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

AAAI Track

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: AAAI Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/XVYB5151

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. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/XVYB5151

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 [4, 16]. 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) [11] 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) [15, 31] has addressed partial observability [3, 21–23] and cost-sensitive communication [1, 2, 8, 17, 18], 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 [11], 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 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 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 [11]) and adaptive RL methods (MultiHead PPO [6], P-DQN [30], HyAR [14], and LIAM [20]), 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 in Online Supplementary Material for illustrative examples of learned strategies.

Contributions of this work

(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 and risk 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.
(4) **Extensive Experimental Evaluation**: We systematically evaluate SHADOW across several pursuer-evader configurations, analyze learning dynamics, communication strategies, and performance under varying threat levels, and environment conditions[2].

[1]RL-based solutions exist for partially observable PEGs with multi-agent coordination [5] and delayed communication [10, 28, 29], 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 Supplementary Materials, 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 [11], where information acquisition comes at the strategic cost of revealing one's own state.[3] Specifically, [11] 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 non-holonomic 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

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

[3]The notion that silence can itself be informative is also explored in [17].

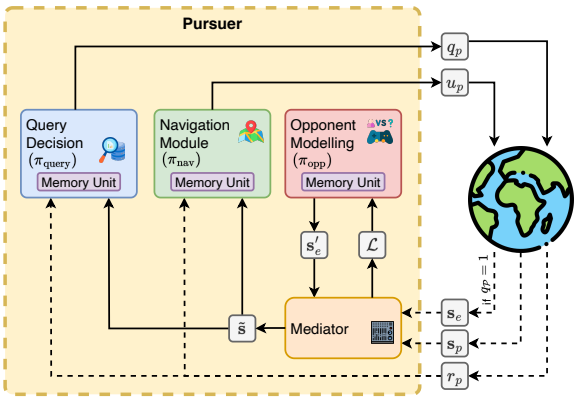


Figure 1: SHADOW Pursuer: The Pursuer operates its navigation control u_p and decides whether to query the opponent's state via a binary action $q_p \in \{0, 1\}$. The environment returns the updated pursuer state s_p and, if $q_p = 1$, the evader's current position s_e. The *Mediator* determines the pursuer's internal state representation $\tilde{s}$, comprising: (i) the current position of the pursuer 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 (s_e), if $q_p = 1$, or inferred by the *Opponent Modeling* module (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 $\tilde{s}$ 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.

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. 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.

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, 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 [29], we learn the navigation policy π_{nav} using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [7], 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) [26], 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 enables policies to learn from both current and past information[5].

This modular design offers two key advantages. First, it provides flexibility in handling the hybrid action structure 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.

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:

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

[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 [24, 27]). 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.

where σ is a scalar representing 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 [13], 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 [9, 25] 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 [12].
- 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 $\alpha_e^{\mathsf{Q}} = 0$ as $\mathcal{E}$ cannot query the state, and $\alpha_e^{\mathsf{T}} = -\alpha_p^{\mathsf{T}} \le 0$ 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 [19, 29]. 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:*

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

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 $\langle (u_p^0, q_p^0); u_e^0 \rangle \in \Omega_{\mathrm{NE}}[\alpha_p^{\mathrm{Q}} = 0]$, 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 Supplementary Material.

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 Appendix B for full hyperparameters 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, $p_{\mathrm{comm}} = 1 - p_{\mathrm{shot}}$. *(iii) Periodic Communication*: the pursuer communicates periodically, each k timesteps. This strategy was proven to be theoretically optimal in the setting of [11][6]. *(iv) MultiHead PPO* [6]: the pursuer leverages a multi-headed actor with PPO to jointly learn the query and navigation policies. *(v) P-DQN* [30]: the pursuer leverages a Parametrized Deep Q-Network to jointly learn the query and navigation policies. *(vi) HyAR* [14]: the pursuer learns the relationship between the discrete action (q_p) and the continuous action (u_p) using a variational autoencoder. *(vii)* LIAM [20]: 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 [32]) 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.

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},

[6] [11] 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.

the speed ratio v_e/v_p, and the communication noise $\boldsymbol{\eta}_q$. This reveals how the pursuer adapts to changing communication risks and differences in physical capabilities.

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%.*

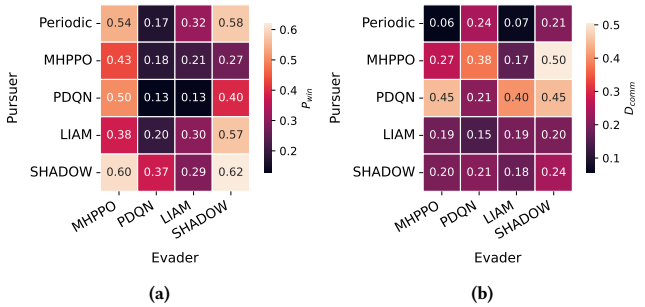
SHADOW vs. Periodic Communication. The Periodic Communication strategy with $k = 40$ achieves the second-highest win rate (P_{win} = 57.6%), trailing SHADOW by 7.1%. This difference is statistically significant, FDR-corrected $p = 0.013$. Furthermore, the Periodic Communication with $k = 40$ incurs a 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).

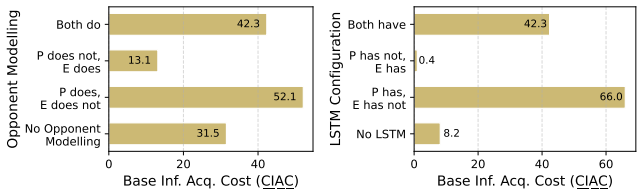
Cross-Strategy Robustness. We evaluate robustness through a pairwise comparison between pursuer and evader strategies. The pursuer is selected from Periodic, MHPPO, PDQN, LIAM, and SHADOW, while the evader is drawn from the same set except for Periodic, which applies only to the pursuer since it is the 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

[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.



(a) (b)

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.



(a) (b)

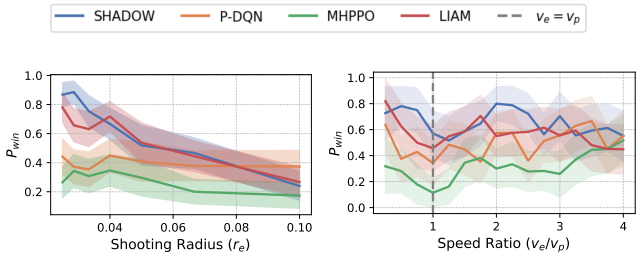
Figure 3: Ablation Study: CIAC under different Opponent Modeling (a) and LSTM (b) configurations.

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



(a) (b)

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}.

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 Figures 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 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 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.

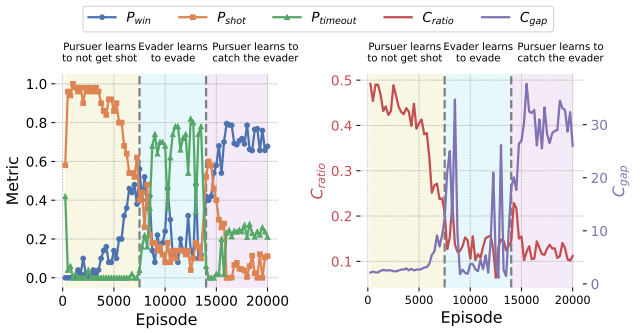


Figure 5: Training dynamics: Evaluation metrics over training episodes. Shaded regions corresponds to distinct learning phases.

6 CONCLUSION

We generalized the PEEC framework proposed in [11] 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 [11] 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.

ACKNOWLEDGMENTS

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. Linear quadratic zero-sum differential games with intermittent and costly sensing. *IEEE Control Systems Letters* (2024).

[2] 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).

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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

[12] 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.

[13] 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.

[14] 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

[15] 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.

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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

[25] 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

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

[27] 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.

[28] 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).

[29] 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).

[30] 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).

[31] Fuhan Yan, Jiuchuan Jiang, Kai Di, Yichuan Jiang, and Zhifeng Hao. 2019. Multi-agent pursuit-evasion problem with the pursuers moving at uncertain speeds. *Journal of Intelligent & Robotic Systems* 95, 1 (2019), 119–135.

[32] 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.