PA-MAPPO: PHASE-AWARE POLICY OPTIMIZATION FOR COOPERATIVE MULTI-UAV PURSUIT UNDER PARTIAL OBSERVABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative multi-UAV pursuit offers a flexible means of intercepting unauthorized UAVs in urban low-altitude airspace. However, limited fields of view and building occlusions require the UAV team to search without current target information, coordinate interception after detection, and reacquire the evading UAV after visual contact is lost. Existing multi-agent reinforcement learning methods often rely on uniform objectives, reactive target following, and a single team-level value estimate, limiting phase-aware decision making and differentiated credit assignment. We propose Phase-Aware Multi-Agent Proximal Policy Optimization (PA-MAPPO), a centralized-training and decentralized-execution framework organized around search, pursuit, and reacquisition. To anticipate evader motion, PA-MAPPO introduces an attention-based LSTM encoder–decoder to predict a multi-step evader trajectory, providing interception guidance during visual contact and after temporary occlusion. To coordinate complementary approaches, we introduce an interception-oriented task representation that encodes relative-motion geometry and interception feasibility. To improve credit assignment, an individual–team dual-critic architecture separately learns agent-specific residual returns and the mean team return. A phase-dependent reward further aligns the learning signals with spatial coverage during search, coordinated encirclement during pursuit, and directed motion toward the last observed target position during reacquisition. Comparisons with representative multi-agent reinforcement learning baselines demonstrate improved capture reliability, mission efficiency, and visual-contact recovery. Ablation studies further verify the contributions of trajectory prediction, task representation, and dual-critic learning.

1 INTRODUCTION

The growing availability of commercial unmanned aerial vehicles (UAVs) has increased the risk of unauthorized operations in urban low-altitude airspace (Peng et al., 2025). Existing counter-UAV systems combine radio-frequency sensing, radar, acoustic sensing, and vision with signal interference or physical interception. Their deployment in populated areas nevertheless remains difficult because buildings constrain flight paths and obstruct sensing (Shi et al., 2018; Peng et al., 2025). Cooperative pursuit by multiple UAVs provides an alternative means of intercepting a non-cooperative target and is an important problem in decentralized swarm control (Peng et al., 2025; Feng et al., 2025). In practice, finite sensing ranges, limited fields of view, and line-of-sight occlusions require the pursuing UAVs to search for the evading UAV, coordinate after detection, and reacquire it after visual contact is lost.

This task can be formulated as a decentralized partially observable Markov decision process (Dec-POMDP) (Oliehoek & Amato, 2016). Centralized training with decentralized execution (CTDE) uses additional information during training while retaining decentralized policies during execution (Lowe et al., 2017). MAPPO is a strong cooperative baseline because it combines CTDE with stable on-policy optimization (Yu et al., 2022; Ding et al., 2025). Although predicted target trajectories can improve decisions under partial observation (Wu et al., 2026), prediction alone does not specify complementary interception positions for different UAVs. Standard MAPPO also uses a uniform

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

objective across search, pursuit, and reacquisition, while a single team-level critic may provide insufficient credit information for heterogeneous cooperative behaviors (Wang et al., 2025).

To address this, we propose Phase-Aware MAPPO (PA-MAPPO), which integrates target-motion prediction, interception geometry, differentiated value estimation, and phase-dependent rewards within a CTDE framework. The resulting policy supports anticipatory interception and coordinated encirclement across the three task phases.

Our main contributions are as follows:

- We introduce an attention-based LSTM encoder–decoder that predicts a multi-step evader trajectory, providing prospective interception guidance during visual contact and short-horizon motion information after temporary occlusion.
- We construct an interception-oriented task representation that encodes relative-motion geometry and interception feasibility, encouraging the pursuing UAVs to approach from complementary directions.
- We propose an individual–team dual-critic architecture that separately learns agent-specific residual returns and the mean team return, providing differentiated credit signals for local maneuvering and team cooperation.
- We design a phase-dependent reward that promotes spatial coverage during search, angularly distributed and contracting encirclement during pursuit, and directed target reacquisition after visual contact is lost.

2 Problem Formulation

We consider cooperative multi-UAV pursuit in an urban workspace containing static obstacles. A team of pursuers must search for, track, and capture a non-cooperative evader under motion constraints and intermittent target visibility.

2.1 Pursuit Scenario

Let $\mathcal{I} = \{1, \dots, N\}$ denote the pursuer set and let e denote the evader. All UAVs move in a bounded planar workspace $\mathcal{W} = [-L, L]^2$. The workspace contains M rectangular obstacles $\mathcal{B} = \{B_m\}_{m=1}^{M}$ that block both UAV motion and line of sight. The collision-free region is

$$\mathcal{W}_{\text{free}} = \mathcal{W} \setminus \bigcup_{m=1}^{M} B_m.$$

At time t, pursuing UAV i has position $\mathbf{p}_i^t \in \mathbb{R}^2$, velocity $\mathbf{v}_i^t$, and heading ψ_i^t. The corresponding variables of the evading UAV are $\mathbf{p}_e^t$, $\mathbf{v}_e^t$, and ψ_e^t. Define the pursuer–evader distance as

$$d_i^t = \left\| \mathbf{p}_i^t - \mathbf{p}_e^t \right\|_2,$$

where $\| \cdot \|_2$ denotes the Euclidean norm. Capture occurs when

$$\min_{i \in \mathcal{I}} d_i^t \le r_{\text{cap}},$$

where r_{cap} is the capture radius.

An episode terminates upon capture or when any pursuing or evading UAV collides with an obstacle or leaves the workspace. Reaching the maximum horizon T is treated as truncation. A commanded motion rejected or redirected by the collision-avoidance mechanism is recorded as a safety intervention and does not by itself terminate the episode.

2.2 Planar Kinematic Model

Each pursuer selects a continuous action $\mathbf{a}_i^t = (a_{x,i}^t, a_{y,i}^t) \in [-1, 1]^2$. If $\|\mathbf{a}_i^t\|_2$ exceeds a small threshold, the desired heading is $\bar{\psi}_i^t = \operatorname{atan2}(a_{y,i}^t, a_{x,i}^t)$. Otherwise, the desired heading remains

unchanged. The state update is

$$\psi_i^{t+1} = \psi_i^t + \mathrm{clip}\big(\mathrm{wrap}(\bar{\psi}_i^t - \psi_i^t), -\Delta\psi_p, \Delta\psi_p\big)\,,$$
$$\mathbf{p}_i^{t+1} = \mathbf{p}_i^t + d_p \begin{bmatrix}\cos\psi_i^{t+1}\\ \sin\psi_i^{t+1}\end{bmatrix},$$

where d_p is the displacement per decision step and $\Delta\psi_p$ is the maximum heading change. If the candidate position is occupied, the simulator selects the feasible direction closest to the commanded direction. The evader follows the same planar kinematic model with displacement d_e, while its heading is determined by a predefined evasive rule.

2.3 Partial Observations and Task Phases

Each pursuer carries a forward-facing sensor with sensing range R and field-of-view angle θ_{fov}. Let ρ_i^t and α_i^t denote the distance and relative bearing from pursuer i to the evader. Direct visibility is defined as

$$c_i^t = \mathbb{I}\left[\rho_i^t \le R,\ |\alpha_i^t| \le \frac{\theta_{\mathrm{fov}}}{2},\ \mathrm{LOS}(\mathbf{p}_i^t, \mathbf{p}_e^t) = 1\right],$$

where $\mathrm{LOS} = 0$ if an obstacle intersects the line segment between the pursuer and the evader. When at least one pursuer detects the evader, the detected target information is shared with the team.

The observation of pursuer i is grouped as

$$o_i^t = \left[o_{i,\mathrm{self}}^t,\ o_{i,\mathrm{target}}^t,\ o_{i,\mathrm{obstacle}}^t,\ o_{i,\mathrm{team}}^t,\ o_{i,\mathrm{memory}}^t\right].$$

These components describe the pursuer motion state, visible or communicated target information, local obstacle measurements, teammate information, and stored target information.

Let $v_t = \mathbb{I}[\sum_i c_i^t > 0]$ indicate team-level target visibility, and let q_t indicate whether a valid last-seen target record exists. The current task phase is

$$\xi_t = \begin{cases}\text{search}, & v_t = 0,\ q_t = 0,\\ \text{pursuit}, & v_t = 1,\\ \text{reacquisition}, & v_t = 0,\ q_t = 1.\end{cases}$$

When the target is lost after detection, its last observed position is retained for a fixed period. Detecting the target again returns the team to pursuit. If the stored information expires, the team resumes search.

2.4 Dec-POMDP Formulation and Objective

We formulate cooperative pursuit as a Dec-POMDP

$$\mathcal{M} = \left\langle \mathcal{I}, \mathcal{S}, \rho_0, \{\mathcal{A}_i\}_{i=1}^N, \mathcal{P}, \{\Omega_i\}_{i=1}^N, \mathcal{O}, \mathcal{R}, \gamma \right\rangle.$$

Here, $\mathcal{I} = \{1, \dots, N\}$ is the set of pursuing UAVs, and $\mathcal{S}$ is the global state space. A state is represented as

$$s^t = \left(\{\mathbf{p}_i^t, \mathbf{v}_i^t, \psi_i^t\}_{i=1}^N, \mathbf{p}_e^t, \mathbf{v}_e^t, \psi_e^t, \mathcal{B}, \mathbf{m}^t\right),$$

where $\mathbf{m}^t$ contains the task phase, coverage record, last-observed target state, and associated timers required to preserve the Markov property. The distribution ρ_0 specifies the initial UAV states and obstacle layout.

Each pursuing UAV has a continuous action space $\mathcal{A}_i = [-1, 1]^2$, and the joint action is $\mathbf{a}^t = (\mathbf{a}_1^t, \dots, \mathbf{a}_N^t)$. The transition kernel $\mathcal{P}(s^{t+1} \mid s^t, \mathbf{a}^t)$ is induced by the UAV kinematics, evasion strategy, collision handling, and task-phase transitions. The observation space Ω_i contains the locally available information of UAV i, while $\mathcal{O}(\mathbf{o}^{t+1} \mid s^{t+1}, \mathbf{a}^t)$ accounts for sensing range, field of view, line-of-sight occlusion, and target-information sharing.

The implementation produces an agent-dependent reward r_i^t, and the cooperative reward is defined as

$$\bar{r}^t = \frac{1}{N}\sum_{i=1}^N r_i^t.$$

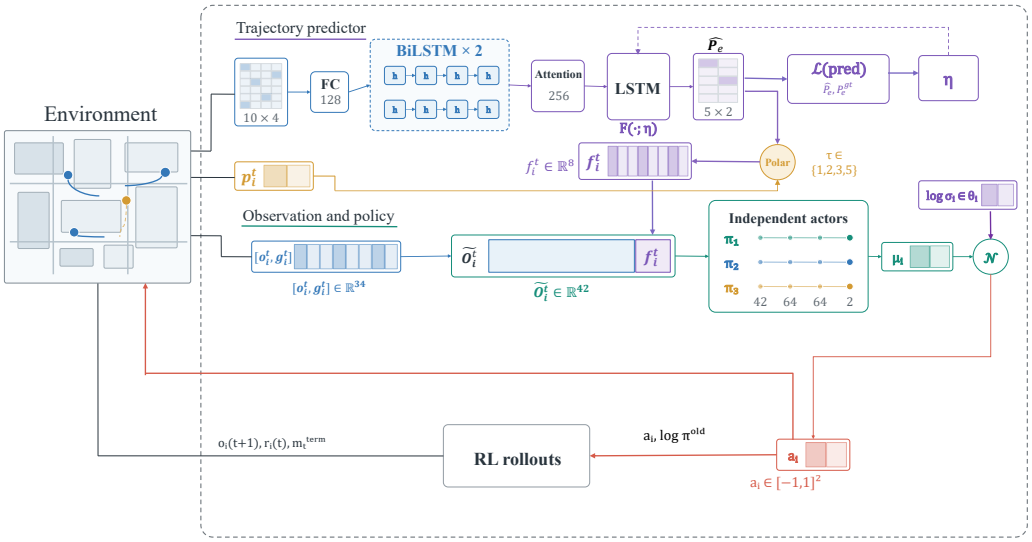

Figure 1: Decentralized interaction in PA-MAPPO.

The reward function $\mathcal{R}$ evaluates phase-dependent task progress, capture, and safety. With independently parameterized decentralized policies $\boldsymbol{\pi}_\Theta = \prod_{i=1}^{N} \pi_{\theta_i}$, the objective is

$$J(\Theta) = \mathbb{E}_{\boldsymbol{\pi}_\Theta, \mathcal{P}} \left[\sum_{t=0}^{T-1} \gamma^t \bar{r}^t \right],$$

where $\gamma \in [0, 1)$ is the discount factor. During execution, each actor uses only its own observation; centralized information is used only by the critics during training.

3 Methodology

3.1 Overview

Figures 1 and 2 illustrate the decentralized interaction and centralized optimization processes of PA-MAPPO. During interaction, the trajectory predictor maps the latest contiguous team-observed evader history to a five-step world-frame trajectory. Pursuer-relative prediction features $\mathbf{f}_i^t$ and the interception-oriented representation $\mathbf{g}_i^t$ augment the local observation $\mathbf{o}_i^t$. Each independently parameterized actor then selects an action from its own augmented observation. During centralized training, the collected rollouts update the individual critic, team critic, and decentralized actors, while the trajectory predictor is trained separately using supervised history–future pairs.

3.2 Multi-Step Evader Trajectory Prediction

Let $\tau_t \leq t$ be the latest time at which at least one pursuing UAV observes the evading UAV. The predictor uses the latest contiguous team-visible history

$$\mathcal{H}_t = \left(\mathbf{h}_e^{\tau_t - K + 1}, \ldots, \mathbf{h}_e^{\tau_t}\right), \qquad \mathbf{h}_e^{\tau} = [\mathbf{p}_e^{\tau}; \mathbf{v}_e^{\tau}],$$

where $K = 10$. Positions are expressed relative to $\mathbf{p}_e^{\tau_t}$, and position displacements and velocities are normalized by the evader motion scale s_m.

A linear projection maps each history element to a 128-dimensional feature. A two-layer bidirectional LSTM produces temporal representations $\{\mathbf{z}_j\}_{j=1}^{K}$, which are summarized by additive attention:

$$\alpha_j = \frac{\exp(f_{\text{att}}(\mathbf{z}_j))}{\sum_{\ell=1}^{K} \exp(f_{\text{att}}(\mathbf{z}_\ell))}, \qquad \mathbf{c}_t = \sum_{j=1}^{K} \alpha_j \mathbf{z}_j.$$

The context $\mathbf{c}_t$ initializes an autoregressive LSTM decoder. Starting from the last observed velocity as a constant-velocity prior, the decoder predicts

$$\widehat{\mathbf{P}}_{\tau_t} = F_\eta(\mathcal{H}_t) = \left(\widehat{\mathbf{p}}_e^{\tau_t+1}, \ldots, \widehat{\mathbf{p}}_e^{\tau_t+H}\right), \qquad H = 5.$$

The predictor is trained separately from PPO using simulator-generated history–future pairs and a horizon-weighted Smooth-L1 loss. Only team-observed history is used during execution; no unobserved evader state is inserted after occlusion.

The shared world-frame prediction is converted into pursuer-relative range–bearing features:

$$\widetilde{\rho}_{i,k}^t = \frac{\left\|\widehat{\mathbf{p}}_e^{\tau_t+k} - \mathbf{p}_i^t\right\|_2}{D_{\max}}, \qquad \widetilde{\beta}_{i,k}^t = \frac{1}{\pi}\operatorname{atan2}\left(\widehat{p}_{e,y}^{\tau_t+k} - p_{i,y}^t, \widehat{p}_{e,x}^{\tau_t+k} - p_{i,x}^t\right).$$

Using $\mathcal{K} = \{1, 2, 3, 5\}$,

$$\mathbf{f}_i^t = \operatorname{concat}_{k\in\mathcal{K}}\left[\widetilde{\rho}_{i,k}^t, \widetilde{\beta}_{i,k}^t\right] \in \mathbb{R}^8.$$

Thus, the UAVs share one predicted trajectory but obtain different policy features through their respective coordinate transformations. The decoder recurrence and supervised objective are detailed in Appendix A.1.

3.3 Interception-Oriented Task Representation

Trajectory prediction describes prospective target motion, whereas the task representation characterizes the current interception geometry. When the evading UAV is visible to the team, define

$$\mathbf{r}_i^t = \mathbf{p}_i^t - \mathbf{p}_e^t, \qquad d_i^t = \|\mathbf{r}_i^t\|_2, \qquad \widehat{\mathbf{r}}_i^t = \frac{\mathbf{r}_i^t}{d_i^t + \epsilon}, \qquad \widehat{\mathbf{v}}_e^t = \frac{\mathbf{v}_e^t}{\|\mathbf{v}_e^t\|_2 + \epsilon}.$$

The interception-oriented representation is

$$\mathbf{g}_i^t = v_t \begin{bmatrix} (\widehat{\mathbf{r}}_i^t)^\top \widehat{\mathbf{v}}_e^t \\ \widehat{r}_{i,x}^t \widehat{v}_{e,y}^t - \widehat{r}_{i,y}^t \widehat{v}_{e,x}^t \\ \mathbb{I}[d_{\min} \le d_i^t \le d_{\max}]\mathbb{I}[(\widehat{\mathbf{r}}_i^t)^\top \widehat{\mathbf{v}}_e^t > 0] \\ (-\widehat{\mathbf{r}}_i^t)^\top \widehat{\mathbf{v}}_e^t \end{bmatrix}.$$

Its four components encode the longitudinal relation, lateral relation, interception feasibility, and approach direction. The representation is set to zero when the target is unavailable or its velocity is negligible.

The augmented actor input is

$$\widetilde{\mathbf{o}}_i^t = \operatorname{concat}\left(\mathbf{o}_i^t, \mathbf{g}_i^t, \mathbf{f}_i^t\right).$$

Each actor maps only its own augmented observation to a continuous action; no centralized critic input or another UAV's observation is required during execution.

3.4 Individual–Team Dual-Critic Learning

Figure 2 summarizes the centralized optimization process. The augmented observations are first concatenated as

$$\overline{\mathbf{o}}_t = \operatorname{concat}\left(\widetilde{\mathbf{o}}_1^t, \ldots, \widetilde{\mathbf{o}}_N^t\right).$$

The team and individual critics are

$$V_t^{\text{team}} = V_{\phi_{\text{team}}}(\overline{\mathbf{o}}_t), \qquad V_{i,t}^{\text{ind}} = V_{\phi_{\text{ind}}}\left(\overline{\mathbf{o}}_t, \widetilde{\mathbf{o}}_i^t, \mathrm{id}_i\right),$$

where id_i is the one-hot identity vector of pursuing UAV i. The individual critic shares parameters across UAVs, with id_i distinguishing their value estimates.

The critics use different reward signals:

$$\bar{r}^t = \frac{1}{N}\sum_{j=1}^N r_j^t, \qquad r_{i,\text{res}}^t = r_i^t - \bar{r}^t.$$

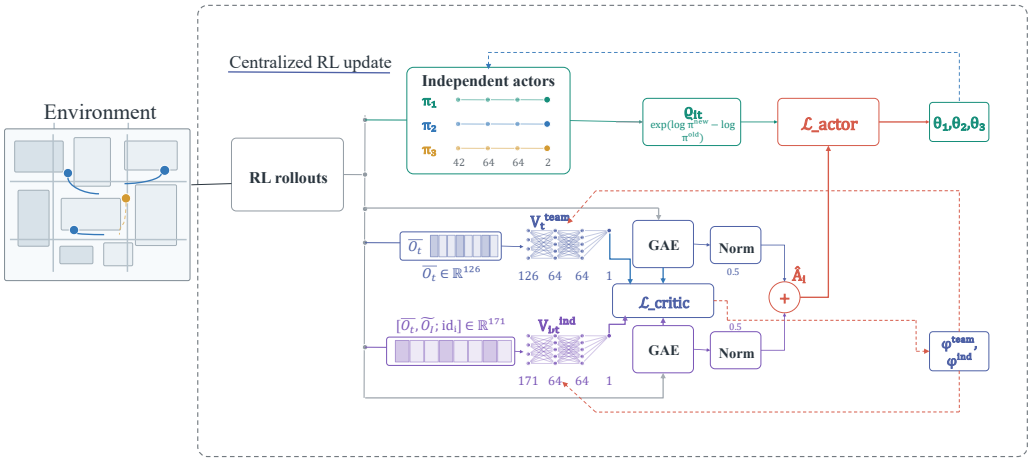

Figure 2: Centralized optimization in PA-MAPPO.

The team critic learns cooperative task progress from $\bar{r}^t$, whereas the individual critic learns each UAV's deviation from the team average through $r^t_{i,\mathrm{res}}$. Their separately normalized advantages are combined as

$$\widehat{A}_{i,t} = w_{\mathrm{ind}}\,\mathrm{Norm}\left(\widehat{A}^{\mathrm{ind}}_{i,t}\right) + w_{\mathrm{team}}\,\mathrm{Norm}\left(\widehat{A}^{\mathrm{team}}_{t}\right), \qquad w_{\mathrm{ind}} + w_{\mathrm{team}} = 1.$$

We use $w_{\mathrm{ind}} = w_{\mathrm{team}} = 0.5$. Each critic is fitted to its own return target using a clipped value loss. The complete TD residuals, generalized advantage estimators, return targets, and value-clipping objective are given in Appendix A.2.

3.5 Phase-Dependent Rewards and Policy Update

Sparse capture feedback is insufficient for learning search, coordinated pursuit, and target reacquisition. We therefore select the shaping reward according to the current phase ξ_t:

$$r^t_i = r_{\mathrm{step}} + r^t_{i,\mathrm{phase}} + r^t_{i,\mathrm{safety}} + R_{\mathrm{vis}} b^t_{i,\mathrm{vis}} + R_{\mathrm{cap}} b^t_{\mathrm{cap}},$$

where

$$r^t_{i,\mathrm{phase}} = \begin{cases} r^t_{i,\mathrm{search}}, & \xi_t = \mathrm{search}, \\ r^t_{i,\mathrm{pursuit}}, & \xi_t = \mathrm{pursuit}, \\ r^t_{i,\mathrm{search}} + r^t_{i,\mathrm{reacq}}, & \xi_t = \mathrm{reacquisition}. \end{cases}$$

Here, r_{step} is a small time penalty, while $b^t_{i,\mathrm{vis}}$ and b^t_{cap} indicate a new detection and successful capture.

Search reward. Before detection, the team is encouraged to expand coverage and avoid redundant search:

$$r^t_{i,\mathrm{search}} = \lambda_{\mathrm{cov}} \Delta A^t_i + \lambda_{\mathrm{new}} b^t_{i,\mathrm{new}} - \lambda_{\mathrm{overlap}} m^t_{i,\mathrm{overlap}} + \lambda_{\mathrm{sep}} \frac{\min(\bar{d}^t_{\mathrm{pp}}, D_{\mathrm{sep}})}{D_{\mathrm{sep}}} + \lambda_{\mathrm{mov}} \|\mathbf{v}^t_i\|_2.$$

The terms measure coverage increase, entry into a new grid cell, redundant coverage, saturated pairwise separation, and movement, respectively.

Pursuit reward. When the target is visible, the reward combines individual approach and team-level encirclement:

$$r^t_{i,\mathrm{pursuit}} = \lambda_d r^t_{i,d} + \lambda_h r^t_{i,h} + \lambda_\theta u^t_\theta + \lambda_a \left[1 - \frac{\bar{d}^t}{D_{\mathrm{enc}}}\right]_+ + \lambda_c \mathbb{I}[\Delta \bar{d}^t > \delta_c] \Delta \bar{d}^t + r^t_{i,\mathrm{sta}},$$

where

$$\bar{d}^t = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}^t_i - \mathbf{p}^t_e\|_2, \qquad \Delta \bar{d}^t = \bar{d}^{t-1} - \bar{d}^t.$$

The terms $r_{i,d}^t$ and $r_{i,h}^t$ measure progress toward the current target or an interception point and alignment with the desired motion direction. Let K_t be the number of pursuing UAVs inside the effective encirclement radius and $\Delta\vartheta_k^t$ their ordered angular gaps around the evading UAV. Angular uniformity is measured by

$$u_\theta^t = \begin{cases} 1 - \min\left(1, \dfrac{1}{2\pi}\displaystyle\sum_{k=1}^{K_t}\left|\Delta\vartheta_k^t - \dfrac{2\pi}{K_t}\right|\right), & K_t \geq 2, \\ 0, & K_t < 2. \end{cases}$$

Distance and contraction terms prevent the UAVs from obtaining a high angular score while remaining far from the target.

Reacquisition reward. After visual contact is lost, the search reward remains active while additional guidance directs the UAVs toward the last observed target position $\mathbf{p}_{\mathrm{ls}}^t$. Let

$$d_{i,\mathrm{ls}}^t = \|\mathbf{p}_i^t - \mathbf{p}_{\mathrm{ls}}^t\|_2.$$

The additional reward is

$$r_{i,\mathrm{reacq}}^t = \lambda_{\mathrm{pot}}\left(d_{i,\mathrm{ls}}^{t-1} - d_{i,\mathrm{ls}}^t\right) + \lambda_{\mathrm{dir}} w_i^t \left[(\widehat{\mathbf{v}}_i^t)^\top \widehat{\mathbf{d}}_{i,\mathrm{ls}}^t\right]_+ + \lambda_{\mathrm{near}}\left[1 - \frac{d_{i,\mathrm{ls}}^t}{D_{\mathrm{near}}}\right]_+ + r_{i,\mathrm{mode}}^t.$$

The mode term switches from directed approach to local patrol after a pursuing UAV reaches the loss region.

Safety shaping is active in every phase and penalizes obstacle proximity, boundary proximity, abrupt heading changes, and terminal collisions. The actors use the standard clipped PPO objective with the combined individual–team advantage, while the trajectory predictor receives no gradient from PPO. Detailed reward definitions and stabilization terms are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Simulation environment. Experiments are conducted in a 400×400 planar urban workspace containing three pursuing UAVs, one evading UAV, and nine 60×60 buildings that obstruct motion and line of sight. Each pursuing UAV has a sensing range of 250 units and a 120° field of view. The pursuing and evading UAVs move at 3.0 and 3.15 units per step, respectively, and capture occurs within 12 units. Training episodes use a 300-step horizon, whereas evaluation episodes are truncated after 500 steps.

Training and evaluation protocol. Building layouts and initial UAV states are randomized during training subject to collision-free initialization. Each algorithm is trained with three independent random seeds and evaluated on the same 300 fixed test cases per seed, producing 900 evaluation episodes per algorithm. Metrics are first averaged over the 300 cases for each seed and then aggregated across the three seed-level means. Error bars denote 95% bootstrap confidence intervals computed using 4000 resamples of the seed-level means. Network and optimization settings are reported in Appendix A.4.

Evasion strategy. The evading UAV follows a repulsive-force strategy whose direction depends on the pursuing UAVs, nearby buildings, and workspace boundaries. This produces adaptive evasive motion while preventing the evading UAV from entering buildings or leaving the workspace.

4.2 Comparative Experiment

We compare PA-MAPPO with MAPPO (Yu et al., 2022), HAPPO (Kuba et al., 2022), HASAC (Liu et al., 2024), and FACMAC (Peng et al., 2021), representing on-policy CTDE, sequential heterogeneous-agent optimization, maximum-entropy heterogeneous-agent learning, and factored centralized value learning, respectively.

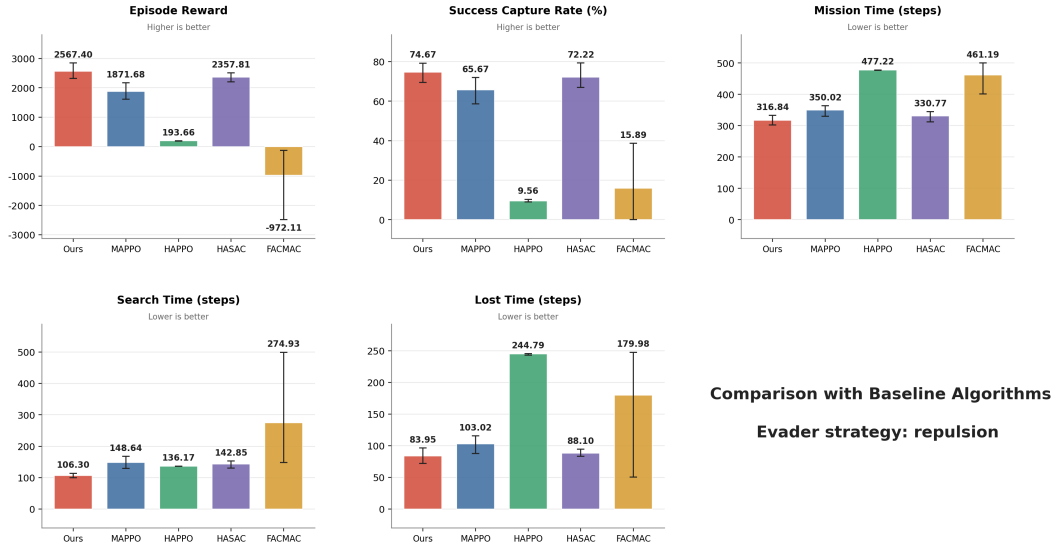

Figure 3: Comparison with representative multi-agent reinforcement learning algorithms under repulsive-force evasion. Bars show means over three training seeds, and error bars denote 95% confidence intervals.

We report Episode Reward, Success Capture Rate (SCR), Mission Time, Search Time, and Lost Time. Episode Reward is the accumulated team reward, and SCR is the proportion of episodes ending in capture. Mission Time and Search Time measure the steps to capture and first team-level detection, respectively, and are set to the maximum horizon for unsuccessful episodes. Lost Time counts the steps without team-level visual contact after the first detection in episodes containing at least one loss event. Higher Episode Reward and SCR are better, whereas lower time-based metrics are preferred.

Figure 3 shows that PA-MAPPO achieves the best mean result on all metrics. Its Episode Reward reaches 2567.40, exceeding HASAC and MAPPO by 8.89% and 37.17%, respectively. Its SCR of 74.67% exceeds HASAC by 2.45 percentage points and MAPPO by 9.00 percentage points. PA-MAPPO also obtains the lowest Mission Time (316.84), Search Time (106.30), and Lost Time (83.95), indicating more reliable capture, faster detection, and improved recovery after visual-contact loss.

4.3 Ablation Study

We evaluate three PA-MAPPO variants. *Conventional Twin Critics* replaces the proposed value decomposition with two centralized critics trained on the same team return. *w/o LSTM* removes multi-step trajectory prediction, whereas *w/o Task Embedding* removes the interception-oriented task representation.

As shown in Figure 4, Conventional Twin Critics causes the largest degradation: Episode Reward decreases from 2567.40 to 1834.55, SCR decreases from 74.67% to 25.83%, and Lost Time increases from 83.95 to 149.54 steps. Removing the LSTM reduces SCR to 65.83% and increases Lost Time to 111.24 steps, demonstrating the value of prospective motion information during pursuit and occlusion. Removing the task representation reduces SCR to 60.83% and increases Mission Time to 346.61 steps, showing that predicted positions alone do not specify complementary interception directions.

4.4 Behavior Analysis

Figure 5 illustrates the intended phase transitions. The pursuing UAVs initially follow different routes to expand coverage, switch to directed interception after detection, and continue toward the

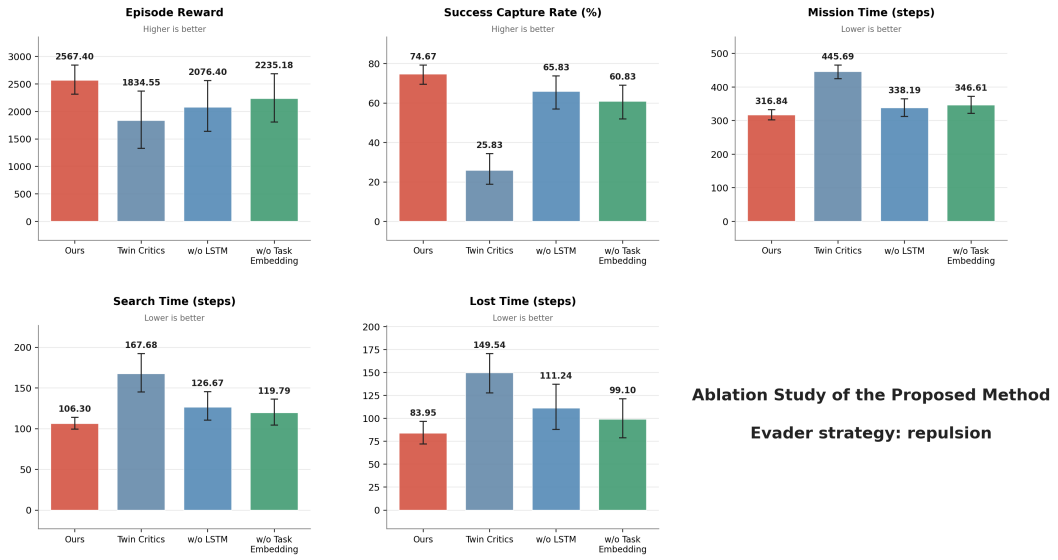

Figure 4: Ablation results under repulsive-force evasion. *Conventional Twin Critics* removes the individual–team value decomposition, *w/o LSTM* removes trajectory prediction, and *w/o Task Embedding* removes the interception-oriented task representation.

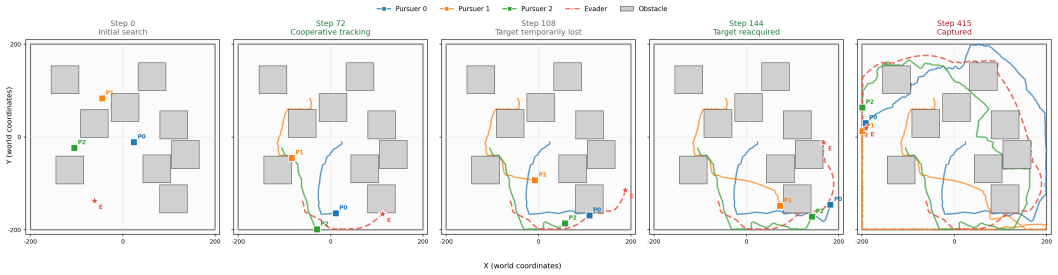

Figure 5: Representative PA-MAPPO rollout: initial search at step 0, pursuit after detection at step 72, visual-contact loss at step 108, target reacquisition at step 144, and capture at step 415.

predicted target region after occlusion. The team reacquires the evading UAV at step 144 and completes capture at step 415, illustrating coordinated behavior across search, pursuit, and reacquisition.

5 Conclusion

This paper introduced PA-MAPPO for cooperative multi-UAV pursuit under limited fields of view and repeated target occlusions. The framework combines multi-step evader-trajectory prediction, an interception-oriented task representation, individual–team dual-critic learning, and phase-dependent rewards. These mechanisms provide prospective motion guidance, complementary interception geometry, differentiated credit signals, and objectives aligned with search, pursuit, and reacquisition. Experiments against representative multi-agent reinforcement learning baselines demonstrate improvements in capture reliability, task efficiency, and visual-contact recovery. Ablation results further verify the contributions of trajectory prediction, interception-oriented task representation, and individual–team value decomposition.

References

Zhengkun Ding, Xingmei Wang, Chengtao Cai, Luyu Jia, and Zhou Xu. Multi-UAV intelligent decision-making method with layer delay dual-center MAPPO for air combat. *Applied Intelligence*, 55:811, 2025. doi: 10.1007/s10489-025-06654-4.

Yukai Feng, Zhengxing Wu, Jian Wang, Sijie Li, Yupei Huang, Junzhi Yu, and Min Tan. Decentralized reinforcement learning for multiple robotic fish in cooperative pursuit task. *IEEE Transactions on Cognitive and Developmental Systems*, 17(4):1022–1034, 2025. doi: 10.1109/TCDS.2025.3540071.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2109.11251.

Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, Qiang Fu, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2306.10715. URL https://arxiv.org/abs/2306.10715.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.02275.

Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016. doi: 10.1007/978-3-319-28929-8.

Brandon Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip H. S. Torr, Wendelin Boehmer, and Shimon Whiteson. FACMAC: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12208–12221, 2021.

Zhe Peng, Guohua Wu, Biao Luo, and Ling Wang. Multi-UAV cooperative pursuit strategy with limited visual field in urban airspace: A multi-agent reinforcement learning approach. *IEEE/CAA Journal of Automatica Sinica*, 12(7):1350–1367, 2025. doi: 10.1109/JAS.2024.124965.

Xiufang Shi, Chaoqun Yang, Weige Xie, Chao Liang, Zhiguo Shi, and Jiming Chen. Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges. *IEEE Communications Magazine*, 56(4), 2018. doi: 10.1109/MCOM.2018.1700430.

Yiqin Wang, Yufeng Wang, Feng Tian, Jianhua Ma, and Qun Jin. Intelligent games meeting with multi-agent deep reinforcement learning: A comprehensive review. *Artificial Intelligence Review*, 58:165, 2025. doi: 10.1007/s10462-025-11166-1.

Yixian Wu, Hongbin Li, Shunyu Tian, Ze Ji, and Changyun Wei. Cooperative encirclement of multi-USVs under communication constraints via prediction-augmented multi-agent reinforcement learning. *Ocean Engineering*, 364:126872, 2026. doi: 10.1016/j.oceaneng.2026.126872.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2103.01955.

A Additional Method Details

A.1 Trajectory-Predictor Recurrence and Loss

Let $\mathbf{c}_t \in \mathbb{R}^{256}$ be the attention-weighted encoder context. The decoder hidden and memory states are initialized by

$$\mathbf{d}_0 = \tanh(W_h \mathbf{c}_t + \mathbf{b}_h)\,, \qquad \mathbf{m}_0 = \tanh(W_m \mathbf{c}_t + \mathbf{b}_m)\,.$$

The normalized last observed velocity provides the constant-velocity prior

$$\boldsymbol{\delta}_0 = \frac{\mathbf{v}_e^{\tau_t}}{s_m}.$$

For prediction horizon $k \in \{1, \dots, H\}$, let $\mathbf{e}_k$ denote the learned horizon embedding. The autoregressive decoder applies

$$(\mathbf{d}_k, \mathbf{m}_k) = \mathrm{LSTMCell}\left([\boldsymbol{\delta}_{k-1}; \mathbf{e}_k], (\mathbf{d}_{k-1}, \mathbf{m}_{k-1})\right),$$

followed by the bounded velocity update

$$\boldsymbol{\delta}_k = \boldsymbol{\delta}_{k-1} + \frac{1}{2}\tanh\left(f_\Delta([\mathbf{d}_k; \mathbf{c}_t])\right).$$

The predicted world-frame position is obtained by accumulating the normalized displacements:

$$\widehat{\mathbf{p}}_e^{\tau_t+k} = \mathbf{p}_e^{\tau_t} + s_m \sum_{\ell=1}^{k} \boldsymbol{\delta}_\ell.$$

Thus, the decoder generates an ordered trajectory rather than independently predicting or repeating a single future position.

For training, define normalized ground-truth and predicted offsets as

$$\mathbf{q}_k = \frac{\mathbf{p}_e^{\tau_t+k} - \mathbf{p}_e^{\tau_t}}{s_m}, \qquad \widehat{\mathbf{q}}_k = \frac{\widehat{\mathbf{p}}_e^{\tau_t+k} - \mathbf{p}_e^{\tau_t}}{s_m}.$$

Their motion increments are

$$\Delta\mathbf{q}_1 = \mathbf{q}_1, \qquad \Delta\mathbf{q}_k = \mathbf{q}_k - \mathbf{q}_{k-1},$$

with the predicted increments defined analogously. The horizon weights increase linearly from 1 to 2:

$$\omega_k = 1 + \frac{k-1}{H-1}.$$

The prediction objective is

$$\mathcal{L}_{\mathrm{pred}} = \frac{\sum_{k=1}^{H} \omega_k \ell_{\mathrm{SL1}}\left(\widehat{\mathbf{q}}_k, \mathbf{q}_k\right)}{\sum_{k=1}^{H} \omega_k} + 0.25\, \ell_{\mathrm{SL1}}\left(\Delta\widehat{\mathbf{q}}_{1:H}, \Delta\mathbf{q}_{1:H}\right),$$

where ℓ_{SL1} denotes the Smooth-L1 loss averaged over spatial dimensions and training samples. The first term prioritizes accurate future positions, while the second regularizes the predicted motion increments.

A.2 Individual–Team Advantage Estimation

The team-average reward and agent-specific residual reward are

$$\bar{r}^t = \frac{1}{N}\sum_{j=1}^{N} r_j^t, \qquad r_{i,\mathrm{res}}^t = r_i^t - \bar{r}^t.$$

Let $d_{i,t} \in \{0, 1\}$ indicate termination for pursuing UAV i, and let

$$d_t = \max_i d_{i,t}$$

be the team termination indicator. The individual and team TD residuals are

$$\delta_{i,t}^{\mathrm{ind}} = r_{i,\mathrm{res}}^t + \gamma(1 - d_{i,t})V_{i,t+1}^{\mathrm{ind}} - V_{i,t}^{\mathrm{ind}},$$

$$\delta_t^{\mathrm{team}} = \bar{r}^t + \gamma(1 - d_t)V_{t+1}^{\mathrm{team}} - V_t^{\mathrm{team}}.$$

Separate generalized advantage estimates are computed as

$$\widehat{A}_{i,t}^{\mathrm{ind}} = \sum_{\ell=0}^{T-t-1} (\gamma\lambda)^\ell \left[\prod_{u=0}^{\ell-1}(1 - d_{i,t+u})\right] \delta_{i,t+\ell}^{\mathrm{ind}},$$

$$\widehat{A}_t^{\mathrm{team}} = \sum_{\ell=0}^{T-t-1} (\gamma\lambda)^\ell \left[\prod_{u=0}^{\ell-1}(1 - d_{t+u})\right] \delta_{t+\ell}^{\mathrm{team}}.$$

Normalization is performed separately for every individual advantage stream and for the team advantage. The policy advantage is

$$\widehat{A}_{i,t} = w_{\text{ind}} \operatorname{Norm}\left(\widehat{A}^{\text{ind}}_{i,t}\right) + w_{\text{team}} \operatorname{Norm}\left(\widehat{A}^{\text{team}}_{t}\right).$$

The corresponding critic return targets are

$$\widehat{R}^{\text{ind}}_{i,t} = \widehat{A}^{\text{ind}}_{i,t} + V^{\text{ind,old}}_{i,t}, \qquad \widehat{R}^{\text{team}}_{t} = \widehat{A}^{\text{team}}_{t} + V^{\text{team,old}}_{t}.$$

For either critic, the clipped prediction is

$$V^{\text{clip}}_t = V^{\text{old}}_t + \operatorname{clip}\left(V_t - V^{\text{old}}_t, -\epsilon, \epsilon\right),$$

and the clipped value loss is

$$\mathcal{L}^{\text{clip}}_V = \mathbb{E}_t\left[\max\left((V_t - \widehat{R}_t)^2, (V^{\text{clip}}_t - \widehat{R}_t)^2\right)\right].$$

The total critic objective is

$$\mathcal{L}_{\text{critic}} = w_{\text{ind}}\mathcal{L}^{\text{clip}}_V\left(V^{\text{ind}}, \widehat{R}^{\text{ind}}\right) + w_{\text{team}}\mathcal{L}^{\text{clip}}_V\left(V^{\text{team}}, \widehat{R}^{\text{team}}\right).$$

For pursuing UAV i, define the probability ratio

$$\varrho_{i,t} = \exp\left(\log \pi_{\theta_i}(\mathbf{a}^t_i \mid \widetilde{\mathbf{o}}^t_i) - \log \pi_{\theta^{\text{old}}_i}(\mathbf{a}^t_i \mid \widetilde{\mathbf{o}}^t_i)\right).$$

The actor objective is

$$\mathcal{L}_{\text{actor},i} = -\mathbb{E}_t\left[\min\left(\varrho_{i,t}\widehat{A}_{i,t}, \operatorname{clip}(\varrho_{i,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_{i,t}\right) + \beta_{\text{ent}}\mathcal{H}\left(\pi_{\theta_i}(\cdot \mid \widetilde{\mathbf{o}}^t_i)\right)\right].$$

A.3 Detailed Reward Components

We use $[x]_+ = \max(x, 0)$. Let $\mathcal{V}^t_i$ be the set of grid cells visited by pursuing UAV i up to time t, and let

$$\mathcal{C}^t = \bigcup_{i=1}^{N} \mathcal{V}^t_i$$

be the cumulative team coverage. The coverage increase is

$$\Delta A^t = |\mathcal{C}^t| - |\mathcal{C}^{t-1}|.$$

If g^t_i is the grid cell occupied by UAV i, define

$$b^t_{i,\text{new}} = \mathbb{I}[g^t_i \notin \mathcal{V}^{t-1}_i].$$

The team coverage increase is attributed to the pursuing UAVs that enter previously unvisited cells:

$$\Delta A^t_i = \frac{b^t_{i,\text{new}}}{\sum_{j=1}^{N} b^t_{j,\text{new}} + \epsilon}\Delta A^t.$$

The overlap count is

$$m^t_{i,\text{overlap}} = \left[\sum_{j=1}^{N} \mathbb{I}[g^t_j = g^t_i] - 1\right]_+.$$

The mean pairwise separation is

$$\bar{d}^t_{\text{pp}} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \left\|\mathbf{p}^t_i - \mathbf{p}^t_j\right\|_2.$$

The resulting search reward is

$$r^t_{i,\text{search}} = \lambda_{\text{cov}}\Delta A^t_i + \lambda_{\text{new}} b^t_{i,\text{new}} - \lambda_{\text{overlap}} m^t_{i,\text{overlap}} + \lambda_{\text{sep}} \frac{\min(\bar{d}^t_{\text{pp}}, D_{\text{sep}})}{D_{\text{sep}}} + \lambda_{\text{mov}} \|\mathbf{v}^t_i\|_2.$$

During pursuit, let $\mathbf{p}_{i,\mathrm{goal}}^t$ denote either the current evader position or the geometry-based interception point assigned to UAV i. Define

$$d_{i,\mathrm{goal}}^t = \left\|\mathbf{p}_i^t - \mathbf{p}_{i,\mathrm{goal}}^t\right\|_2, \qquad r_{i,d}^t = d_{i,\mathrm{goal}}^{t-1} - d_{i,\mathrm{goal}}^t.$$

Let

$$\widehat{\mathbf{d}}_{i,\mathrm{goal}}^t = \frac{\mathbf{p}_{i,\mathrm{goal}}^t - \mathbf{p}_i^t}{d_{i,\mathrm{goal}}^t + \epsilon}, \qquad \widehat{\mathbf{v}}_i^t = \frac{\mathbf{v}_i^t}{\|\mathbf{v}_i^t\|_2 + \epsilon}.$$

The heading-alignment term is

$$r_{i,h}^t = (\widehat{\mathbf{v}}_i^t)^\top \widehat{\mathbf{d}}_{i,\mathrm{goal}}^t.$$

For angular coordination, only pursuing UAVs within the effective encirclement radius are included. If their ordered bearing angles are

$$\vartheta_{(1)}^t \le \cdots \le \vartheta_{(K_t)}^t,$$

the circular gaps are

$$\Delta\vartheta_k^t = \begin{cases} \vartheta_{(k+1)}^t - \vartheta_{(k)}^t, & k < K_t, \\ 2\pi + \vartheta_{(1)}^t - \vartheta_{(K_t)}^t, & k = K_t. \end{cases}$$

Angular uniformity is

$$u_\theta^t = \begin{cases} 1 - \min\left(1, \dfrac{1}{2\pi}\displaystyle\sum_{k=1}^{K_t}\left|\Delta\vartheta_k^t - \frac{2\pi}{K_t}\right|\right), & K_t \ge 2, \\ 0, & K_t < 2. \end{cases}$$

The stabilization term penalizes lagging, retreat near the target, and repeated large heading changes:

$$r_{i,\mathrm{sta}}^t = -\lambda_{\mathrm{lag}}\left[d_i^t - d_{\min}^t - D_{\mathrm{lag}}\right]_+ - \lambda_{\mathrm{ret}}\mathbb{I}[d_i^t < D_{\mathrm{ret}}]\left[d_i^t - d_i^{t-1}\right]_+ - \lambda_{\mathrm{osc}} b_{i,\mathrm{osc}}^t,$$

where

$$d_{\min}^t = \min_j d_j^t$$

and $b_{i,\mathrm{osc}}^t$ indicates repeated large heading reversals within a fixed temporal window. Combining these terms gives

$$r_{i,\mathrm{pursuit}}^t = \lambda_d r_{i,d}^t + \lambda_h r_{i,h}^t + \lambda_\theta u_\theta^t + \lambda_a\left[1 - \frac{\bar{d}^t}{D_{\mathrm{enc}}}\right]_+ + \lambda_c\mathbb{I}[\Delta\bar{d}^t > \delta_c]\Delta\bar{d}^t + r_{i,\mathrm{sta}}^t.$$

During reacquisition, define the unit direction to the last observed target position as

$$\widehat{\mathbf{d}}_{i,\mathrm{ls}}^t = \frac{\mathbf{p}_{\mathrm{ls}}^t - \mathbf{p}_i^t}{d_{i,\mathrm{ls}}^t + \epsilon}.$$

The distance-dependent directional weight is

$$w_i^t = \min\left(1, \frac{d_{i,\mathrm{ls}}^t}{D_{\mathrm{dir}}}\right).$$

The reacquisition term is

$$r_{i,\mathrm{reacq}}^t = \lambda_{\mathrm{pot}}\left(d_{i,\mathrm{ls}}^{t-1} - d_{i,\mathrm{ls}}^t\right) + \lambda_{\mathrm{dir}} w_i^t\left[(\widehat{\mathbf{v}}_i^t)^\top \widehat{\mathbf{d}}_{i,\mathrm{ls}}^t\right]_+ + \lambda_{\mathrm{near}}\left[1 - \frac{d_{i,\mathrm{ls}}^t}{D_{\mathrm{near}}}\right]_+ + r_{i,\mathrm{mode}}^t.$$

The mode-dependent term is

$$r_{i,\mathrm{mode}}^t = \begin{cases} R_{\mathrm{approach}}\left[1 - \dfrac{d_{i,\mathrm{ls}}^t}{D_{\mathrm{reacq}}}\right]_+, & \text{during directed approach}, \\ R_{\mathrm{patrol}}, & \text{during local patrol}. \end{cases}$$

Let $d_{i,\mathrm{obs}}^t$ and $d_{i,\mathrm{bnd}}^t$ be the distances from UAV i to the nearest obstacle and workspace boundary. A compact representation of the safety shaping is

$$r_{i,\mathrm{safety}}^t = -\lambda_{\mathrm{obs}}\left[1 - \frac{d_{i,\mathrm{obs}}^t}{D_{\mathrm{obs}}}\right]_+ - \lambda_{\mathrm{bnd}}\left[1 - \frac{d_{i,\mathrm{bnd}}^t}{D_{\mathrm{bnd}}}\right]_+ - \lambda_{\mathrm{turn}}\left|\mathrm{wrap}(\psi_i^t - \psi_i^{t-1})\right| - R_{\mathrm{col}} b_{i,\mathrm{col}}^t.$$

Rejected or redirected motion commands are recorded as safety interventions but are not treated as terminal collisions.

A.4 Network Architectures and Hyperparameters

Table 1 summarizes the network dimensions. The actor dimension includes the base observation, four-dimensional interception-oriented representation, and eight-dimensional prediction feature.

Table 1: Network architectures used by PA-MAPPO.

Component	Configuration
Trajectory history	$K = 10$ observations with (x, y, v_x, v_y) input features
Encoder projection	$4 \to 128$ with layer normalization, ReLU, and dropout
Temporal encoder	Two-layer bidirectional LSTM with hidden size 128 per direction
Attention context	256 dimensions
Autoregressive decoder	Horizon embedding size 32; LSTMCell hidden size 128; prediction horizon $H = 5$
Policy prediction feature	Horizons $\{1, 2, 3, 5\}$; range–bearing dimension 8
Task representation	Interception-oriented dimension 4
Independent actor	$42 \to 64 \to 64 \to 2$
Team critic	$126 \to 64 \to 64 \to 1$
Individual critic	$171 \to 64 \to 64 \to 1$

The individual-critic input consists of the 126-dimensional centralized observation, the 42-dimensional local augmented observation, and the three-dimensional one-hot identity vector. The team critic receives only the centralized observation.

Table 2: Optimization and training hyperparameters.

Hyperparameter	Value
Training budget per run	8×10^6 environment steps
Independent training seeds	3
Training episode horizon	300 steps
Evaluation episode horizon	500 steps
Actor and critic learning rate	1×10^{-3}
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clipping parameter ϵ	0.20
Optimization epochs per update	10
Episode batch size	32
Episode mini-batch size	8
Initial entropy coefficient	0.005
Entropy-coefficient range	$[0.001, 0.1]$
Individual/team advantage weights	$0.5/0.5$
Actor/critic hidden dimensions	$64/64$
Actor/critic gradient-norm limit	10
Predictor learning rate	1×10^{-3}
Predictor mini-batch size	32
Predictor dropout	0.20
Predictor gradient-norm limit	1
Predictor replay capacity	$10\,000$ history–future pairs
Evaluation cases per seed	300
Bootstrap resamples	4000

The principal reward settings are listed in Table 3. Positive coefficient magnitudes are reported; negative signs are introduced explicitly by the corresponding penalty terms.

Table 3: Principal reward parameters used in the reported experiments.

Reward parameter	Value
Step penalty r_{step}	-0.001
New-detection reward R_{vis}	0.8
Capture reward R_{cap}	45.0
Terminal collision penalty R_{col}	50.0
Search-grid cell width	50.0
Coverage coefficient λ_{cov}	0.05
New-cell coefficient λ_{new}	0.10
Overlap coefficient λ_{overlap}	0.02
Separation coefficient λ_{sep}	0.12
Separation saturation distance D_{sep}	120.0
Movement coefficient λ_{mov}	0.05
Angular-uniformity coefficient λ_{θ}	0.35
Effective angular-coverage radius	60.0
Encirclement approach coefficient λ_{a}	0.08
Contraction threshold δ_{c}	0.5
Last-position potential coefficient λ_{pot}	0.40
Last-position direction coefficient λ_{dir}	0.70
Near-loss-position coefficient λ_{near}	0.25
Near-loss-position radius D_{near}	30.0
Directed reacquisition radius D_{reacq}	70.0
Reacquisition base reward	0.60
Reacquisition approach coefficient	0.35
Obstacle-proximity threshold	60.0
Boundary-proximity threshold	50.0