework designed to address them.

CG_1 *Variable Observation Space*: The observation dimension received from the environment depends on the query action. Without querying, agents perceive only their local state. When queried, both agents receive the full environment state, complicating learning and representation

CG_2 *Non-Markovian Dependencies*: Agents must condition their decisions on past observations and elapsed time. Policies must integrate temporal information and memory to handle delayed effects and shifting strategies.

CG_3 *Parameter Generalization*: The environment is governed by parameters like speed, bounds on acceleration, and capture radius. Policies must adapt across parameter settings without retraining, enabling robust adaptation to new scenarios.

[3] The notion that silence can itself be informative is also explored in [27].

CG$_4$ *Hybrid Action Space*: The agents need to simultaneously execute both a continuous navigation action and a discrete query decision.

CG$_5$ *Strategic Querying*: As for the PEEC formulation, querying the opponent's state is beneficial but risky, as it reveals its own position. The agent must learn to query selectively, weighing informational gain against the cost of exposure.

3.1 SHADOW Architecture

Deriving a closed-form solution to our PEEC game is challenging. We therefore design SHADOW, an RL-based method to learn policies of the players. Figure 1 shows the architecture of a SHADOW pursuer. The *Navigation Module* determines continuous navigation control input u_p at each timestep. The *Query Decision Module* decides whether to query q_p the evader's current position, trading off information gain against potential risk of being discovered or eliminated. The *Opponent Modeling Module* predicts the evader's position $\mathbf{s}'_e$ when no query is made, and is updated via $\mathcal{L}$ when ground-truth observations are available. Each module contains a recurrent *Memory Unit* (e.g., LSTM) to capture temporal dependencies. The *Mediator* integrates all available information (past positions, query outcomes, and timing) into a compact internal state $\tilde{\mathbf{s}}$ that serves as input to both decision modules.

Due to the asymmetric configuration of our PEEC, a SHADOW evader shares the same architecture as the pursuer, except it lacks the *Query Decision Module* as only the pursuer can access the adversary's position.

We now describe each component in greater detail. While the following discussion focuses on the pursuer, the same principles apply to the evader when equipped with a full SHADOW architecture.

Mediator. The *Mediator* translates raw observations from the environment into an internal representation for the agents, addressing CG$_1$. The observation of a pursuer $\mathcal{P}$ facing an evader $\mathcal{E}$ includes: (i) $\mathcal{P}$'s current position $\mathbf{s}_p(t)$, (ii) the elapsed time since the last observation $(t - t_0)$, (iii) $\mathcal{E}$'s last observed position $\mathbf{s}_e(t_0)$, and (iv) the estimated current position of the evader $\mathbf{s}'_e(t)$ and its associated uncertainty $\sigma(t)$, both inferred by the *Opponent Modeling* module. When the pursuer queries the evader's state, the *Mediator* updates the estimated position of the opponent and sets the uncertainty $\sigma(t) = 0$, this allows the different components of SHADOW to implicitly interact. This formulation ensures a consistent structure in the agent's observations, regardless of the pursuer's query decision, while allowing both the pursuer and evader to implicitly assess the reliability of the estimated opponent's state[4]

In addition to positional information, the *Mediator* incorporates key environmental parameters (e.g., the agents' velocities, the capture radius, and the shooting radius) into the internal state $\tilde{\mathbf{s}}$. While these elements assume some prior knowledge of the opponent's capabilities, they also enable agents to generalize across varying scenarios, supporting adaptive policy learning and addressing CG$_3$.

Navigation & Query Decision. The pursuer operates in a hybrid action space involving two components: a binary decision $q_p(t) \in \{0, 1\}$ determining whether to query the evader's position at time t,

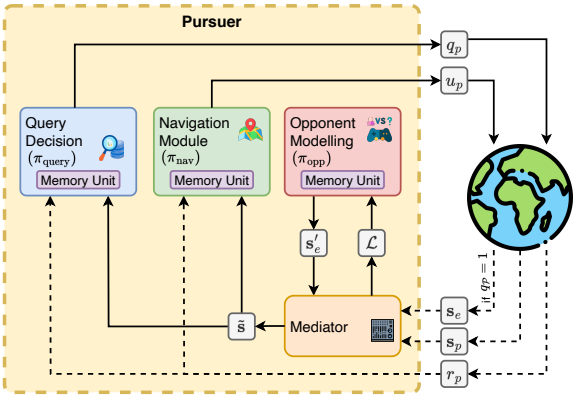


Figure 1: SHADOW Pursuer: The Pursuer operates its navigation control u_p and decides whether to query the opponent's state via a binary action $q_p \in \{0, 1\}$. The environment returns the updated pursuer state $\mathbf{s}_p$ and, if $q_p = 1$, the evader's current position $\mathbf{s}_e$. The *Mediator* determines the pursuer's internal state representation $\tilde{\mathbf{s}}$, comprising: (i) the current position of the pursuer $\mathbf{s}_p$, (ii) the elapsed time since the last observation, (iii) the last observed position of the evader, and (iv) the estimated current position of the evader which is either returned by the environment ($\mathbf{s}_e$), if $q_p = 1$, or inferred by the *Opponent Modeling* module ($\mathbf{s}'_e$) if $q_p = 0$. The *Mediator* also provides feedback $\mathcal{L}$ to the *Opponent Modeling* module, indicating prediction error when the true position of the adversary becomes available ($q_p = 1$). Finally, the pursuer's internal state $\tilde{\mathbf{s}}$ and reward r_p are passed to the *Query Decision* and *Navigation Module* to decide next actions. All networks include a *Memory Unit* (e.g., LSTM) responsible for encoding the temporal observation history.

and a continuous decision $u_p(t)$ controlling its navigation policy. To address this, we use a decoupled learning framework, where separate RL agents are responsible for each decision.

Following [47], we learn the navigation policy π_{nav} using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [13], which is well suited for continuous control tasks. Specifically, the pursuer $\mathcal{P}$ receives its internal state observation $\tilde{\mathbf{s}}$ (processed via the *Mediator*) and outputs an action $u_p = \pi_{\text{nav}}(\tilde{\mathbf{s}})$ to control movement dynamics.

Simultaneously, the pursuer's query policy π_{query} is trained using Proximal Policy Optimization (PPO) [41], a policy-gradient algorithm robust to stochastic discrete actions. At each timestep, $\mathcal{P}$ takes a query decision $q_p = \pi_{\text{query}}(\tilde{\mathbf{s}})$, determining whether to access the opponent's real position.

Both the TD3 and PPO policies leverage a sequence model, specifically an LSTM layer, which acts as a memory unit to encode the temporal observation history. This design addresses challenge CG$_2$, enabling policies to learn from both current and past information[5].

[4] Our Mediator design yields fixed-length representations, enabling sample-efficient standard RL algorithms (TD3, PPO) rather than complex variable-dimensional methods (e.g., attention mechanisms or set encoders [39, 43]). This trades architectural flexibility for training stability.

[5] While we include the elapsed time since last communication in the state of the agents, the LSTM-based memory may also represent previous observations and their effect on agent dynamics.

This modular design offers two key advantages. First, it provides flexibility in handling the hybrid action structure (challenge CG_4) by allowing each sub-policy—navigation and querying—to specialize in its respective action modality. Second, it enables targeted optimization: although the modules are trained independently, they coordinate implicitly through the shared Mediator state. For example, queries provide true positions and thus enable confident manoeuvring. Similarly, the query module observes the consequences of recent manoeuvrers and can trigger communication when distance or uncertainty increase. We compare this design against monolithic baselines that handle hybrid action spaces jointly in the experiments.

Opponent Modeling. The opponent modeling module estimates the position of the opponent and quantifies the associated uncertainty. This information can support both navigation (π_{nav}) and query (π_{query}) decision-making: accurate predictions may enable more effective maneuvering (for both the pursuer and the evader) when direct observations are unavailable, and reduce the need to query when there is confidence in the opponent's estimated position, thus addressing challenge CG_5.

We model this component as a TD3 agent which predicts the evader's position $\mathbf{s}'_e(t)$ and related uncertainty, from $\mathcal{E}$'s last observed state $\mathbf{s}_e(t_0)$ and the time since last observation $t - t_0$. Formally, the agent learns the following policy:

$$\pi_{\text{opp}} : (\mathbf{s}_e(t_0), t - t_0) \to (\mathbf{s}'_e(t), \sigma)$$

where $\sigma \in \mathbb{R}$ is the predicted uncertainty. The model minimizes a Gaussian Negative Log-Likelihood (NLL) loss:

$$\mathcal{L} = \frac{|\mathbf{s}_e(t) - \mathbf{s}'_e(t)|}{2\sigma + \varepsilon} + \frac{1}{2}\log(\sigma + \varepsilon),$$

where the first term penalizes inaccurate predictions (scaled by uncertainty), while the second prevents trivial solutions with overly large σ. A small constant parameter ε ensures numerical stability. Notably, π_{opp} does not receive the opponent current state as input, but it is only used during training to evaluate the loss. The predicted uncertainty σ quantifies reliability and can modulate decisions in π_{nav} and π_{query}. Its effects and relationship with the elapsed time since the last communication ($t - t_0$) are examined in the experiments.

Since both $\mathcal{P}$ and $\mathcal{E}$ adapt their strategies during training (potentially in response to the predictions of the opponent model), it is essential to co-train π_{opp} jointly with the navigation (π_{nav}) and query decision (π_{query}) policies. This ensures mutual adaptation and prevents policy misalignment due to static or outdated opponent predictions, justifying the use of an RL-based agent over a fixed pre-trained model for prediction of opponent state.

4 A CONCRETE PEEC GAME

To evaluate our learning framework, we instantiate a concrete PEEC game in a two-dimensional environment. The game involves a pursuer $\mathcal{P}$ and an evader $\mathcal{E}$ interacting on a bounded planar map $M \subset \mathbb{R}^2$. While the evader moves covertly, the pursuer can choose to query the evader's full state $\mathbf{s}_e$ at the expense of revealing its own state $\mathbf{s}_p$.

State. The state of the game $\mathbf{s} = (\mathbf{s}_p, \mathbf{s}_e)$ consists of the local state of each player. The state of each player $i \in \{\mathcal{P}, \mathcal{E}\}$ is defined as $\mathbf{s}_i = (x_i, y_i, \psi_i) \in M \times [-\pi, \pi)$ and includes their location (x_i, y_i) and heading angle ψ_i.

Dynamics & Actions. Following [22], the dynamics of agent $i \in \{\mathcal{P}, \mathcal{E}\}$ are given by

$$x_i = v_i \cos\psi_i,\ y_i = v_i \sin\psi_i,\ \psi_i = u_i / v_i$$

The pair $(x_i(t), y_i(t)) \in M$ denotes the position of player i at time t, $\psi_i(t)$ is its heading, v_i is the constant velocity of player i, and $u_i(t) \in [-U_i, +U_i]$ is its lateral acceleration, which acts as a control input.

Querying & Observability. In addition to controlling its lateral acceleration, the pursuer can *query* the evader's state by contacting its control unit. We denote by $q_p(t) \in \{0, 1\}$ the binary control variable for querying at time t. When $q_p(t) = 1$, both $\mathcal{P}$ and $\mathcal{E}$ receive information regarding the respective opponent state $\mathbf{s}_i(t)$. Otherwise, the agents only retain their local state $\mathbf{s}_o(t)$. Because real communication channels are rarely perfect, prior research [16, 40] often models imperfect state transmission by adding stochastic noise, modelling occlusions and sensor uncertainty. Following this approach, upon querying the state, each player observes a perturbed position of the opponent, $\mathbf{s}_i(t) + \mathbf{w}_q$, where $\mathbf{w}_q\, \mathcal{N}(0, \boldsymbol{\eta}_q)$ is Gaussian noise.

Game Evolution and Terminal Condition. The game starts at time $t = 0$ from an initial state $\mathbf{s}(0)$. Agents continuously evolve their trajectories by selecting $u_p(t)$ and $u_e(t)$, and the pursuer optionally issues queries via $q_p(t)$. The game terminates at the earliest time $T_f \le T$ when one of the following conditions is met: (i) The pursuer catches the evader, i.e., the Euclidean distance $r(t) = d(\mathcal{P}, \mathcal{E})$ falls below a capture threshold r_c. (ii) The evader survives until the terminal time $t = T$. (iii) The pursuer communicates ($q_p(t) = 1$) and is eliminated with probability $p_e = 2^{-r(t)/r_e}$, where r_e is the evader's shooting radius, i.e., the distance at which the elimination probability equals 50%. Notably, shooting is not modeled as a strategic decision of the evader. We assume that the evader shoots anytime the pursuer reveals its position, but it might miss the target depending on their distance. This assumption is reasonable for resource-rich evaders, which do not have any incentive to withhold fire.

Pursuer Payoff Function. The pursuer's payoff function P_p includes an integral cost over time and a terminal reward:

$$P_p = R_p^f - \int_0^{T_f} (\alpha_p^{\mathsf{T}} \mathbb{1} + \alpha_p^{\mathsf{Q}} \mathbb{1}_{q_{\mathcal{P}}} + \alpha_p^{\mathsf{B}} \mathbb{1}_{\partial M} + \alpha_p^{\mathsf{A}} |u_p|) dt$$

Here, $\alpha_p^{\mathsf{T}}, \alpha_p^{\mathsf{Q}}, \alpha_p^{\mathsf{B}}, \alpha_p^{\mathsf{A}} \ge 0$ are fixed coefficients that determine the cost profile:

- The *time penalty* α_p^{T} encourages faster pursuit [20].
- The *query penalty* α_p^{Q} reflects the cost or risk associated with revealing the pursuer's position.
- The *boundary penalty* α_p^{B} penalizes collisions with the map boundary ∂M, causing physical damage to the UAV.
- The *acceleration penalty* α_p^{A} models energy or resource consumption due to lateral control effort.

The *terminal reward* R_p^f is given by:

$$R_p^f(\mathbf{s}(T_f)) = \begin{cases} r_p, & \mathcal{P} \text{ catches } \mathcal{E} \\ 0, & T_f = T \\ -p_p, & \mathcal{P} \text{ is eliminated} \end{cases}$$

Evader Payoff Function. The evader integral payoff function takes a similar form, except $\alpha_e^{\mathrm{Q}} = 0$ as $\mathcal{E}$ cannot query the state, and $\alpha_e^{\mathsf{T}} = -\alpha_p^{\mathsf{T}} \leq 0$ to promote evasion. In addition, the evader's terminal reward $R_{\mathcal{E}}^{f}(\mathbf{s}_f)$ is negative, i.e., $r_e = -r_p$ in case it gets caught, and zero otherwise. Since shooting is not modeled as a strategic decision, we do not reward the evader when the pursuer is eliminated.

Nash Equilibrium (NE). A pair of control strategies $\langle (u_p^*, q_p^*); u_e^* \rangle$ is an NE if players' payoffs are minimized,

$$(u_p^*, q_p^*) \in \arg\min_{(u_p, q_p)} P_p(\langle (u_p, q_p), u_e^* \rangle; \mathbf{s}_0)$$

$$u_e^* \in \arg\min_{u_e} P_e(\langle (u_p^*, q_p^*), u_e \rangle; \mathbf{s}_0)$$

We denote the set of all NE solutions by Ω_{NE}. It is important to note that we do not assume a system-level payoff $P_S := P_p - P_e$ that one player minimizes and the other maximizes, as in zero-sum settings [33, 47]. While such games admit elegant minimax solutions, they rely on strong assumptions about goal alignment. In our PEEC game, each agent optimizes its own payoff, reflecting potentially conflicting objectives. Even under simplified assumptions, we are not aware of a closed-form NE for the proposed game.

Next, we propose a formal definition for the non-monetary implicit cost of information acquisition in PEEC games.

Definition 1 (Critical Information Acquisition Cost (CIAC))**.** *The* Critical Information Acquisition Cost (CIAC) *is the threshold communication penalty α_c^{Q}, for which there exists an NE where the pursuer obtains a non-negative payoff:*

$$\alpha_c^{\mathrm{Q}} = \sup\{\alpha_p^{\mathrm{Q}} \mid \max_{\Omega_{\mathsf{NE}}[\alpha_p^{\mathrm{Q}}]} \mathbb{E}[P_p | \alpha_p^{\mathrm{Q}}] \geq 0\}$$

By definition, when the communication penalty α_p^{Q} exceeds α_c^{Q}, a pursuer facing a rational evader cannot afford to communicate while ensuring a positive payoff. Conversely, when $\alpha_p^{\mathrm{Q}} < \alpha_c^{\mathrm{Q}}$, there exist a non-trivial querying strategy that yields a positive payoff. Intuitively, one may think of α_c^{Q} as the "*effective*" cost of communication, taking information disclosure and risk of elimination into account.

Proposition 2. *With a zero-sum assumption (i.e., $P_e \equiv -P_p$) and $r_e = 0$, $\alpha_c^{\mathrm{Q}} \geq 0$ is a maximum.*

Intuitively, since $\mathbb{E}[P_p | \alpha_p^{\mathrm{Q}}]$ is linear in α_p^{Q} and so monotonic and continuous, and $\mathbb{E}[P_p | \alpha_p^{\mathrm{Q}} = 0] > -\infty$, α_c^{Q} exists. Furthermore, fixing $\alpha_p^{\mathrm{Q}} < 0$, the pursuer may rapidly and repeatedly query the state sufficiently many times to ensure a positive payoff.

Although CIAC represents the "true" information-acquisition cost, it is computationally challenging to estimate because it requires solving for equilibria across the entire range of communication penalties. For this reason, we introduce a tractable lower bound signal, defined below:

Definition 3 (Base Information Acquisition Cost (CIAC))**.** *Given an NE $\langle (u_p^0, q_p^0); u_e^0 \rangle \in \Omega_{\mathsf{NE}}[\alpha_p^{\mathrm{Q}} = 0]$, the* Base Information Acquisition Cost (CIAC) *is the maximal penalty $\underline{\alpha}_c^{\mathrm{Q}}$ a pursuer is willing to pay per query while ensuring a non-negative payoff:*

$$\underline{\alpha}_c^{\mathrm{Q}} = \frac{\mathbb{E}[P_p \mid \alpha_p^{\mathrm{Q}} = 0]}{\mathbb{E}[N_p^{\mathrm{Q}}]},$$

where N_p^{Q} is the number of pursuer queries.

Proposition 4. *Assuming zero-sum (i.e., $P_e \equiv -P_p$), $\underline{\alpha}_c^{\mathrm{Q}} \leq \alpha_c^{\mathrm{Q}}$.*

Intuitively, since $\langle (u_p^0, q_p^0); u_e^0 \rangle$ is an NE, the evader has no incentive to deviate, and therefore, as long as $\alpha_p^{\mathrm{Q}} < \underline{\alpha}_c^{\mathrm{Q}}$, the pursuer can ensure an expected positive payoff without changing its strategy. Therefore, $\underline{\alpha}_c^{\mathrm{Q}} \leq \alpha_c^{\mathrm{Q}}$.

Formal proofs of Propositions 2,4 are in the Appendix.

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

We instantiate our PEEC game using a SHADOW pursuer following the architecture described above. The *Evader* is also SHADOW-operated but omits the *Query Decision* model π_{query}. As our PEEC formulation is asymmetric, only the pursuer can query the full state of the game. For all experimental settings, models were trained for 20,000 episodes using a mini-batch size of 32, and evaluated on the same $N = 500$ held-out episodes. Unless otherwise specified, we retained the default hyperparameters provided in the original implementations of each algorithm. See Table 2 in Appendix B for full hyper-parameters and hardware details. Statistical significance is assessed using Mann-Whitney U test with FDR correction for multiple hypothesis testing.

5.2 Experimental Protocol

To evaluate SHADOW's effectiveness in learning adaptive query and navigation policies, we designed five experimental tracks: *Baseline Comparison*, *Ablation Study*, *Sensitivity Analysis*, *Uncertainty Dynamics*, and *Training Dynamics Analysis*. A few illustrative examples of game trajectories are provided in Appendix A.

Baseline Comparison We examine whether SHADOW outperforms three heuristic approaches and four RL-based strategies: *(i) No Communication*: the pursuer never communicates. *(ii) Random Communication*: the pursuer uses the inverse probability of getting shot to decide when to query the evader's state, $p_{\text{comm}} = 1 - p_{\text{shot}}$. *(iii) Periodic Communication*: the pursuer communicates periodically, each k timesteps. This strategy was proven to be theoretically optimal in the setting of [19][6]. *(iv) MultiHead PPO* [12]: the pursuer leverages a multi-headed actor with PPO to jointly learn the query and navigation policies. *(v) P-DQN* [51]: the pursuer leverages a Parametrized Deep Q-Network to jointly learn the query and navigation policies. *(vi) HyAR* [23]: the pursuer learns the relationship between the discrete action (q_p) and the continuous action (u_p) using a variational autoencoder. *(vii)* LIAM [35]: each agent learns a model of its opponent through an encoder–decoder architecture that reconstructs the opponent's position from its own partial observation. Full details on baselines configuration are in Appendix C. We do not include multi-agent reinforcement learning methods (e.g., MAPPO [54]) because these baselines are designed for scenarios with *multiple* physical agents with shared objectives. In our PEEC, the pursuer is a *single* physical agent whose decision-making we decompose into functional modules (navigation and query), not separate entities.

[6] [19] proposed a simplified PEEC game with strong assumptions, including symmetric agent goals (i.e., zero-sum formulation), guaranteed pursuer survivability (i.e., the pursuer cannot be eliminated when discovered), and favorable dynamics (e.g., higher maneuverability for the pursuer).

Model	End-State Outcomes			Communication Strategy			Behavioral Efficiency		
	P_{win}	P_{shot}	P_{timeout}	C_{ratio}	C_{gap}	D_{comm}	T_{len}	$\bar{S}_P$	$\bar{S}_E$
No communication	0.184 ± 0.034	N/A	0.816 ± 0.034	N/A	N/A	N/A	143.2 ± 44.37	0.192 ± 0.003	0.233 ± 0.012
Random communication	0.020 ± 0.012	0.980 ± 0.012	0.000 ± 0.000	0.492 ± 0.014	2.018 ± 0.054	0.194 ± 0.012	32.82 ± 2.149	0.149 ± 0.013	0.313 ± 0.019
Periodic (k=5)	0.182 ± 0.033	0.818 ± 0.033	0.000 ± 0.000	0.204 ± 0.003	5	0.149 ± 0.010	43.46 ± 2.363	0.126 ± 0.009	0.223 ± 0.010
Periodic (k=10)	0.264 ± 0.038	0.736 ± 0.038	0.000 ± 0.000	0.104 ± 0.003	10	0.143 ± 0.010	57.05 ± 2.819	0.112 ± 0.007	0.198 ± 0.011
Periodic (k=20)	0.480 ± 0.043	0.520 ± 0.043	0.000 ± 0.000	0.056 ± 0.003	20	0.159 ± 0.011	86.37 ± 4.857	0.125 ± 0.006	0.181 ± 0.009
Periodic (k=30)	0.546 ± 0.041	0.439 ± 0.043	0.015 ± 0.001	0.039 ± 0.004	30	0.191 ± 0.013	119.2 ± 6.361	0.159 ± 0.006	0.191 ± 0.009
Periodic (k=40)	0.576 ± 0.052	0.416 ± 0.036	0.010 ± 0.008	0.033 ± 0.004	40	0.206 ± 0.012	320.1 ± 22.93	0.168 ± 0.005	0.123 ± 0.005
Periodic (k=50)	0.276 ± 0.039	0.274 ± 0.039	0.450 ± 0.043	0.026 ± 0.004	50	0.244 ± 0.011	283.3 ± 29.01	0.162 ± 0.005	0.121 ± 0.005
MultiHead PPO	0.272 ± 0.039	0.056 ± 0.020	0.672 ± 0.041	0.066 ± 0.019	10.32 ± 0.020	0.504 ± 0.021	258.5 ± 39.85	0.162 ± 0.001	0.185 ± 0.012
P-DQN	0.396 ± 0.043	0.042 ± 0.017	0.564 ± 0.043	0.046 ± 0.001	62.13 ± 5.546	0.448 ± 0.018	441.1 ± 37.59	0.196 ± 0.012	0.244 ± 0.013
HyAR	0.246 ± 0.037	0.002 ± 0.003	0.752 ± 0.037	0.002 ± 0.001	55.01 ± 648.1	0.686 ± 0.065	52.94 ± 8.399	0.897 ± 0.012	0.238 ± 0.011
LIAM	0.570 ± 0.044	0.428 ± 0.043	0.002 ± 0.003	0.069 ± 0.015	32.19 ± 4.027	0.197 ± 0.014	165.9 ± 12.73	0.254 ± 0.013	0.393 ± 0.005
SHADOW (Ours)	**0.620 ± 0.042**	0.350 ± 0.041	0.030 ± 0.015	0.172 ± 0.023	29.42 ± 4.988	0.240 ± 0.014	272.8 ± 20.04	0.168 ± 0.011	0.231 ± 0.017

Table 1: Baseline Comparison: SHADOW against seven baselines. All experiments use a SHADOW-controlled evader, while the pursuer's strategy varies across rows. P_{win} is our primary performance measure. The others are included solely for analysis.

Performance is assessed via metrics in three categories: (i) *End-State Outcomes* includes the percentage of evaluation episodes the pursuer wins P_{win}, gets shot P_{shot} or runs out of time P_{timeout}; (ii) *Communication Strategy* includes the average percentage of communication events C_{ratio}, the average time between queries C_{gap}, the average distance between agents at the last communication D_{comm}, and CIAC; and (iii) *Behavioral Efficiency* includes the average episode duration T_{len} and steering costs $\bar{S}_P, \bar{S}_E$. See Appendix D for full metrics details. Across all experiments, the primary performance metric used for hyperparameter tuning and baseline comparison is P_{win}.

Ablation Study We assess SHADOW's key components by systematically removing the *Opponent Modeling* module and the LSTM-based *Memory Unit*. For opponent modeling, we test four configurations in which either agent, both, or neither use the module. We similarly toggle the LSTM layer for both agents. This enables us to assess how temporal modeling contributes to effective movement.

Sensitivity Analysis We measure SHADOW's performance under varying environmental conditions, i.e., the shooting radius r_{e}, the speed ratio v_e/v_p, and the communication noise $\boldsymbol{\eta}_q$.

Training & Uncertainty Dynamics We investigate how the pursuer and evader strategies evolve during training by tracking their metrics over time. In addition, we analyze how the uncertainty σ predicted by the opponent modeling module influences the pursuer's communication decisions.

5.3 Baseline Comparison

Table 1 reports the performance of SHADOW compared to the 7 baselines. All results assume that the evader is controlled by SHADOW, while the pursuer's strategy varies across baselines. *SHADOW consistently outperforms competitors from the pursuer's perspective, achieving the highest win rate P_{win} = 62%.*

SHADOW vs. Periodic Communication. The Periodic Communication strategy with $k = 40$ achieves the second-highest win rate (P_{win} = 57.6%), trailing SHADOW by 7.1%. This difference is statistically significant, FDR-corrected $p = 0.013$. Furthermore, the Periodic Communication with $k = 40$ incurs a substantially higher probability of being shot (P_{shot} = 41.6%) compared to SHADOW (P_{shot} = 35%, a 15.8% reduction, FDR-corrected $p = 0.031$). Additionally, it results in longer episodes (T_{len} = 320.1 vs. 272.8 for SHADOW, a 14.7% decrease, FDR-corrected $p = 0.034$) and a smaller average communication distance (D_{comm} = 0.206 vs. 0.240 for SHADOW, a 14.2% increase, FDR-corrected $p = 1.85 \times 10^{-4}$), indicating that the agent tends to communicate at closer ranges, potentially increasing risk of being shot.

Periodic strategies also exhibit higher sensitivity to interval choice: increasing from $k = 40$ to $k = 50$ causes P_{win} to collapse from 57.6% to 27.6%, while P_{timeout} surges from 1% to 45%. This abrupt transition reflects a threshold effect—the additional 10 timesteps of silence provide sufficient time for the evader to escape beyond interception range before the next query. Such brittleness underscores a key advantage of adaptive communication policies like SHADOW, which dynamically modulate query timing based on uncertainty and distance rather than fixed intervals. Finally, SHADOW induces a higher average steering cost on the evader ($\bar{S}_E$ = 0.231) compared to all periodic strategies (FDR-corrected $p > 3.81 \times 10^{-72}$). This behavior depends on SHADOW's higher communication frequency (C_{ratio} = 17.2%) forcing the evader into more evasive maneuvers.

SHADOW vs. RL Baselines. Most RL-based baselines, i.e., MultiHead PPO, P-DQN and HyAR, learn more conservative pursuer behaviors. These agents communicate far less than SHADOW (C_{ratio} = 6.6%, 4.6% and 0.2%, respectively), resulting in lower probability of being shot (P_{shot} = 5.6%, 4.2% and 0.2%, respectively). However, this comes at the cost of the win rate (P_{win} = 27.2%, 39.6%, and 24.6%, respectively). By contrast, LIAM, similar to SHADOW, exhibits the opposite trade-off as it adopts an aggressive querying strategy[7]: it achieves a competitive win rate (P_{win} = 57.0%), but at the cost of a substantially higher shooting probability (P_{shot} = 42.8%). This similarity between SHADOW and LIAM depends on their common design, i.e., these methods differ in the opponent modelling strategy but share the same architecture of the query decision module and the navigation module. However, SHADOW still outperforms LIAM by 8% (P_{win} = 62% vs. 57%) while maintaining a relatively safer behaviour (35% vs 42.8% shooting probability, a 18.2% reduction).

Cross-Strategy Robustness. We evaluate robustness through a pairwise comparison between pursuer and evader strategies. The pursuer

[7] Although this behaviour maximizes the pursuer's capture rate P_{win}, both SHADOW and LIAM could learn "safer" solutions by appropriately tuning the terminal reward R_p^f to prioritize survival over elimination.

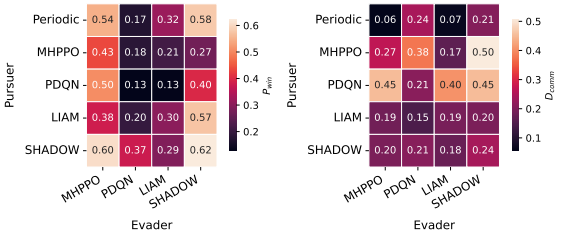


Figure 2: Baseline Comparison: Performance in terms of (a) pursuer win rate P_{win} and (b) average distance at last communication D_{comm}, across different pursuer-evader combinations.

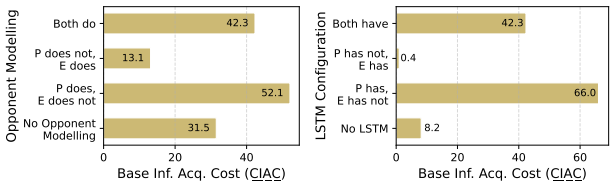


Figure 3: Ablation Study: CIAC under different Opponent Modeling (a) and LSTM (b) configurations.

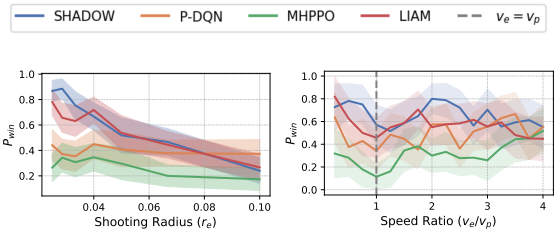


Figure 4: Sensitivity analysis: Effect of the shooting radius r_{e} (a) and the speed ratio v_e/v_p (b) on pursuer win rate P_{win}.

is selected from Periodic, MHPPO, PDQN, LIAM, and SHADOW, while the evader is drawn from the same set except for Periodic, which applies only to the pursuer since it is the sole agent capable of communication. Figures 2a and 2b present pursuer win rate P_{win} and average distance at last communication D_{comm}, respectively, for each pursuer-evader pairing. Two key findings emerge from this analysis. First, SHADOW pursuers consistently achieve the highest P_{win} across most evader types: 60% against MHPPO evaders, 37% against PDQN evaders, and 62% against SHADOW evaders (as also shown in Table 1). The only exception occurs against LIAM evaders, where SHADOW achieves 29% win rate, slightly below LIAM (30%) and pursuers enacting periodic communication (32%). We also observe that LIAM and PDQN evaders are more challenging opponents—indeed, Figure 2a also reveals that all pursuers achieve their lowest win rates when facing PDQN and LIAM evaders.

Second, examining D_{comm} in Figure 2b, SHADOW and LIAM pursuers demonstrate similar behaviors in D_{comm} across different opponents: SHADOW varies only from 0.18 (against LIAM evaders) to 0.24 (against SHADOW evaders), while LIAM ranges from 0.15 to 0.20. In contrast, other pursuers exhibit larger variability: for example, Periodic pursuers maintain $D_{\text{comm}} = 0.06 - 0.07$ against MHPPO and LIAM evaders but jump to 0.21 against SHADOW. This variability reveals that fixed-interval policies cannot strategically select communication distances—they query at predetermined times regardless of opponent position or strategy. Conversely, SHADOW maintains consistent communication distances by adaptively timing queries based on game state.

5.4 Ablation Study

Figure 3a reports CIAC under 4 opponent modeling configurations. Equipping either agent with an opponent modeling module consistently yields a strategic advantage for that agent. When only the pursuer uses opponent modeling (*P does, E does not*), CIAC reaches 52.1—the highest observed value, representing a substantial increase over the baseline of 31.5 when neither agent models its opponent (*No Opponent Modeling*). Conversely, when only the evader uses opponent modeling, CIAC drops to 13.1, a notable decrease compared to 42.3 when both agents use opponent modeling (*Both do*).

These results suggest that a pursuer that uses opponent modeling learns to communicate more strategically, accepting communication costs to gain higher returns. This holds regardless of the evader's configuration. The gain stems from two complementary factors: (i) less frequent but more effective communication, and (ii) increased likelihood of catching the evader. For instance, when the evader lacks opponent modeling, the pursuer's P_{win} rises from 47.2% to 62% (FDR-corrected $p = 2.60 \times 10^{-6}$), while C_{ratio} drops from 24% to 15% (FDR-corrected $p = 4.18 \times 10^{-5}$), and C_{gap} grows from 10.13 to 38.73 ($p = 4.98 \times 10^{-10}$). See Appendix E for more details.

A similar pattern emerges when applying opponent modeling to baselines powered by MultiHead PPO and P-DQN: in asymmetric settings, the agent with opponent modeling consistently outperforms its counterpart, while in the symmetric setting (where both agents use opponent modeling), the pursuer maintains a strategic edge. See Figure 8 in Appendix E for more details.

Finally, equipping either agent with an LSTM memory improves its performance. As shown in Figure 3b, equipping the pursuer with an LSTM yields a pursuer willing to pay more for information regarding the evader. For instance, when neither agent uses memory, CIAC = 8.2. This value rises sharply to CIAC = 66.0 when only the pursuer uses the LSTM. This improvement is primarily driven by a large increase in the pursuer's success rate, with P_{win} jumping from 11.2% to 88.8% (+87.3%, FDR-corrected $p = 3.73 \times 10^{-79}$). See Figure 7c, 7d in Appendix E for full details.

5.5 Sensitivity Analysis

Shooting Radius. We examine how the shooting radius r_{e} affects game's outcome. Figure 4a shows the pursuer win rate P_{win} as a function of r_{e}. SHADOW exhibits a clear decreasing trend: as communication becomes riskier (i.e., higher r_{e}), the pursuer's performance drops significantly. This is confirmed by regression analysis, where the (negative) effect size is substantial and highly significant

($\beta = -28.3$, FDR-corrected $p = 1.45 \times 10^{-15}$, $R^2 = 0.144$). Similar results hold for LIAM: $\beta = -7.45$, FDR-corrected $p = 6.82 \times 10^{-4}$, $R^2 = 0.094$. In contrast, P-DQN and MultiHead PPO do not adapt their communication strategy to changing risk: the same regression does not yield statistically significant results ($p > 0.10$).

We further investigate how r_e modulates communication behavior. For SHADOW, an increase in r_e leads to lower communication frequency (C_{ratio}) and shorter communication gaps (D_{comm}) from the shooting radius r_e. Specifically, when we fit a regression model to infer C_{ratio} and D_{comm}, we obtain the following results: *(i)* For C_{ratio}: $\beta = -0.886$, FDR-corrected $p = 0.010$, $R^2 = 0.027$. *(ii)* For D_{comm}: $\beta = 1.60$, FDR-corrected $p = 4.82 \times 10^{-6}$, $R^2 = 0.074$. In contrast, P-DQN, MHPPO and LIAM show no significant relationship between r_e and either C_{ratio} or D_{comm} ($p > 0.59$). This indicates that their policies are insensitive to variations in communication risk and fail to optimize the communication-efficiency trade-off as SHADOW does. See Appendix F.1 for further details.

Speed Ratio. We assess how the relative speed of the evader and pursuer, expressed as the ratio v_e/v_p, affects the win rate P_{win} for SHADOW, P-DQN, MHPPO, and LIAM, as shown in Figure 4b. We do not observe any trend: across all models, the regression results indicate no significant relationship between P_{win} and the speed ratio (FDR-corrected $p > 0.2$). Despite this, we find that the speed ratio influences communication behavior: for SHADOW and MHPPO, higher speed ratio (i.e., faster evaders) lead to increased CIAC ($\beta = 7.48$, FDR-corrected $p = 0.018$, $R^2 = 0.372$ for SHADOW; $\beta = 20.44$, FDR-corrected $p = 0.005$, $R^2 = 0.518$ for MHPPO), indicating that slower pursuers increasingly rely on communication to compensate the speed difference (see Figure 10b, Appendix F.2).

Communication Noise. We evaluate the effect of communication noise $\boldsymbol{\eta}_q$ on both the pursuer win rate P_{win} and CIAC for SHADOW, P-DQN, MHPPO and LIAM. Regardless of the method, neither metric exhibits a significant trend as noise increases ($p > 0.7$), see Appendix F.3 for full results. This suggests that noisier communication channels do not necessarily degrade the pursuer's ability to capture the evader.

5.6 Training & Uncertainty Dynamics

We analyze the evolution of the pursuer's strategy during training by examining outcome and communication metrics across episodes. Figure 5 reports the trends of P_{win}, P_{shot}, P_{timeout}, C_{ratio}, and C_{gap} during training. We identify three phases (shaded regions) of training. During an initial phase (up to episode 7,500), the pursuer rapidly learns to reduce the risk of being shot. This is reflected by a steep decline in P_{shot} and C_{ratio}, showing that the agent quickly recognizes the dangers of communication.

During the intermediate phase (episodes 7,500 to 14,500), the agent enters a transient regime: the pursuer frequently times out as the evader learns an effective movement policy, i.e., its reward increases steeply (see Figure 13 in Appendix H for further details). In this phase, the pursuer is actively experimenting with different communication strategies, i.e., while C_{ratio} stabilizes around 15%, C_{gap} exhibits substantial fluctuations between 2 and 35.

In the final phase (after episode 14,500), P_{win} improves and stabilizes above 60%. This improvement is *not* accompanied by a change in C_{ratio}, which remains steady. Instead, the key adaptation occurs

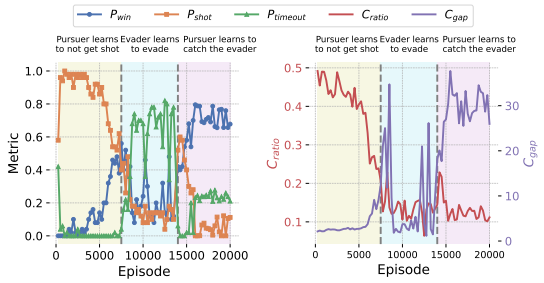


Figure 5: Training dynamics: Evaluation metrics over training episodes. Shaded regions corresponds to distinct learning phases.

in C_{gap}, which stabilizes at a value of 30, indicating that the pursuer has learned to distribute evader's queries more strategically.

At the end of the training process, we analyze how uncertainty σ predicted by the opponent modeling module influences communication decisions. We find that σ increases systematically in the timesteps leading up to a communication event ($q_p = 1$): $\beta = 0.012$, FDR-corrected $p = 3.81 \times 10^{-16}$, while no such trend is observed before non-communication actions ($q_p = 0$). This suggests that the pursuer monitors the reliability of its internal predictions and initiates communication when uncertainty accumulates. Furthermore, regression analysis (full details in Appendix G) demonstrates that both σ and the elapsed time since last communication ($t - t_0$) independently predict the error of the opponent modeling module, suggesting they capture complementary aspects of uncertainty.

6 CONCLUSION

We generalized the PEEC framework proposed in [19] to model the strategic tension between information gathering and concealment in adversarial settings, and proposed **SHADOW**, an RL-based approach tailored to this challenge. Our formulation allows for asymmetric dynamics, non-zero-sum objectives, and realistic exposure risks where agents may be eliminated upon discovery. Empirical results show that SHADOW outperforms seven baselines across a range of threat levels and environmental dynamics. Moreover, SHADOW pursuers adaptively modulate communication based on communication risk, leveraging predicted uncertainty to query more when uncertainty is high and remain silent when confidence is low.

Limitations. Like any study, ours has some limitations. First, our experiments focus on single pursuer-evader scenarios, extending SHADOW to multi-agent settings would require addressing coordination protocols and collective exposure. Second, although we relax many assumptions from prior PEEC work [19] on implicit exposure cost (i.e., zero-sum objectives, guaranteed pursuer survivability, and noiseless communication), our environment remains simplified compared to real-world applications. Incorporating 3D dynamics and spatial obstacles that constrain movement and observability represent important avenues for future research.

ACKNOWLEDGEMENTS

This work was partly funded by Army Research Organization Grant W911NF2320240 and by the Israel Science Foundation under grant 2544/24.

REFERENCES

[1] Shubham Aggarwal, Tamer Başar, and Dipankar Maity. 2024. Best response strategies for asymmetric sensing in linear-quadratic differential games. *arXiv preprint arXiv:2406.05632* (2024).

[2] Shubham Aggarwal, Tamer Başar, and Dipankar Maity. 2024. Linear quadratic zero-sum differential games with intermittent and costly sensing. *IEEE Control Systems Letters* (2024).

[3] Saad A Aleem, Cameron Nowzari, and George J Pappas. 2015. Self-triggered pursuit of a single evader with uncertain information. *arXiv preprint arXiv:1512.06184* (2015).

[4] Claude Berge. 1997. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity*. Courier Corporation.

[5] Pierre Bernhard and A-L Colomb. 1988. Saddle point conditions for a class of stochastic dynamical games with imperfect information. *IEEE Trans. Automat. Control* 33, 1 (1988), 98–101.

[6] Craig J. Bester, Steven D. James, and George D. Konidaris. 2019. Multi-Pass Q-Networks for Deep Reinforcement Learning with Parameterised Action Spaces. *arXiv preprint arXiv:1905.04388* (2019). arXiv:1905.04388 http://arxiv.org/abs/1905.04388

[7] Sourabh Bhattacharya and Seth Hutchinson. 2010. On the existence of nash equilibrium for a two-player pursuit—evasion game with visibility constraints. *The International Journal of Robotics Research* 29, 7 (2010), 831–839.

[8] Shaunak D Bopardikar, Francesco Bullo, and Joao P Hespanha. 2008. On discrete-time pursuit-evasion games with sensing limitations. *IEEE Transactions on Robotics* 24, 6 (2008), 1429–1439.

[9] Francesco Borra, Luca Biferale, Massimo Cencini, and Antonio Celani. 2022. Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number. *Physical Review Fluids* 7, 2 (2022), 023103.

[10] Shaofei Chen, Feng Wu, Lincheng Shen, Jing Chen, and Sarvapali D Ramchurn. 2015. Multi-agent patrolling under uncertainty and threats. *PloS one* 10, 6 (2015), e0130154.

[11] Cristino De Souza, Rhys Newbury, Akansel Cosgun, Pedro Castillo, Boris Vidolov, and Dana Kulić. 2021. Decentralized multi-agent pursuit using deep reinforcement learning. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4552–4559.

[12] Yannis Flet-Berliac and Philippe Preux. 2019. Merl: Multi-head reinforcement learning. *arXiv preprint arXiv:1909.11939* (2019).

[13] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*. PMLR, 1587–1596.

[14] Autonomous Agents Research Group. 2022. Local Information Agent Modelling (LIAM). https://github.com/uoe-agents/LIAM.

[15] Abhishek Gupta, Ashutosh Nayyar, Cédric Langbort, and Tamer Basar. 2014. Common information based Markov perfect equilibria for linear-Gaussian games with asymmetric information. *SIAM Journal on Control and Optimization* 52, 5 (2014), 3228–3260.

[16] G. Hexner, I. Rusnak, and H. Weiss. 2019. A Pursuit-Evasion Game with Incomplete Information. In *2019 27th Mediterranean Conference on Control and Automation (MED)*. 583–588. https://doi.org/10.1109/MED.2019.8798566

[17] György Hexner, Ilan Rusnak, and Haim Weiss. 2019. A pursuit-evasion game with incomplete information. In *2019 27th Mediterranean Conference on Control and Automation (MED)*. IEEE, 583–588.

[18] Penglin Hu, Quan Pan, Chunhui Zhao, and Yaning Guo. 2024. Transfer reinforcement learning for multi-agent pursuit-evasion differential game with obstacles in a continuous environment. *Asian Journal of Control* 26, 4 (2024), 2125–2140.

[19] Yunhan Huang and Quanyan Zhu. 2022. A Pursuit-Evasion Differential Game with Strategic Information Acquisition. arXiv:2102.05469 [eess.SY] https://arxiv.org/abs/2102.05469

[20] Chiraz Ben Jabeur, Hassene Seddik, Khaled Khnissi, and Ahmad Hably. 2025. Robotic pursuit evasion problem in a constrained game area using deep reinforcement learning and self-play training. https://www.researchsquare.com/article/rs-6279213/v1. Preprint on Research Square. DOI: https://doi.org/10.21203/rs.3.rs-6279213/v1.

[21] Ilya Kostrikov. 2018. PyTorch Implementations of Reinforcement Learning Algorithms. https://github.com/ikostrikov/pytorch-a2c-ppo-acktr-gail.

[22] Mangal Kothari, Joel G Manathara, and Ian Postlethwaite. 2014. A cooperative pursuit-evasion game for non-holonomic systems. *IFAC Proceedings Volumes* 47, 3 (2014), 1977–1984.

[23] Boyan Li, Hongyao Tang, Yan Zheng, Jianye Hao, Pengyi Li, Zhen Wang, Zhaopeng Meng, and Li Wang. 2021. HyAR: Addressing Discrete-Continuous Action Reinforcement Learning via Hybrid Action Representation. *CoRR* abs/2109.05490 (2021). arXiv:2109.05490 https://arxiv.org/abs/2109.05490

[24] Efrat Sless Lin, Noa Agmon, and Sarit Kraus. 2019. Multi-robot adversarial patrolling: Handling sequential attacks. *Artificial Intelligence* 274 (2019), 1–25.

[25] Viliam Lisỳ, Branislav Bošanskỳ, and Michal Pěchouček. 2012. Anytime algorithms for multi-agent visibility-based pursuit-evasion games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. 1301–1302.

[26] Jiazhen Liu, Peihan Li, Yuwei Wu, Gaurav S Sukhatme, Vijay Kumar, and Lifeng Zhou. 2024. Multi-robot target tracking with sensing and communication danger zones. *arXiv preprint arXiv:2404.07880* (2024).

[27] Dipankar Maity. 2023. Efficient communication for pursuit-evasion games with asymmetric information. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2104–2109.

[28] Dipankar Maity. 2023. Optimal intermittent sensing for pursuit–evasion games. *IEEE Control Systems Letters* 7 (2023), 3000–3005.

[29] Dipankar Maity, Achilleas Anastasopoulos, and John S Baras. 2017. Linear quadratic games with costly measurements. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 6223–6228.

[30] Dipankar Maity and John S Baras. 2016. Strategies for two-player differential games with costly information. In *2016 13th International Workshop on Discrete Event Systems (WODES)*. IEEE, 211–216.

[31] Dipankar Maity, Aneesh Raghavan, and John S Baras. 2017. Stochastic differential linear-quadratic games with intermittent asymmetric observations. In *2017 American Control Conference (ACC)*. IEEE, 3670–3675.

[32] Dipankar Maity, Alexander Von Moll, Daigo Shishika, and Michael Dorothy. 2024. Optimal Evasion from a Sensing-Limited Pursuer. In *2024 American Control Conference (ACC)*. IEEE, 2758–2765.

[33] Jianjun Ni, Liu Yang, Liuying Wu, and Xinnan Fan. 2018. An improved spinal neural system-based approach for heterogeneous AUVs cooperative hunting. *International Journal of Fuzzy Systems* 20 (2018), 672–686.

[34] Yaniv Oshrat, Noa Agmon, and Sarit Kraus. 2020. Adversarial fence patrolling: Non-uniform policies for asymmetric environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10377–10384.

[35] Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. 2021. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 19210–19222.

[36] Alberto Quattrini Li, Raffaele Fioratto, Francesco Amigoni, and Volkan Isler. 2018. A search-based approach to solve pursuit-evasion games with limited visibility in polygonal environments. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1693–1701.

[37] Eric Raboin, Ugur Kuter, and Dana Nau. 2012. Generating strategies for multi-agent pursuit-evasion games in partially observable euclidean space. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. 1201–1202.

[38] I Rhodes and D Luenberger. 1969. Differential games with imperfect state information. *IEEE Trans. Automat. Control* 14, 1 (1969), 29–38.

[39] Sasha Salter, Dushyant Rao, Markus Wulfmeier, Raia Hadsell, and Ingmar Posner. 2021. Attention-Privileged Reinforcement Learning. In *Proceedings of the 2020 Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 155)*, Jens Kober, Fabio Ramos, and Claire Tomlin (Eds.). PMLR, 394–408. https://proceedings.mlr.press/v155/salter21a.html

[40] L. Schenato, Songhwai Oh, S. Sastry, and P. Bose. 2005. Swarm Coordination for Pursuit Evasion Games using Sensor Networks. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. 2493–2498. https://doi.org/10.1109/ROBOT.2005.1570487

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[42] Daigo Shishika, James Paulos, and Vijay Kumar. 2020. Cooperative team strategies for multi-player perimeter-defense games. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2738–2745.

[43] Viacheslav Sinii, Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, and Sergey Kolesnikov. 2024. In-context reinforcement learning for variable action spaces. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1862, 21 pages.

[44] John von Neumann. 1959. Mathematische Annalen 100 (1928), pp. *Contributions to the Theory of Games* 100, 40 (1959), 13.

[45] Xiaoxiao Wang, Peng Yi, and Yiguang Hong. 2025. A Hierarchical Deep Reinforcement Learning Strategy for Collective Pursuit-Evasion Game with Partial Observations. *IEEE Transactions on Artificial Intelligence* (2025).

[46] Wei Wei, JingJing Wang, Jun Du, Zhengru Fang, Chunxiao Jiang, and Yong Ren. 2022. Underwater differential game: Finite-time target hunting task with communication delay. In *ICC 2022-IEEE International Conference on Communications*. IEEE, 3989–3994.

[47] Wei Wei, Jingjing Wang, Jun Du, Zhengru Fang, Yong Ren, and CL Philip Chen. 2023. Differential Game-Based Deep Reinforcement Learning in Underwater Target Hunting Task. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[48] Isaac E Weintraub, Meir Pachter, and Eloy Garcia. 2020. An introduction to pursuit-evasion differential games. In *2020 American Control Conference (ACC)*. IEEE, 1049–1066.

[49] Axing Xi and Yuanli Cai. 2022. A nonlinear finite-time robust differential game guidance law. *Sensors* 22, 17 (2022), 6650.

[50] Axing Xi and Yuanli Cai. 2024. Deep Reinforcement Learning-Based Differential Game Guidance Law against Maneuvering Evaders. *Aerospace (MDPI Publishing)*

11, 7 (2024).
[51] Jiechao Xiong, Qing Wang, Zhuoran Yang, Peng Sun, Lei Han, Yang Zheng, Haobo Fu, Tong Zhang, Ji Liu, and Han Liu. 2018. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space. *arXiv preprint arXiv:1810.06394* (2018).
[52] Can Xu, Yin Zhang, Weigang Wang, and Ligang Dong. 2022. Pursuit and evasion strategy of a differential game based on deep reinforcement learning. *Frontiers in Bioengineering and Biotechnology* 10 (2022), 827408.
[53] Fuhan Yan, Jiuchuan Jiang, Kai Di, Yichuan Jiang, and Zhifeng Hao. 2019. Multiagent pursuit-evasion problem with the pursuers moving at uncertain speeds. *Journal of Intelligent & Robotic Systems* 95, 1 (2019), 119–135.
[54] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of PPO in cooperative multi-agent games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1787, 14 pages.
[55] Zixu Zhang and Jaime F Fisac. 2021. Safe occlusion-aware autonomous driving via game-theoretic active perception. *arXiv preprint arXiv:2105.08169* (2021).

A EXAMPLES

Figure 6 shows four representative episodes in which the pursuer successfully captures the evader. In all cases, both agents are controlled using the SHADOW framework. These examples are selected to highlight the diversity and adaptiveness of the learned strategies, particularly the non-trivial communication behavior of the pursuer.

A consistent observation across all examples is that the pursuer's communication strategy is not periodic. Initially, the pursuer typically attempts to intercept the evader by exploring the surrounding region near its starting position. If this initial search proves unfruitful, the pursuer triggers its first communication request to obtain an updated position of the evader. We note that in most cases, the pursuer does not initiate communication at the beginning of the episode. This depends on the assumption that the initial position of the evader is known to the pursuer, and vice-versa. The only exception is the example in Figure 6d, where the pursuer starts on the extreme edge of the map with a heading directed toward the boundary. Given that the environment does not support toroidal wrapping, this configuration likely prompted an early communication request to avoid an inefficient trajectory.

Furthermore, a recurring behavioral motif is the looping or circular movement of the pursuer around the inferred position of the evader. This pattern appears in all examples and reflects a strategy to maximize the chance of intercepting the evader while minimizing unnecessary communication. The looping behavior continues until the pursuer performs another communication or closes the distance sufficiently to trigger capture.

The distance dynamics between the agents, shown in the inset plots, further illustrate the non-linear nature of the pursuit process. In Examples 6a and 6c, the distance fluctuates due to temporary misjudgments in the evader's inferred position. In Example 6b, the distance initially increases as the evader escapes, but is followed by a sharp decrease once the pursuer commits to a sequence of closely spaced communications. This suggests that the pursuer inferred a containment opportunity, ultimately cornering the evader in the lower-right region of the map. Interestingly, Example 6b also demonstrates that multiple consecutive communication actions—although seemingly redundant—might be effective when coordinated with spatial awareness.

B IMPLEMENTATION DETAILS

To promote generalization and avoid overfitting to specific scenarios, we introduced randomization in the initial conditions of each training episode. In particular, the initial positions of the pursuer and the evader were sampled uniformly over the map. Moreover, two key environment parameters were randomized: the shooting radius $r_e \sim \mathcal{U}(1, 4)$, and the speed ratio between the evader and the pursuer $v_e/v_p \sim \mathcal{U}(0.1, 4)$. We assume that both agents share the same maximum lateral acceleration, ensuring symmetry in maneuverability.

To monitor performance during training, we conducted validation every 250 episodes by running the current policy over 50 independent simulations. These intermediate evaluations provide insight into learning dynamics and are used to track policy convergence across experimental conditions.

All experiments were conducted on a high-performance computing workstation equipped with an Intel 10th Gen i9-10980XE processor, 256 GB of RAM, and an NVIDIA RTX A6000 GPU with 48 GB of dedicated memory. The simulation environment and learning algorithms were implemented in `Python 3.10` using `PyTorch 2.3`. In Table 2, we report the values of each hyperparameter of SHADOW and the environment. Notably, the query module π_{query} runs in parallel with the navigation module π_{nav} on the same device, introducing no meaningful computational overhead. As a result, SHADOW's computational complexity is essentially inherited from its underlying RL components: TD3 for π_{nav} and the opponent model π_{opp}, and PPO for π_{query}. The training time for the best configuration of SHADOW is approximately 55 hours on our hardware.

C BASELINES

We examine whether SHADOW can outperform three heuristic and four learning-based strategies in terms of effectiveness and efficiency. Specifically, we compare SHADOW against seven baselines:

i. *No Communication*: the pursuer never communicates.
ii. *Random Communication*: the pursuer uses a distance-dependent probability function to decide when to query the evader's state. It computes the current distance between the pursuer and the last observed evader $d(P, E)$, then estimates the probability of getting shot p_{shot}. The probability of communication is then set to $p_{\text{comm}} = 1 - p_{\text{shot}}$.
iii. *Periodic Communication*: the pursuer communicates periodically, each k timesteps. Notably, this strategy has proven to be theoretically optimal in similar partially observable control settings [19], where shooting was not taken into account, and thus serve as a meaningful non-learning benchmark.
iv. *MultiHead PPO* [12]: the pursuer leverages a multi-headed actor with PPO to jointly learn the query and navigation policies.
v. *P-DQN* [51]: the pursuer leverages a Parametrized Deep Q-Network to jointly learn the query and navigation policies. Notably, this baseline is specifically designed for hybrid action spaces, as in our environment the pursuer has to take a discrete action (query q_p) and a continuous action (lateral acceleration u_p).
vi. *HyAR* [23]: the pursuer uses discrete action (query action q_p) embeddings to capture high-level choices and a conditional

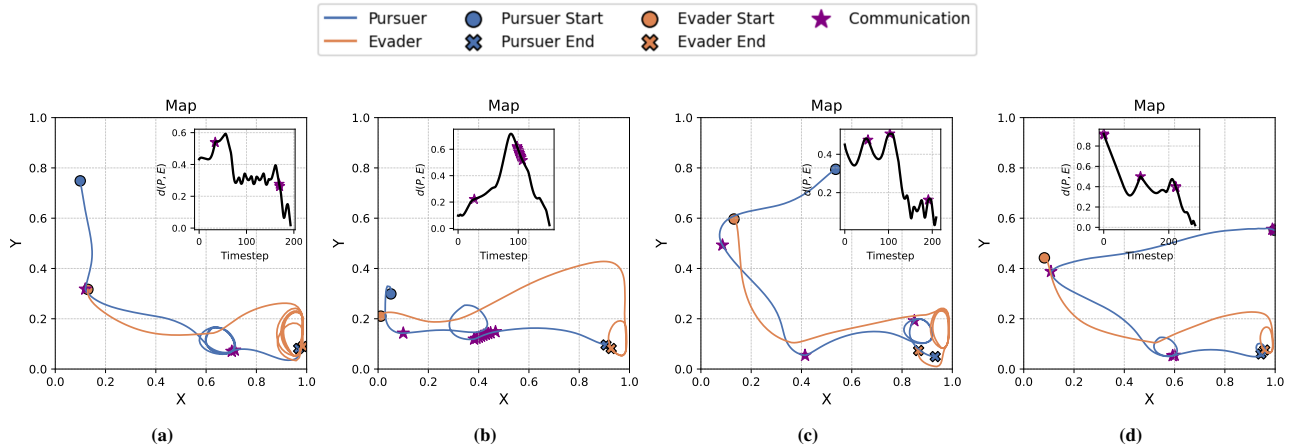


Figure 6: Examples of pursuer-evader interactions in our PEEC game: Each subplot illustrates a representative game, showing the trajectories of the pursuer and evader over the 2D map. The inset plot in each subplot reports the pursuer–evader distance as a function of the timestep.

Category	Parameter	Value
Environment	Maximum Mission Time (s), T	1000
	Map Size (km × km), $M = [0, W] \times [0, H]$	$W = H = 1$
	Reference Speed (knot), v_{ref}	15
	Speed Ratio, v_e/v_p	$v_e/v_p \sim \mathcal{U}(0.1, 4)$
	Agents' Acceleration (rad/s), U_p, u_e	0.9π
	Catching Radius (m), r_c	25
	shooting radius (m), r_s	$r_s \sim \mathcal{U}(25, 100)$
Reward	Time Penalty, $\alpha_p^{\text{T}}, \alpha_e^{\text{T}}$	0.5, -0.5
	Communication Penalty, α_p^{Q}	0
	Hit Boundary Penalty, $\alpha_p^{\text{B}} = \alpha_e^{\text{B}}$	10
	Lateral Acceleration Penalty, $\alpha_p^{\text{A}} = \alpha_e^{\text{A}}$	0.5
	Terminal Reward - Shooting Penalty	100
	Terminal Reward - Catch Bonus	1000
Sequential TD3	Soft Update Rate	0.005
	Policy Noise	0.2
	Noise Clip	0.5
	Policy Delay	5
	Exploration Noise	0.1
Sequential PPO	Policy Clip	0.2
	λ for Generalized Advantage Estimation	0.95
Shared	No. Training Steps	10^7
	Evaluation Frequency	250
	No. Evaluation Episodes	50
	Replay buffer size	10^6
	Batch size	32
	Hidden dimension	256
	Learning rate, η	$3e^{-4}$
	Discount factor, γ	0.99

Table 2: Environment and model hyperparameters.

variational autoencoder to generate continuous actions (lateral acceleration u_p) conditioned on these choices.

vii. *LIAM* [35]: the pursuer learns a model of its opponent through an encoder-decoder architecture that reconstructs the evader's position from its own partial observations. For fair comparison, the pursuer uses the same modular architecture as SHADOW: TD3 for navigation policy, PPO for query policy, and LIAM's encoder-decoder mechanism replacing SHADOW's opponent modeling module.

We adapted publicly available and widely adopted implementations of Multihead PPO [21], Parameterized DQN (P-DQN) [6] and LIAM [14]. We developed our own implementation of HyAR since we did not find any official code-base for this approach. These models were adapted to operate under the same observation and action spaces as SHADOW, and trained under equivalent conditions to ensure fair comparison.

D METRICS

End-State Outcomes $P_{\text{timeout}} + P_{\text{win}} + P_{\text{shot}} = 1$ denote the fraction of the N evaluation episodes in which the pursuer succeeds, is eliminated, or runs out of time (i.e., the evader escapes), respectively.

Communication Strategy The average fraction of timesteps in which the pursuer communicates is defined as:

$$C_{\text{ratio}} = \mathbb{E}_{i=1}^{N}\left[\mathbb{E}_{t=1}^{T_f^{(i)}}\left[q_t^{(i)}\right]\right],$$

where $\mathbb{E}_i^N[X_i]$ denotes an average $N^{-1}(X_1 + \ldots + X_N)$.

Let $\{t_j^{(i)}\}_{j=1}^{N_{\text{Q}}^{(i)}}$ be the ordered timesteps when communication occurs in episode i. The average number of timesteps between consecutive communications is:

$$C_{\text{gap}} = \mathbb{E}_{i=1}^{N}\left[\mathbb{E}_{j=2}^{N_{\text{Q}}^{(i)}}\left[(t_j^{(i)} - t_{j-1}^{(i)})\right]\right]$$

The average distance between agents at the last communication is $D_{\text{comm}} = \frac{1}{N}\sum_{i=1}^{N} r^{(i)}(t_{M_i}^{(i)})$.

Behavioral Efficiency We assess the physical effort and timing of each episode. The average episode duration, in timesteps, is $T_{\text{len}} = \mathbb{E}_{i=1}^{N}\left[T_f^{(i)}\right]$.

The average steering cost per timestep for the pursuer and evader are defined as $\bar{S}_P = \mathbb{E}_i\left[\mathbb{E}_t\left[u_p^{(i)}(t)\right]\right]$ and evader $\bar{S}_E = \mathbb{E}_i\left[\mathbb{E}_t\left[u_e^{(i)}(t)\right]\right]$, respectively.

Information Acquisition Cost We assess the lower bound on the cost of information acquisition as the average pursuer payoff over the average number of queries:

$$\underline{\text{CIAC}} = \frac{\mathbb{E}_i\left[P_p^{(i)}\right]}{\mathbb{E}_i\left[N_{\text{Q}}^{(i)}\right]}$$

E ABLATION STUDY

E.1 Opponent Modeling

Figures 7a shows outcome-related metrics under four configurations of opponent modeling. We observe that equipping any agent with an opponent modeling module consistently provides a strategic advantage to that agent. When the pursuer is the sole user of opponent modeling (*P does, E does not*), it achieves the highest win rate (P_{win} = 0.658), while the evader struggles to evade capture, resulting in the lowest timeout (P_{timeout} = 0.02) and shot (P_{shot} = 0.322) rates. In contrast, when only the evader employs opponent modeling (*P does not, E does*), its ability to avoid capture increases significantly, reflected in a high timeout rate (P_{timeout} = 0.446) and a markedly reduced pursuer win rate (P_{win} = 0.17).

Interestingly, in the symmetric setting, where neither agents utilize opponent modeling (*No Opponent Modeling*), the game is balanced, with near-equal win and shot rates (P_{win} = 0.472 and P_{shot} = 0.528, respectively), indicating neither agent holds a significant predictive advantage. Conversely, when both agents utilize opponent modeling (*Both do*), the game turns in favour of the pursuer (P_{win} = 0.620).

To explain this difference, we investigate how opponent modeling influences the pursuer's communication strategy. We find in Figure 7b that the opponent modeling module reduces reliance on frequent updates, enabling the pursuer to act more autonomously and efficiently. For example, in configuration *P does, E does not*, we observe the lowest communication ratio (C_{ratio} = 0.15) and the largest average gap between communications (C_{gap} = 38.73). By contrast, the baseline configuration *No Opponent Modeling* yields a communication ratio C_{ratio} = 0.24 (37.5% increase w.r.t. 0.15 obtained for the *P does, E does not* configuration) and a average gap between communications C_{gap} = 10.13 (73.8% decrease w.r.t. 48.73 for the *P does, E does not* configuration).

Figure 8 extends our ablation study by examining the impact of the opponent modeling module on two alternative baseline architectures: MultiHead PPO and P-DQN. While these methods differ in their absolute performance, the overarching trends observed with SHADOW remain consistent. For example, for MultiHead PPO in asymmetric settings, the agent equipped with opponent modeling consistently gains a strategic advantage over its counterpart.

In addition, when both agents employ opponent modeling, the pursuer improves its win rate, similar to the behavior observed in SHADOW. This trend is particularly evident in the case of P-DQN, where the pursuer's win rate increases from 26.8% (without opponent modeling) to 39.6% (with opponent modeling enabled for both agents). These findings reinforce the conclusion that, irrespective of the evader's configuration, incorporating opponent modeling is beneficial for the pursuer.

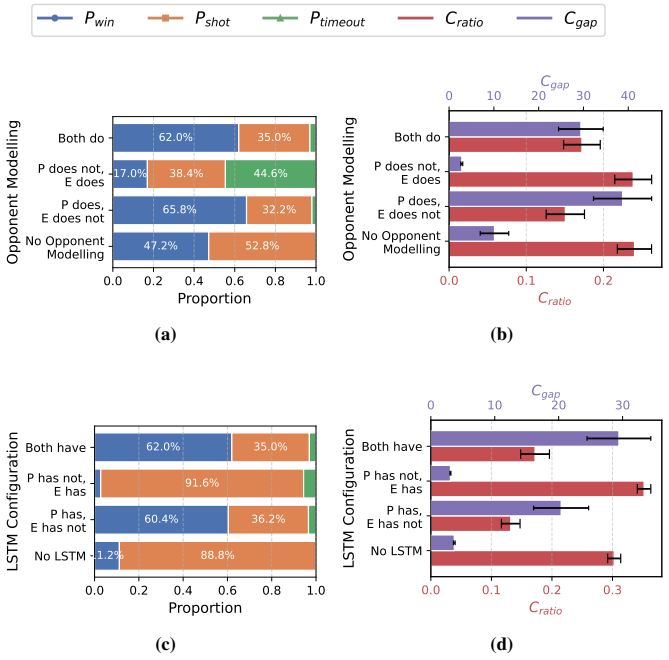


Figure 7: Ablation Study: Game outcome and communication-related metrics for different opponent modeling configurations (a-b) and LSTM configurations (c-d).

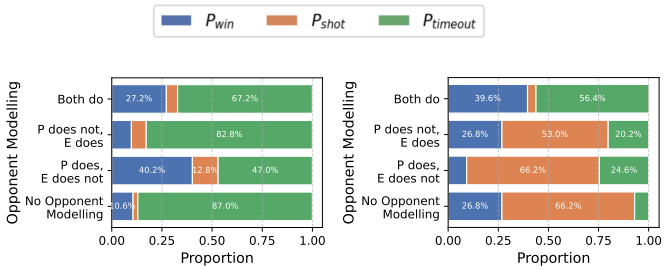


Figure 8: Ablation Study: Outcome-related metrics (P_{win}, P_{shot}, P_{timeout}) for the four configurations of opponent modeling for MultiHead PPO (a) and P-DQN (b) baselines.

E.2 LSTM-based Configurations

Next, we investigate the role of the LSTM-based sequence encoder used to learn the query decision policy and navigation policy. Figure 7c shows outcome-related metrics under the four LSTM configurations. The results indicate a substantial impact of temporal abstraction on the pursuer's performance. For example, when neither agent uses an LSTM, the pursuer rarely succeeds (P_{win} = 0.112), being always shot down before interception (P_{shot} = 0.888). In contrast, when the pursuer is equipped with an LSTM and the evader is not, the pursuer's win rate increases dramatically to 0.604—mirroring the effect of the opponent modeling module.

The LSTM also benefits the evader: in the asymmetric setup where only the evader uses an LSTM, the pursuer's evasion rate,

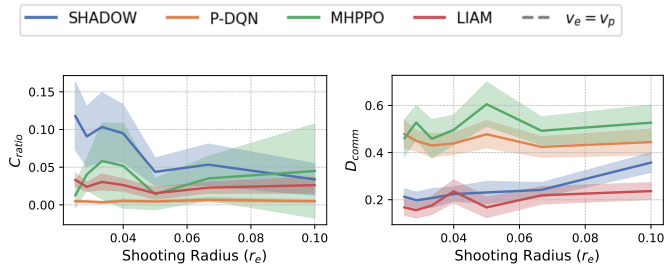


Figure 9: Sensitivity analysis: Effect of the shooting radius r_e on the communication ratio C_{ratio} and the average distance at last communication D_{comm}.

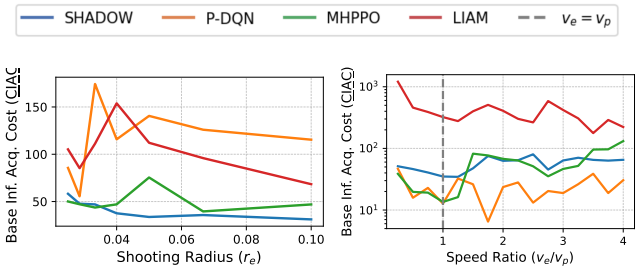


Figure 10: Sensitivity analysis: Effect of the shooting radius r_e (a) and the speed ratio v_e/v_p (b) on CIAC.

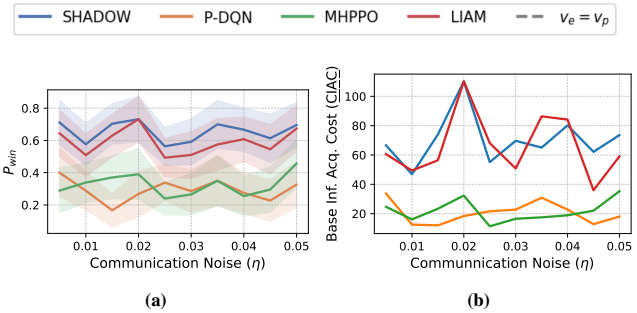


Figure 11: Sensitivity analysis: Effect of the communication noise η on the pursuer win rate P_{win} and the Base Information Acquisition Cost CIAC.

i.e., P_{timeout}, increases from 0 to 0.056. Finally, the full symmetric configuration (both agents using LSTM) yields the highest performance overall ($P_{\text{win}} = 0.620$), suggesting that access to temporal patterns benefits both agents, but still results in a net advantage for the pursuer.

Beyond win/loss outcomes, Figure 7d shows that the LSTM encoder has the same effect on the communication policy learnt by the pursuer, i.e., it allows the agent to communicate less frequently (lower C_{ratio} and larger C_{gap}). This is likely because this sequential model acts like a memory to remember recent communications[8].

F SENSITIVITY ANALYSIS

F.1 Shooting radius

Figure 9 shows the effect of the shooting radius r_e on the communication ratio C_{ratio} and the average distance at last communication D_{comm} for SHADOW, P-DQN, MHPPO and LIAM.

As discussed in the main paper, SHADOW exhibits clear and significant trends: as r_e increases, communication involves more risk, and so the pursuer communicates less frequently (lower C_{ratio}) and with larger intervals between communications (larger D_{comm}). These behavioral adaptations reflect a risk-aware communication policy that dynamically adjusts to the cost of information disclosure. In contrast, MHPPO, P-DQN and LIAM do not show meaningful variation: their C_{ratio} and D_{comm} remain stable as r_e varies, reinforcing the conclusion that these methods are insensitive to changes in communication risk.

Next, we investigate how the shooting radius r_e influences the information acquisition cost directly (CIAC). Figure 10a shows CIAC as a function of r_e for SHADOW, P-DQN, MHPPO and LIAM. Across all values of r_e, pursuers controlled by P-DQN, MHPPO and LIAM exhibit higher CIAC, which depends on their lower communication frequency (see Table 1). Notably, only SHADOW displays a decreasing trend: as r_e increases, CIAC decreases. In other words, the greater the threat of being shot, the more selectively the pursuer communicates. This behavioral adaptation is supported by a regression analysis using r_e as the independent variable ($\beta = -330.9$, FDR-corrected $p = 0.024$, $R^2 = 0.669$). In contrast, P-DQN, MHPPO and LIAM show no significant sensitivity to risk level, with FDR-corrected $p > 0.42$ in both cases.

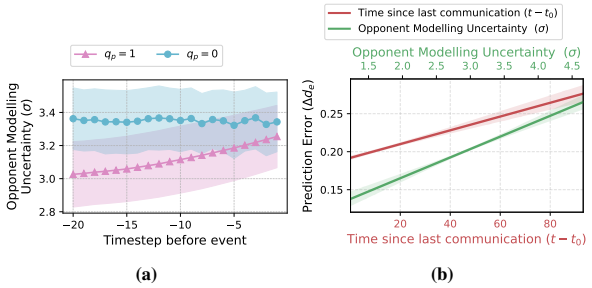
F.2 Speed Ratio

We examine the impact of the evader-to-pursuer speed ratio v_e/v_p on *CoI*. As shown in Figure 10b, SHADOW and MHPPO both exhibit a positive association, indicating that as the evader becomes relatively faster, CIAC increases ($\beta = 7.48$, FDR-corrected $p = 0.018$, $R^2 = 0.372$ for SHADOW; $\beta = 20.44$, FDR-corrected $p = 0.005$, $R^2 = 0.518$ for MHPPO). This suggests that greater speed disparities in favor of the evader compel the pursuer to rely more heavily on communication. A likely explanation is that faster evaders require more adaptive and coordinated responses, increasing the value of communication. In contrast, P-DQN and LIAM shows no statistically significant trend ($p > 0.60$), suggesting limited behavioral responsiveness to changing evader dynamics.

F.3 Communication Noise

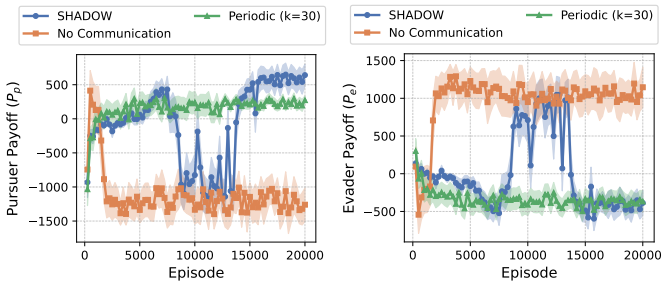
Figure 11 shows the pursuer win rate P_{win} and the Base Information Acquisition Cost CIAC as a function of the communication noise $\boldsymbol{\eta}_q \equiv \eta$ for SHADOW, P-DQN, MHPPO and LIAM. Neither metric exhibits a significant trend as noise increases (FDR-corrected $p > 0.7$).

[8]While we include the elapsed time since last communication in the state of the game, the LSTM-based memory may also represent previous communications and their effect on the agent's dynamics.



(a) (b)

Figure 12: Uncertainty Dynamics: (a) Temporal evolution of predicted uncertainty σ in the 20 timesteps preceding two types of events: communication actions ($q_p = 1$) and non-communication actions ($q_p = 0$). (b) Prediction error Δd_e of the opponent modeling module as a function of the predicted uncertainty σ and the time since the last communication ($t - t_0$).



(a) (b)

Figure 13: Training Dynamics: Cumulative reward obtained by the pursuer (a) and the evader (b) across training episodes under three strategies: SHADOW, periodic communication, and no communication (baseline).

G UNCERTAINTY DYNAMICS

We examine the dynamics of the uncertainty σ predicted by the opponent modeling module over time. Our analysis focuses on two conditions: (i) the 20 timesteps leading up to a communication event ($q_p = 1$), and (ii) the 20 timesteps preceding a generic non-communication event ($q_p = 0$), which serves as a control group. Figure 12a shows the temporal trends of σ under both conditions. A linear regression identifies a significant upward trend in uncertainty prior to communication events ($\beta = 0.012$, FDR-corrected $p = 3.81 \times 10^{-16}$), while no significant trend is observed in the control condition ($p = 0.258$). These findings suggest that the pursuer tends to query the evader's state when uncertainty about its position accumulates, highlighting the role of the opponent modeling module in guiding information-seeking behavior.

We further analyze the relationship between the predicted uncertainty σ and the time since the last communication ($t - t_0$). Both variables are intuitively related to the evader's position: Indeed, longer intervals since the last observation are expected to increase uncertainty, and similarly, high σ explicitly signals unreliable predictions from the opponent modeling module. Our goal is to understand whether these two variables provide overlapping or complementary information. To investigate this, we measure the prediction error $\Delta d_e = \|\mathbf{s}_e - \mathbf{s}'_e\|_2$, defined as the distance between the evader's true position $\mathbf{s}_e$ and the predicted position $\mathbf{s}'_e$. We then fit a regression model to predict this error using both σ and $t - t_0$. As shown in Figure 12b, both variables are significantly and positively associated with prediction error ($\beta = 0.0219$, FDR-corrected $p = 2.91 \times 10^{-48}$ for σ; $\beta = 0.0466$, FDR-corrected $p = 5.34 \times 10^{-36}$ for $t - t_0$). These results suggest that σ and the elapsed time since last communication capture different aspects of uncertainty, and imply that the pursuer may benefit from using both signals to guide its behavior.

H TRAINING DYNAMICS

Figure 13 shows the cumulative rewards achieved by the pursuer and the evader throughout the training process. We compare three settings: (i) our proposed SHADOW method, (ii) a baseline with periodic communication (fixed interval $k = 30$), and (iii) a no-communication baseline.

In the No Communication scenario, the pursuer is unable to achieve any meaningful success. Its reward rapidly converges to values below 1,000 by episode 2,500, indicating that it consistently fails to intercept the evader. Meanwhile, the evader accumulates high rewards, reflecting its repeated success in escaping. However, this case is degenerate: due to the complete absence of communication, the pursuer is effectively blind, and the evader's high reward is not the result of a refined evasive strategy but rather a static failure of the learning process to progress.

In the periodic communication setting, training converges quickly (by episode 5,000). The pursuer achieves a moderate reward plateau of approximately 200, indicating that fixed communication intervals provide enough situational awareness to enable occasional successes, albeit without strategic adaptability.

In contrast, SHADOW exhibits a slower but more structured learning trajectory, consistent with the multi-stage training dynamics described in the main paper: Initially, the pursuer prioritizes minimizing the risk of being shot; only in later episodes does it fine-tune its communication timing to improve interception performance. This results in a longer convergence period (approximately 15,000 episodes), but ultimately leads to a higher average reward (approximately 500) than both baselines.

I COMPREHENSIVE RELATED WORK

Pursuit-Evasion Differential Games (PEGs) [48] form a canonical framework for adversarial motion planning and have been studied extensively in the robotics literature, with applications that span from autonomous missile interception [49] and air-combat maneuvering [22] to self-driving cars [55] and multi-robot security patrol [24, 34, 42]. A growing body of work relaxes the idealized assumption of perfect information. We group prior works on PEGs with imperfect information into three categories that we scan below.

Exogenous visibility limits. In this category, access to state information is constrained by external environmental factors. [7] proved existence of Nash equilibria despite obstacle-induced blind spots

and line-of-sight constraints. Discrete-time counterparts show that scheduled noisy measurements qualitatively change optimal strategies [8], while incomplete-information variants highlight the need for randomized evader policies when the pursuer's observations are noisy [17]. In all of these works, the access to the state is limited by the environment, which may return a noisy observation, the local observation up to a given radius, or the state at predetermined set of discrete timesteps. In particular, players cannot decide when to acquire additional information.

Internal sensing cost. Another line of work endows agents with sense or communicate actions that incur purely internal costs. In this setting, players can strategically decide to acquire information, but such queries come at a certain cost that can model e.g., battery capacity [27], limited bandwidth [3], or link-establishment cost [32]. The cost can be either integrated as part of the player's payoff, imposing a soft constraint, or as a hard constraint, such as a budget that limits the number of state queries. To this end, [29] considered a switched-link linear-quadratic (LQ) differential game, were a fixed cost per measurement is introduced. Some extensions cover discrete-event formulations [30, 31], finite-budget intermittent sensing [28], and asymmetric settings where only the disadvantaged player can request costly observations [1, 2]. A related study features a static remote sensor whose limited transmissions—or deliberate silence—shape the game's information flow [27]. In general, these papers yield limited sensing policies and quantify the performance loss under scarce measurements.

Implicit exposure cost. As mentioned earlier, only [19] explicitly considers information acquisition at the expense of information disclosure. While pioneering, their work assumed a relatively structured PEG that allows closed-form solutions and theoretical analysis on the one hand, but is limited in its applications on the other hand.

In this work, we alleviate assumptions taken in [19], and resort to Deep Reinforcement Learning (DRL) techniques in order to achieve a robust methodology for PEECs that applies to a broader and more realistic class of PEECs. In particular, we do not restrict ourselves to LQG games, and consider non-holonomic, non-linear dynamics and non-quadratic payoffs. Moreover, we do not have a global system payoff to be minimized by the evader and maximized by the pursuer. Instead, each player has its own payoff that it aims to minimize, and in turn the game may not be zero sum. Even in this broader setting, we empirically observe that without explicit communication cost, the pursuer should communicate frequently, even when it is slower than the evader and less maneuverable to some extent. We therefore strengthen the evader and allow it to eliminate a pursuer that discloses its location by communicating, with some probability that is proportional to the distance between them.

Reinforcement-Learning Solutions to PEGs (Without Exposure Cost). Recently, an increasing number of works utilize DRL tools to solve non-LQ PEGs with great success, yet they do not consider binding sensing to reveal information. Examples include multi-agent RL for underwater target-hunting with communication delay [46, 47]; deep-deterministic-policy-gradient guidance against maneuvering evaders [50]; microswimmer pursuit–evasion at low Reynolds number [9]; decentralized multi-agent RL that scales to physical quadrotor pursuits [11]; and more [18, 45, 52]. These studies demonstrate that model-free methods can address nonlinear dynamics, partial observability, and multi-agent coordination, yet none integrates an exposure/detection effect.

J OMITTED PROOFS

Proposition 2. *With a zero-sum assumption (i.e., $P_e \equiv -P_p$) and $r_e = 0$, $\alpha_c^{\mathrm{Q}} \geq 0$ is a maximum.*

Proof. We first prove the existence of a maximum α_c^{Q}.

Since we assume a two-player zero-sum game, by the Min-Max Theorem [44], the payoff in all Nash-equilibria is the same, and equals:

$$V(\alpha_p^{\mathrm{Q}}) := \max_{\Omega_{\mathsf{NE}}[\alpha_p^{\mathrm{Q}}]} \mathbb{E}[P_p \mid \alpha_p^{\mathrm{Q}}] = \max_{(u_p,q_p)} \min_{u_e} P_p(\langle (u_p, q_p); u_e \rangle \mid \alpha_p^{\mathrm{Q}})$$

.

Since the space of pursuer and defender strategies is compact, it follows from the Maximum (Minimum) Theorem [4] that $V(\alpha_p^{\mathrm{Q}})$ is continuous. Moreover, V is monotonic decreasing with α_p^{Q}. Indeed, let $\alpha_1 > \alpha_2$, and let $\langle (u_p^1, q_p^1); u_e^1 \rangle \in \Omega_{\mathsf{NE}}[\alpha_1]$. Then we have:

$$\begin{aligned} V(\alpha_2) &= \max_{(u_p^2,q_p^2)} \min_{u_e^2} P_p(\ldots \mid \alpha_2) \\ &\geq \min_{u_e^2} P_p(\langle (u_p^1, q_p^1); u_e^2 \rangle \mid \alpha_2) \\ &= P_p(\langle (u_p^1, q_p^1); u_e^1 \rangle \mid \alpha_2) \\ &\geq P_p(\langle (u_p^1, q_p^1); u_e^1 \rangle \mid \alpha_1) = V(\alpha_1), \end{aligned}$$

Where in the first inequality we use the same pursuer strategy for the lower communication penalty, the equality is then since the communication penalty is independent of the evader strategy and since $\langle (u_p^1, q_p^1); u_e^1 \rangle$ is in equilibrium, and the last inequality is due to P_p being linearly decreasing with the communication penalty, when the strategy profile is fixed.

Finally, $\lim_{\alpha_p^{\mathrm{Q}} \to -\infty} V(\alpha_p^{\mathrm{Q}}) = \infty$. This is because when α_p^{Q} is negative, the pursuer gets a reward of $|\alpha_p^{\mathrm{Q}}| \to \infty$ for a single query.

Therefore, the set $\Omega_V := \{\alpha_p^{\mathrm{Q}} \mid V(\alpha_p^{\mathrm{Q}}) \geq 0\} = V^{-1}([0, \infty))$ has a maximum, α_c^{Q}.

Next, let $\alpha_p^{\mathrm{Q}} < 0$ and assume $r_e = 0$. We will construct a pursuit strategy with a positive payoff. Let $d_0 = \mathbb{E}[D_0] > 0$ be the expected distance between the pursuer and the evader at $t = 0$. By Markov inequality, with probability at least 0.5, the initial distance is at least $d_0/2$. Let $t_0 > 0$ be the minimum time it takes the pursuer and evader to meet given initial distance is $\geq d_0/2$. For instance, in our game, $t_0 = \frac{d_0/2 - r_c}{(v_p + v_e)}$ as the players move at a constant speed. Finally, denote by $-P_{\max}$ the maximal negative payoff the pursuer can get. Consider the query strategy $q_p(\alpha_p^{\mathrm{Q}})$ that queries the state every

$$dt_0 := \frac{t_0}{3\lceil P_{\max}/\alpha_p^{\mathrm{Q}} \rceil}$$

times starting with $t = 0$, and let u_p be arbitrary. Then the pursuer expected payoff equals:

$$\begin{aligned} \mathbb{E}[P_p] = &\mathbb{E}[P_p \mid D_0 > d_0/2] \Pr[D_0 > d_0/2] + \\ &\mathbb{E}[P_p \mid D_0 \leq d_0/2] \Pr[D_0 \leq d_0/2] \geq \\ &(3-1) P_{\max} \cdot 0.5 - P_{\max} = 0, \end{aligned}$$

where we bound the second term by $-P_{\max}$, and for the first term, given $D_0 > d_0/2$ (w.p. at least 0.5), the pursuer is ensured to perform at least $3P_{\max}/\alpha_p^{\mathrm{Q}}$ queries, and therefore receive a communication reward of $3P_{\max}$ overall.

Thus, for any negative communication penalty, there exists a maximin strategy for the pursuer yielding a non-negative expected payoff. Therefore, the supremum over $V^{-1}([0,\infty))$ must be non-negative, and so $\alpha_c^{\mathrm{Q}} \geq 0$. □

Proposition 4. *Assuming zero-sum (i.e., $P_e \equiv -P_p$), $\underline{\alpha}_c^{\mathrm{Q}} \leq \alpha_c^{\mathrm{Q}}$.*

Proof. Recall for the case of a zero-sum game, we defined:

$$V(\alpha_p^{\mathrm{Q}}) := \max_{(u_p,q_p)} \min_{u_e} P_p(\langle (u_p,q_p); u_e\rangle \mid \alpha_p^{\mathrm{Q}}),$$

and α_c^{Q} is defined as the maximum communication penalty for which V is non-negative. It is therefore sufficient to prove that $V(\underline{\alpha}_c^{\mathrm{Q}}) \geq 0$. Indeed:

$$\begin{aligned} V(\underline{\alpha}_c^{\mathrm{Q}}) &\geq \min_{u_e'} \mathbb{E}[P_p(\langle (u_p,q_p); u_e'\rangle \mid \underline{\alpha}_c^{\mathrm{Q}})] \\ &= \mathbb{E}[P_p(\langle (u_p,q_p); u_e'\rangle \mid \underline{\alpha}_p^{\mathrm{Q}})] \\ &= \mathbb{E}[P_p(\langle (u_p,q_p); u_e'\rangle \mid 0)] - \underline{\alpha}_p^{\mathrm{Q}} \mathbb{E}[N_p^{\mathrm{Q}}] \geq 0, \end{aligned}$$

where the first inequality is due to fixing a pursuer strategy, the following equality is since the evader is not penalized for pursuer communication, and the last inequality is by the definition of $\underline{\mathsf{CIAC}}$. □