2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Vision-based Pursuit-Evasion Robot Policies

Andrea Bajcsy*, Antonio Loquercio*, Ashish Kumar, Jitendra Malik

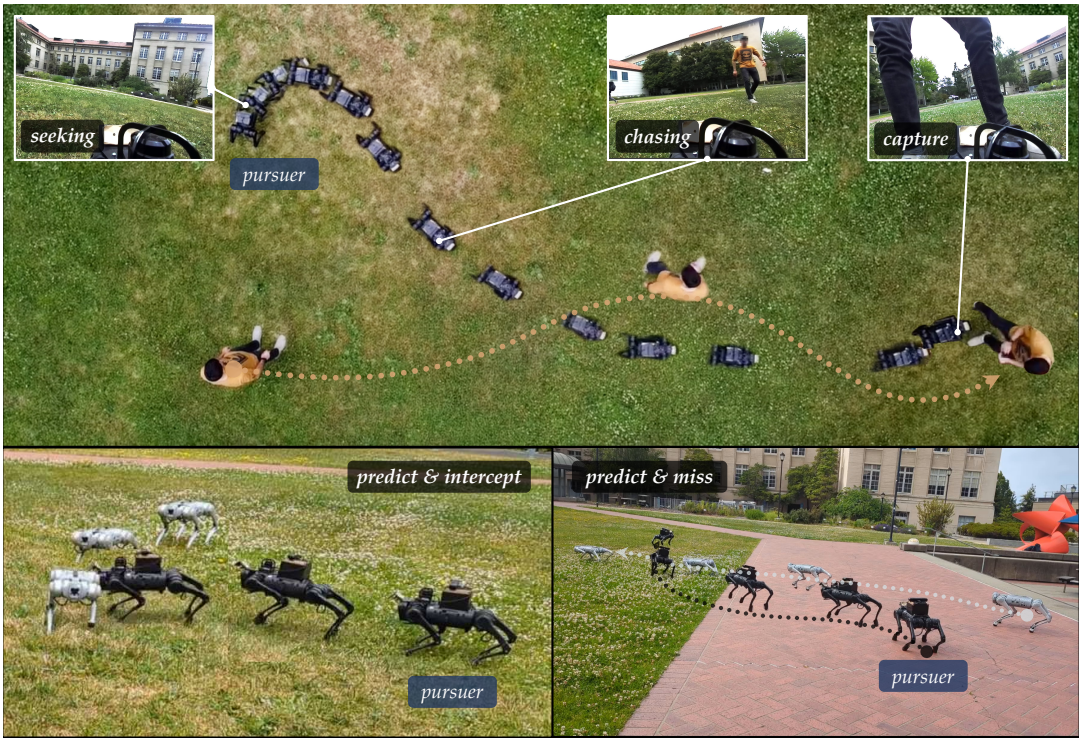

Fig. 1: Our approach deployed in a pursuit-evasion interaction in the wild. Our policy (black robot acting as pursuer) automatically synthesizes behaviors like slowing down and information gathering, accelerating upon detection, and prediction and interception. Video results and code at `https://abajcsy.github.io/vision-based-pursuit/`.

***Abstract*— Learning strategic robot behavior—like that required in pursuit-evasion interactions—under real-world constraints is extremely challenging. It requires exploiting the dynamics of the interaction, and planning through both physical state and latent intent uncertainty. In this paper, we transform this intractable problem into a supervised learning problem, where a fully-observable robot policy generates supervision for a partially-observable one. We find that the quality of the supervision signal for the partially-observable pursuer policy depends on two key factors: the balance of diversity and optimality of the evader's behavior, and the strength of the modeling assumptions in the fully-observable policy. We deploy our policy on a physical quadruped robot with an RGB-D camera on pursuit-evasion interactions in the wild. Despite all the challenges, the sensing constraints bring about creativity: the robot is pushed to gather information when uncertain, predict intent from noisy measurements, and anticipate in order to intercept.**

I. Introduction

Robot learning has accelerated progress for embodied agents acting "in the wild": quadrupedal and wheeled robots navigate through hard-to-model terrains [1]–[5], quadrotors fly at their limits [6], [7], and robotic arms deftly manipulate deformable objects [8]. However, these successes are limited to robots in isolation; in reality, robots deployed at scale will inevitably interact with other agents, like people or robots.

In-the-wild multi-agent interactions raise significant challenges: not only does a robot have to account for perception-induced uncertainty of the physical state (e.g., ego state, positions of others), but it must also account for uncertainty in other agents' future behavior. This problem setting is traditionally modeled by decentralized partially-observable Markov decision processes (Dec-POMDPs) or partially-observable stochastic games (POSGs). While in theory, solutions to these formulations would automatically yield desirable behaviors like information gathering when uncertain, in practice they are notoriously intractable.

Nevertheless, human and animal behavior exhibits these abilities [9]. Pursuit-evasion interactions are a canonical example: the pursuer gathers information about the hidden evader by turning and scanning the environment; upon detection, the pursuer has to continuously strategize about its next move without perfect knowledge of how the evader will react, all from onboard sensors. In this work, we take a step

*equal contribution. All authors with UC Berkeley. This work was supported by the DARPA Machine Common Sense program and by the ONR MURI award N00014-21- 1-2801. Thanks to Noemi Aepli for her help with real-world experiments.

towards building similar capabilities into autonomous robots.

Our key idea is to leverage a fully-observable policy to generate supervision for a partially-observable one. However, the classic paradigm of privileged learning [10] does not apply naively to this setting. Namely, privileged information depends not only on the robot, but also on the other agent's behavior, which is dictated by *more* than just physics; it is dictated by the other agent's intent. Therefore, we design a learning procedure to first build a low-dimensional latent representation of intent from future evader trajectories and then learn to estimate this representation from a history of pursuer actions and observations.

Through extensive empirical analysis, we find that the quality of the supervision signal depends on a delicate balance between the diversity of the agents' behavior and optimality of the interaction. In addition, there are many models for generating the fully-observable supervisor policy (e.g., game-theory [11], multi-agent RL [12]), each with their own potential strengths and weaknesses. We discover that fully-observable policies obtained under strong modeling assumptions (e.g., both agents play under perfect-state Nash equilibrium), are less effective at supervising partially observable ones.

Informed by this analysis, we synthesize a policy that *automatically* takes actions to resolve physical state uncertainty (e.g., looking around to detect the other agent) while also generating predictions about other agents' intent to yield strategic behavior. We deploy this policy on a physical legged robot in a pursuit-evasion game, where it interacts with humans or other legged robots (Fig. 1). Note that the robot only uses onboard sensing, e.g., proprioception and an RGB-D camera, to estimate its state and other agents' physical state and intent.

II. Related Work

Dynamic Games & Multi-Agent RL. Dynamic game theory has a long history of modeling strategic interaction between multiple agents [13]–[16] and has influenced fields like robust control [17] and reinforcement learning [18]–[20]. Both traditional and modern variants of dynamic games have predominantly assumed knowledge of perfect state. While this has been successful in contexts like robustness to physical disturbances [19], [21], [22], it is a limiting assumption for real-world interaction. Partially observable stochastic games provide a mathematical model of strategic interaction under partial observability [23], where all players have only partial information about environment state. However, they are tremendously computationally expensive to solve, and approximations are an active area of research [24]. Limited-FOV pursuit-evasion games have been explored [25]–[29], and while a suite of algorithms exist under varying modeling assumptions, to-date none of these approaches have been demonstrated to work in unstructured interactions (like those between a human and robot) that occur "in the wild" (i.e., unknown a priori environment). To make the optimization tractable, multi-agent reinforcement learning (MARL) algorithms exploit large-scale simulation and neural network representations [19], [30]–[33]. Such approaches have achieved impressive results in simulation interactions like hide-and-seek [34], video games like Starcraft [35], diplomacy [36], and board games like Go [37]. However, to-date, such approaches have not yet scaled to embodied systems acting under real-world sensing constraints More related to our approach is the work in visual tracking, where an agent is trained with reinforcement learning to keep another agent as long as possible in the field of view [38]–[40]. While the objective is similar, we optimize for capture and not maximizing the time the pursuer is in the field of view. At a higher level, our work is inspired by the RoboCup series [41], which is the first real-world multi-agent demonstration in competitive settings.

Latent Intent Modeling. Recent works [42]–[44] learn a latent representation of agent intent via reconstructing a dataset of fully-observed states and rewards. This line of works assumes that the latent intent changes only *between* interaction episodes and not during an interaction. To handle intent changes *during* interaction [45] learns an estimator of a human's latent state to predict their immediate next action. These works overwhelmingly assume that the only hidden state in the interaction is the opponent's intent: the physical state of the robot, the opponent's state, and possibly the opponent's action are assumed to be observable. We address the constraints imposed by on-board robot perception, where the evader's physical state, latent intent, and action are hidden.

Multi-Quadruped Interaction. Progress in low-level control for quadrupedal robots [1], [3], [4] has increased the interest in combining low-level controllers with high-level decision-making [46]. However, multi-agent quadruped interactions have been relatively under-explored. Most relevant is [47] which trains a *centralized* high-level coordination policy with perfect (global) state for two robots pushing a box. [48], [49] present cooperative control of robots via model predictive control and control barrier functions. To the best of our knowledge, our work is the first to demonstrate autonomous interaction between a quadruped and another robotic or human agent truly in the wild.

III. Overview

We seek a robot policy that can strategically interact with another agent in a decentralized fashion (i.e., no explicit communication) and only using proprioception and a single onboard RGB-D camera. Although our technical approach is general, we ground this work in pursuit-evasion games [13], which exhibit core challenges at the heart of real-world multi-agent interaction: partial observability, nonlinear physical dynamics (e.g., quadrupedal dynamics), low-latency decision-making, and a need for strategic planning.

We approach this problem using privileged learning [10]. We found that directly using reinforcement learning to train a policy that reasons *strategically* through *partial observability* was unsuccessful. The key to our approach is to leverage a fully-observable policy to generate supervision for

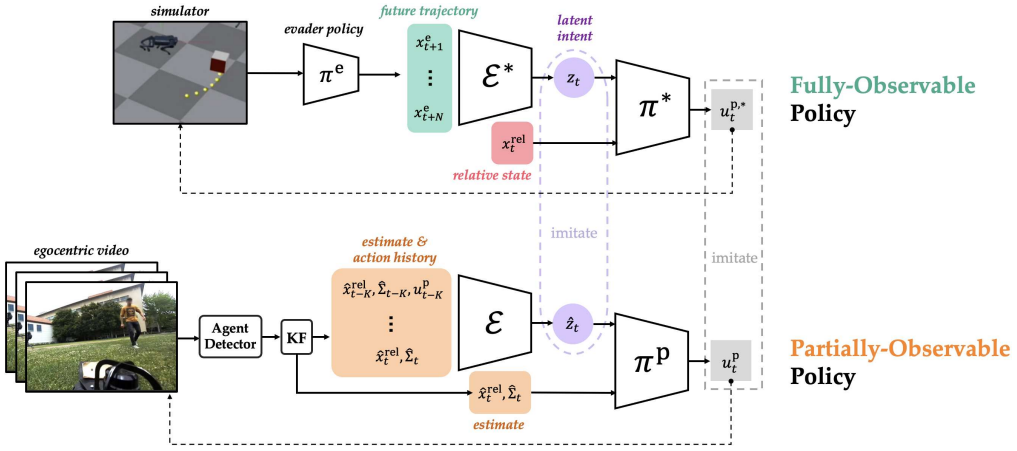

Fig. 3: (top) The fully-observable policy knows the true relative state and gets privileged future evader trajectory from which it learns the evader intent. (bottom) The partially-observable policy must plan through physical and latent intent uncertainty.

the partially-observable one. During privileged training, we leverage a new type of privileged information: the future state trajectory of the evader. We first learn a **fully-observable policy** (π^*) (top, Fig. 3), that gets access the true future N-step state trajectory of the evader and the current true relative state. This enables π^* to quickly learn actions that account for the evader's behavior by using a learned latent intent, z_t, that encodes the future trajectory of the other agent.

We then distill this policy into a **partially-observable policy** (π^{p}) which only uses an egocentric video stream from an onboard RGB-D camera (bottom, Fig. 3). Specifically, π^{p} gets access to a *history* of relative state *estimates and uncertainties* which are generated via a standard Kalman Filter [50]. Even though the Kalman filter is an incredibly coarse approximation of the true system, the uncertainty information captured by the covariance matrices is sufficient for the prediction policy to learn information-gathering behaviors (like turning and looking for the evader), when combined with the teacher policy. In this light, our privileged learning approach can be viewed as an approximation to the optimal policy obtained via solving the underlying, but intractable, decentralized partially-observable Markov decision process (Dec-POMDP).

Our partially-observable policy can be applied zero-shot in the real world using the output of an off-the-shelf object detector [51]. We deploy it in the wild to play a pursuit-evasion game with a human evader and another quadrupedal robot controlled by a human operator.

IV. Approach

Given an evader policy π^{e}, we define the pursuer's planning problem as a finite-horizon, discrete-time optimization problem. We seek a policy for the pursuer $\pi^{\mathrm{p}} : \mathcal{O}^{\mathrm{p}} \to \mathcal{U}^{\mathrm{p}}$ which maps from observations to actions that maximizes:

$$J(\pi^{\mathrm{p}}, \pi^{\mathrm{e}}) = \mathbb{E}_{\tau \sim p(\tau \mid \pi^{\mathrm{p}}, \pi^{\mathrm{e}})} \Big[\sum_{t=0}^{T} \gamma^t r_t \Big]. \tag{1}$$

Here, $\tau = \{(x_t^{\mathrm{p}}, x_t^{\mathrm{e}}, u_t^{\mathrm{p}}, u_t^{\mathrm{e}}, o_t^{\mathrm{p}}, o_t^{\mathrm{e}}, r_t)\}_{t=0}^{T}$ is the joint trajectory of states, actions, observations, and rewards induced by a pair of pursuer and evader policies, drawn from the distribution $p(\tau \mid \pi^{\mathrm{p}}, \pi^{\mathrm{e}})$. The discount factor is denoted by γ. More formally, this optimization defines the solution to a two-agent, finite horizon decentralized partially-observable Markov decision process (Dec-POMDP).

We denote the global physical state of the pursuer as $x^{\mathrm{p}} \in \mathbb{R}^{n_{\mathrm{p}}}$ and the evader to be $x^{\mathrm{e}} \in \mathbb{R}^{n_{\mathrm{e}}}$. Note that the pursuer policy π^{p} does not observe the global state of the agents. The robot's high-level linear and angular velocity commands are denoted by $u^{\mathrm{p}} \in \mathcal{U}^{\mathrm{p}}$ and the pursuer's low-level joint torques are controlled via a pre-computed walking policy. The evader also controls its linear and angular velocity, $u^{\mathrm{e}} \in \mathcal{U}^{\mathrm{e}}$. Both $\mathcal{U}^i, i \in \{\mathrm{p}, \mathrm{e}\}$ are bounded sets, modeling actuation limits. For example, in simulation, the maximum linear speed of the pursuer is 3 m/s, and the evader is 2.5 m/s.

The pursuer is rewarded for minimizing the distance between the two agents at each timestep, and obtains a termination bonus upon capture: *Pursuit:* $r_t = -||x_t^{\mathrm{e}} - x_t^{\mathrm{p}}||_2^2$, *Capture:* $r_T = \alpha \cdot 1\{||x_T^{\mathrm{e}} - x_T^{\mathrm{p}}||_2^2 \leq 0.8\}$ where $\alpha = 100$ is a hyperparameter.

Asymmetries. Our setting has three asymmetries that induce complexity: 1) *information* (agents have limited FOV and partial state), 2) *dynamics* (e.g., robot quadruped interacting with human), and 3) *control bound* asymmetry (e.g., agents with different maximum speeds).

Ego-Centric State. All agents reason about the *relative* physical state in their own body frame. In a slight abuse of notation, we refer to $x^{\mathrm{rel}} := [p_x^{\mathrm{rel}}, p_y^{\mathrm{rel}}, \theta^{\mathrm{rel}}]^\top$ as the true relative planar position and orientation of the exo-centric agent in the ego-centric agent's body frame.

State Estimation. In the wild, the true relative state is not available due to sensing limits. Instead, the pursuer estimates x_t^{rel} from the output of a 3D object detector using the RGB camera [51]. Let $o_t^{\mathrm{p}} \in \mathcal{O}^{\mathrm{p}}$ be the 3D relative position of the evader with respect to the pursuer's camera frame[1]. The pursuer's relative state estimate are the mean and covariance

[1]If the evader is out of the FOV, then $o_t^{\mathrm{p}} = \emptyset$ and no measurement update is performed.

of a Kalman filter: $(\hat{x}_t^{\text{rel}}, \hat{\Sigma}_t) = \text{KF}(o_{0:t}^{\text{p}}, u_{0:t}^{\text{p}})$. While more complex filter designs could be used for even better performance [52], we find that an unoptimized Kalman filter is sufficient for the pursuer to learn information-gathering behaviors.

Evader Policy. The evader policy is key for enabling the pursuer to learn strategic maneuvers. However, where does the evader policy come from? Datasets of quadrupeds interacting with other agents in the wild do not exist, and simulating human-robot or robot-robot interactions that capture the diversity of the real-world is an ongoing challenge for simulation-based robotics. Instead, we take an investigative approach and study three simulated evader policy models: random motion primitives, multi-agent RL, and dynamic game theory. Across all models, we assume the evader has access to the current true relative state in their own body frame[2].

A. Fully-Observable Policy: Teacher

To learn the pursuer teacher policy π^*, we must address the challenge that privileged information depends on the evader's behavior which is dictated by their dynamics and intent.

Future Evader Trajectory & Latent State. The fully-observable policy π^* gets access to both the true pursuer relative state, x_t^{rel}, and the future N states of the evader: $x_{t:t+N}^{\text{e}}$. Since the pursuer reasons in a relative coordinate system, the evader trajectory is converted into the pursuer's body frame starting from state at the start of the prediction horizon. This relative state trajectory, $x_{t:t+N}^{\text{rel}} \in \mathbb{R}^{N\times 3}$, is input into an encoder, $\mathcal{E}^*(x_{t:t+N}^{\text{rel}}) = z_t \in \mathbb{R}^8$ which learns a low-dimensional latent representation. Intuitively, z_t should capture low-dimensional information about the evader's near-term behavior: for example, the evader's goal direction, their policy class (e.g., spline coefficients), or control bounds. At each timestep, z_t is re-inferred.

Design. Although this pursuer policy is clearly not deployable in the real world, it enables us to convert the intractable planning problem in Eq. 1 to a Markov Decision Process (MDP), amenable to off-the-shelf RL methods [53]. We use Proximal Policy Optimization [53] for training. The policy π^* and the privileged encoder $\mathcal{E}^*$ are both three-layer MLPs with $[512, 256, 128]$ hidden units.

B. Partially-Observable Policy: Student

The partially-observable policy, π^{p}, relies on RGB camera observations $o_t^{\text{p}} \in \mathcal{O}^{\text{p}}$. We use a off-the-shelf 3d object detector [51], [54] to convert from the raw RGB image observable o_t^{p} to an detected relative position and heading in the pursuer's camera frame, $y_t \in \mathbb{R}^3$. We use a Kalman Filter to generate an estimated relative state $\hat{x}_t^{\text{rel}}$ and an associated covariance $\hat{\Sigma}_t$. We use a simplified state transition model $\hat{x}_{t+1}^{\text{rel}} = A\hat{x}_t^{\text{rel}} + Bu_t^{\text{p}}$ which ignores the role of the evader (i.e., $u_t^{\text{e}} \equiv 0$) during the the prediction[3] step. The history of relative state estimates and pursuer actions are encoded into the estimated lower-dimensional latent intent $\mathcal{E}(\hat{x}_{0:t}^{\text{rel}}, \hat{\Sigma}_{0:t}, u_{0:t-1}^{\text{p}}) = \hat{z}_t$.

[2]More details in project page.

[3]This removes the need for a velocity estimator which can be hard to design and noisy in reality. While this makes filtering imperfect, we empirically find that the learned policy can compensate for inaccuracies.

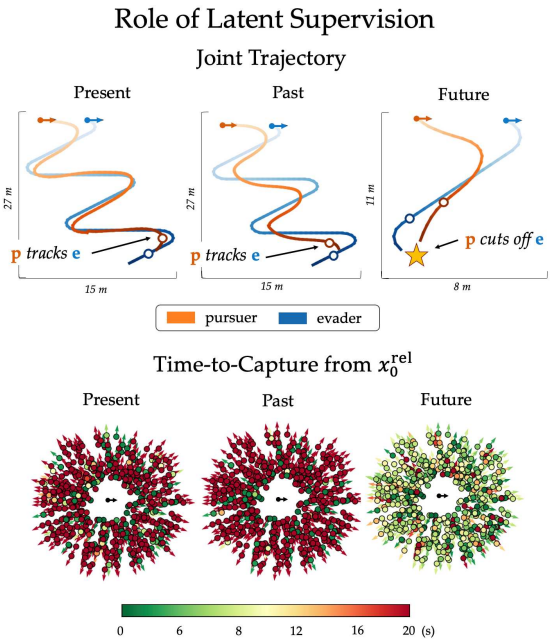

Fig. 4: (top) Joint trajectories reveal that using only the present state or directly a history of the past states leads to sub-optimal performance. Supervision from the future enables prediction and fast capture. (bottom) Time-to-capture as a function x_0^{rel}: future supervision quarters the time.

Design. We use DAGGER [55] and the fully-observable policy π^* to supervise both the latent intent estimate and the action at each timestep. The policy network is a 3-layer MLP with $[512, 256, 128]$ hidden units, and the encoder $\mathcal{E}$ is a 1-layer LSTM with hidden state 256.

V. Simulation Experiments

We first want to understand the design choices that are important to learn a successful pursuer policy: 1) the ability to learn strategic behavior, 2) the evader policy π^{e} that the robot interacts with at *training* time and 3) the evader interaction at *deployment* time. We perform a set of simulation experiments to ablate the design of the pursuer policy (Sec. V-A), study the effect of the evader on distillation (Sec. V-B), and analyze test-time adaptation of the pursuer policy to out-of-distribution opponents (Sec. V-C). We use Isaac Gym [56] for training and evaluation, and report results over 500 random initial conditions.

A. Predictive representations enable strategic behavior

One of the key types of privileged information we leverage is future trajectory state, which leaks information about the future intent of the other agent. In this section, we ask the question *"What is the value of learning predictive representations for action?"* Here, we fix the evader policy to be highly predictable and investigate alternative approaches to inferring the evader's latent intent. The evader always moves in Dubins' paths [57] with a fixed time duration for turning

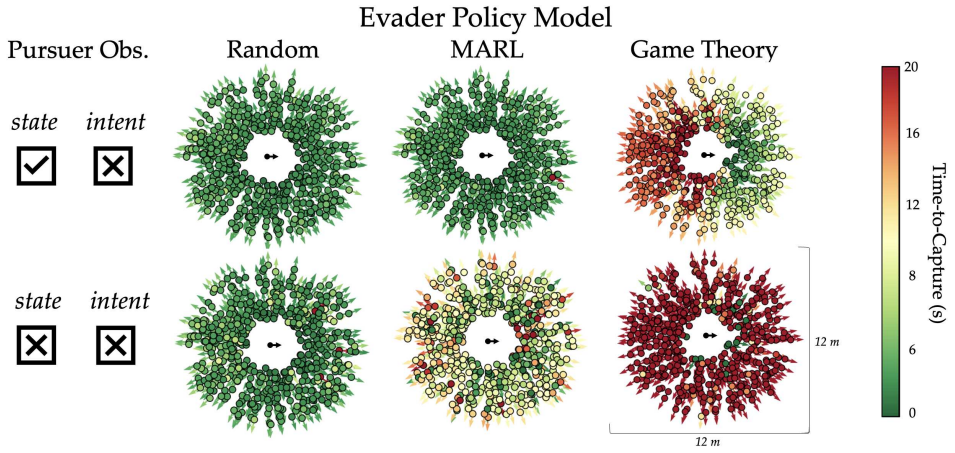

Fig. 5: 500 randomly sampled x_0^{rel} initial positions and relative orientation. Colors indicate the normalized time-to-capture starting from the shown initial condition. (top row) Policy knows perfect physical state but infers latent intent from history of relative states and pursuer actions, trained with three different evader models: heuristic, MARL, and zero-sum game-theoretic. (bottom row) Partially-observable state and intent policy is supervised by the corresponding policy above.

or going straight determined randomly upon the start of the episode (details in project page). If the pursuer has a high-quality understanding of the evader's latent intent, it should be able to intercept it along its weaving path. Throughout this section, all policies observe *ground-truth* relative states but not the evader's latent intent.

We consider three approaches: a **reactive** pursuer policy which only observes the present relative state and does not infer any latent intent, a **lookback** policy which must estimate the latent intent from a history of relative states *without supervision from the future* [42], [45] and a **lookahead** policy (ours), which uses a history of relative states and supervision from the future to predict a latent representation of the evader's future trajectory. The **lookback** and **lookahead** policies use identical LSTM architectures for intent estimation.

The **reactive** policy fails to predict the evader's behavior and is unable to do better than tracking the evader and trailing behind it (left, Fig. 4). While adding a history improves the pursuer's strategy (center, Fig. 4), it still struggles to estimate the evader's intent reliably. In contrast, the **lookahead** policy, trained to predict a latent representation of the evader's future trajectory, learns effective predictive behaviors (right, Fig. 4). In addition, the **lookahead** policy converges 10 times faster than the **lookback** one (see project page). Overall, our experiments show that using the future trajectory as privileged information favors training and enables strategic behaviors.

B. For distillation, balance interaction diversity & optimality

Influence of evader model on pursuer policy. Now that we have a teacher policy architecture, we turn to the role of the evader on the teacher pursuer policy. We compute three fully-observable teacher policies, π^*, trained against three different evader models, π^{e}. With a slight abuse of notation, let x^{rel} be the relative state in the evader's body frame. The **random** evader, $\pi^{\text{e}}_{\text{rand}}(t)$, randomly samples a set of controls to apply each 1-3 seconds. The **multi-agent RL** evader, $\pi^{\text{e}}_{\text{marl}}(x_t^{\text{rel}})$, is trained to evade a pre-trained pursuer policy, equivalent to a single iteration of [19]. Finally, assuming perfect relative state, our setting could be modelled by a zero-sum **game theory** model, whose solution characterizes the optimal pair of policies for the pursuer and the evader [17]. We compute $\pi^{\text{e}}_{\text{game}}(x_t^{\text{rel}})$ via an off-the-shelf dynamic game solver [58] on a simplified dynamics models of both agents (Dubins' cars). Although the resulting policy performance decreases when applied on a simulated legged robot system, it still exhibits the same time-to-capture trends. Details on all evaders in project page. Top row in Fig. 5 shows that with perfect state, the pursuer capture time is indistinguishable between **random** and **MARL** evaders, while the optimal **game theory** evader maximally exploits the interaction.

Distillation to partially-observable policy. After training the fully-observable pursuer policy against each evader, we supervise the corresponding partially-observable pursuer policy (bottom row, Fig. 5). We find that the **game-theoretic** pursuer policy makes for a poor supervisor because the supervision and interaction data operate under a perfect state-feedback Nash equilibrium assumption that is too hard to satisfy for the partially-observable policy. In contrast, robots trained against noisily-optimal (**MARL**) or extremely diverse (**random**) evaders have smaller in-distribution performance drops. This indicates that the interaction assumptions under which the teacher policy is obtained must be feasible for the partially-observable student.

C. To adapt, coverage is more helpful than specialization

Ultimately, the test-time distribution of the evader is unknown a priori. Thus, we ask *"How quickly and how effectively can different partially-observable pursuer policies adapt to an out of distribution evader?"* We simulate interaction between $\pi^{\text{p}}_{\text{rand}}, \pi^{\text{p}}_{\text{marl}}, \pi^{\text{p}}_{\text{game}}$ and the highly predictable Dubins' agent (see here for more details). None of the pursuers were trained on this behavior. We collect batches of joint state trajectories, and then finetune the weights of

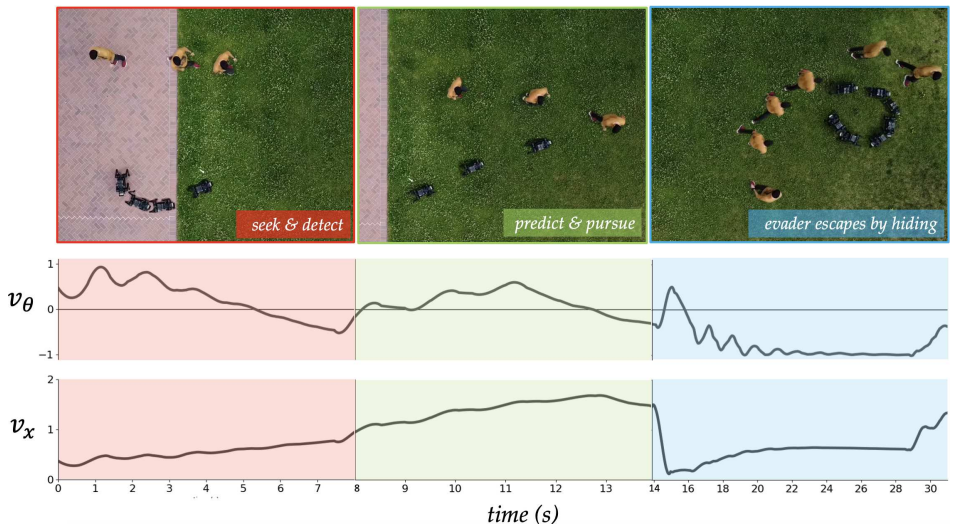

Fig. 6: Single-shot interaction between a vision-based pursuer policy and a human. **Left:** Since the human starts outside the FOV of the robot, the latter turns and seeks until it gets the first detection. **Center:** The robot predicts the human will go straight and gallops to where the person will be. **Right:** Human strategically hides outside robot's FOV to escape.

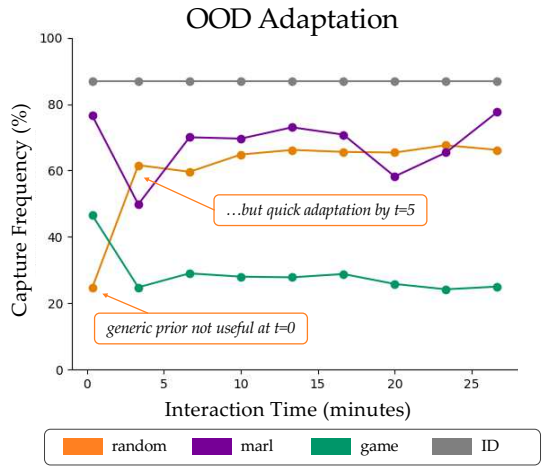

Fig. 7: Coverage is better than specialization for adaptation.

the pursuer's encoder $\mathcal{E}$ by supervising the latent $\hat{z}_t$ at each timestep via the privileged encoder $\mathcal{E}^*$. Since $\pi^{\text{p}}_{\text{rand}}$ has a generic prior on the evader motion, the representation in $\mathcal{E}$ is rich enough to triple its capture frequency in just 5 min. (Fig. 7). $\pi^{\text{p}}_{\text{marl}}$ is good at the start, but, due to its prior on the evader motion, it is less flexible and needs more data to see improvements. $\pi^{\text{p}}_{\text{game}}$, with a stronger prior than $\pi^{\text{p}}_{\text{marl}}$, fails to adapt and reaches a sub-optimal performance with the limited data. This indicates that to quickly adapt to agents "in the wild", which are neither random nor optimal, coverage is more helpful than specialization.

VI. Real-World Results

Real-world interactions are out-of-distribution for two main reasons: (1) the behavior of the evader is unscripted and possibly very different to what was observed in simulation, and (2) the physical dynamics of the evader do not follow the unicycle model as in simulation. We run two sets of experiments to study how our policies react to such conditions.

First, we ablate the pursuer policy and deploy $\pi^{\text{p}}_{\text{rand}}$, $\pi^{\text{p}}_{\text{marl}}$, $\pi^{\text{p}}_{\text{game}}$ on a physical quadruped robot to interact with a human. We observe that $\pi^{\text{p}}_{\text{rand}}$ and $\pi^{\text{p}}_{\text{marl}}$ perform qualitatively similarly. They showcase information-seeking motions when the evader is not in the field of view and predictive strategies when the evader is visible, i.e., heading towards where the evader *will be*, not where *it is* (Fig. 6). However, $\pi^{\text{p}}_{\text{marl}}$ shows slightly better anticipation and faster reaction times. Conversely, $\pi^{\text{p}}_{\text{game}}$ shows inefficient information-seeking motions, taking long detours to reach the evader. Such performance, inline with the experiments from Sec. V, confirms that the diversity of interaction data collected by a game-theoretic supervisor is not high enough to be robust to real-world interactions.

Second, we keep the pursuer policy fixed and ablate the evader dynamics. Concretely, we compare the performance of $\pi^{\text{p}}_{\text{rand}}$ when interacting against a human or another quadruped teleoperated by a human (Fig. 1). In both cases, we observe aspects of strategic behavior. However, such behavior is more apparent during interaction with another robot. This is due to the robot's dynamics being closer to the unicycle model the policy was trained on in simulation.

VII. Conclusion

This paper takes the first steps toward learning vision-based robot policies that can reason strategically through partially-observable physical state and latent intent. We find interesting pursuer behaviors when deployed on a physical quadruped robot with an RGB-D camera: information gathering under uncertainty, intent prediction from noisy state estimates, and anticipation of agents' motion. One limitation of our current approach is that it does not model the affordances of the environment, like obstacles, that could be strategically used by the pursuer. An exciting avenue for future work is for the robot to sense and tractably plan through such environment and terrain geometry.

REFERENCES

[1] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[2] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, and M. Hutter, "Fast traversability estimation for wild visual navigation," *Robotics: Science and Systems*, 2023.

[3] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

[4] A. Loquercio, A. Kumar, and J. Malik, "Learning Visual Locomotion with Cross-Modal Supervision," in *ICRA*, 2023.

[5] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, "Agile autonomous driving using end-to-end deep imitation learning," *arXiv preprint arXiv:1709.07174*, 2017.

[6] A. Loquercio, E. Kaufmann, R. Ranftl, M. Müller, V. Koltun, and D. Scaramuzza, "Learning high-speed flight in the wild," *Science Robotics*, vol. 6, no. 59, p. eabg5810, 2021.

[7] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," *Nature*, vol. 620, no. 7976, pp. 982–987, Aug 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06419-4

[8] C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Iterative residual policy: for goal-conditioned dynamic manipulation of deformable objects," *Robotics: Science and Systems*, 2022.

[9] A. M. Wilson, J. Lowe, K. Roskilly, P. E. Hudson, K. Golabek, and J. McNutt, "Locomotion dynamics of hunting in wild cheetahs," *Nature*, vol. 498, no. 7453, pp. 185–189, 2013.

[10] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020, pp. 66–75.

[11] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[12] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.

[13] R. Isaacs, "Differential games i: Introduction," RAND CORP SANTA MONICA CA SANTA MONICA, Tech. Rep., 1954.

[14] ——, *Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization*. Courier Corporation, 1999.

[15] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[16] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

[17] I. M. Mitchell, A. M. Bayen, and C. J. Tomlin, "A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games," *IEEE Transactions on automatic control*, vol. 50, no. 7, pp. 947–957, 2005.

[18] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[19] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2817–2826.

[20] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in multi-agent systems and applications-1*, pp. 183–221, 2010.

[21] Y. Tang, J. Tan, and T. Harada, "Learning agile locomotion via adversarial training," in *2020 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. IEEE, 2020, pp. 6098–6105.

[22] K.-C. Hsu, D. P. Nguyen, and J. F. Fisac, "Isaacs: Iterative soft adversarial actor-critic for safety," *L4DC*, 2022.

[23] Z. Zhang and J. F. Fisac, "Safe occlusion-aware autonomous driving via game-theoretic active perception," *arXiv preprint arXiv:2105.08169*, 2021.

[24] W. Schwarting, A. Pierson, S. Karaman, and D. Rus, "Stochastic dynamic games in belief space," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 2157–2172, 2021.

[25] I. Suzuki and M. Yamashita, "Searching for a mobile intruder in a polygonal region," *SIAM Journal on computing*, vol. 21, no. 5, pp. 863–888, 1992.

[26] B. P. Gerkey, S. Thrun, and G. Gordon, "Visibility-based pursuit-evasion with limited field of view," *The International Journal of Robotics Research*, vol. 25, no. 4, pp. 299–315, 2006.

[27] S. D. Bopardikar, F. Bullo, and J. P. Hespanha, "On discrete-time pursuit-evasion games with sensing limitations," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1429–1439, 2008.

[28] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbrück, "Steering a predator robot using a mixed frame/event-driven convolutional neural network," in *2016 Second international conference on event-based control, communication, and signal processing (EBCCSP)*. IEEE, 2016, pp. 1–8.

[29] Z. Zhang, X. Wang, Q. Zhang, and T. Hu, "Multi-robot cooperative pursuit via potential field-enhanced reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8808–8814.

[30] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[31] M. Woodward, C. Finn, and K. Hausman, "Learning to interactively learn and assist," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 03, 2020, pp. 2535–2543.

[32] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[33] D.-K. Kim, M. Riemer, M. Liu, J. Foerster, M. Everett, C. Sun, G. Tesauro, and J. P. How, "Influencing long-term behavior in multiagent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 808–18 821, 2022.

[34] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent interaction," *Machine Learning, Cornell University*, 2019.

[35] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[36] M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu *et al.*, "Human-level play in the game of diplomacy by combining language models with strategic reasoning," *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.

[37] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[38] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking and its real-world deployment via reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 6, pp. 1317–1332, 2019.

[39] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "AD-VAT: An asymmetric dueling mechanism for learning visual active tracking," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HkgYmhR9KX

[40] F. Zhong, X. Bi, Y. Zhang, W. Zhang, and Y. Wang, "Rspt: Reconstruct surroundings and predict trajectories for generalizable active object tracking," *arXiv preprint arXiv:2304.03623*, 2023.

[41] M. Asada and H. Kitano, "The robocup challenge," *Robotics and Autonomous Systems*, vol. 29, no. 1, pp. 3–12, 1999.

[42] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh, "Learning latent representations to influence multi-agent interaction," in *Conference on robot learning*. PMLR, 2021, pp. 575–588.

[43] S. Parekh, S. Habibian, and D. P. Losey, "Rili: Robustly influencing latent intent," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 01–08.

[44] W. Z. Wang, A. Shih, A. Xie, and D. Sadigh, "Influencing towards stable multi-agent interactions," in *Conference on robot learning*. PMLR, 2022, pp. 1132–1143.

[45] J. Z.-Y. He, Z. Erickson, D. S. Brown, A. Raghunathan, and A. Dragan, "Learning representations that enable generalization in assistive tasks," in *Conference on Robot Learning*. PMLR, 2023, pp. 2105–2114.

[46] X. Huang, Z. Li, Y. Xiang, Y. Ni, Y. Chi, Y. Li, L. Yang, X. B. Peng, and K. Sreenath, "Creating a dynamic quadrupedal robotic goalkeeper with reinforcement learning," *arXiv preprint arXiv:2210.04435*, 2022.

[47] O. Nachum, M. Ahn, H. Ponte, S. Gu, and V. Kumar, "Multi-agent manipulation via locomotion using hierarchical sim2real," *Conference on Robot Learning*, 2019.

[48] R. T. Fawcett, L. Amanzadeh, J. Kim, A. D. Ames, and K. A. Hamed, "Distributed data-driven predictive control for multi-agent

collaborative legged locomotion," *arXiv preprint arXiv:2211.06917*, 2022.
[49] J. Kim, J. Lee, and A. D. Ames, "Safety-critical coordination for cooperative legged locomotion via control barrier functions," *arXiv preprint arXiv:2303.13630*, 2023.
[50] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
[51] "Zed 2 camera," https://www.stereolabs.com/zed-2/, accessed: 2023-05-08.
[52] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, "Multimodal sensor fusion with differentiable filters," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 444–10 451.
[53] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[54] V. Tadic, A. Toth, Z. Vizvari, M. Klincsik, Z. Sari, P. Sarcevic, J. Sarosi, and I. Biro, "Perspectives of realsense and zed depth sensors for robotic vision applications," *Machines*, vol. 10, no. 3, p. 183, 2022.
[55] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
[56] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 91–100.
[57] L. E. Dubins, "On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents," *American Journal of mathematics*, vol. 79, no. 3, pp. 497–516, 1957.
[58] I. M. Mitchell, "The flexible, extensible and efficient toolbox of level set methods," *Journal of Scientific Computing*, vol. 35, pp. 300–329, 2008.