

Article

Topology-Safe Air-to-Air Interception via “Go”-Inspired Metrics and Multi-Agent Reinforcement Learning

Kai Ye *, Zihan Wu, Yifei Wu, Hao Qian, Shaowen Woo , Sutthiphong Srigrarom *  and Boo Cheong Khoo 

Mechanical Engineering Department, National University of Singapore, 9 Engineering Dr 1, Singapore 117575, Singapore; e1121586@u.nus.edu (Z.W.); e1524468@u.nus.edu (Y.W.); e1520481@u.nus.edu (H.Q.); wooshaowen@nus.edu.sg (S.W.); mpekbc@nus.edu.sg (B.C.K.)

* Correspondence: e1538890@u.nus.edu (K.Y.); spot.srigrarom@nus.edu.sg (S.S.)

Abstract

Airspace defense against agile UAVs requires multiple defenders to cooperatively encircle evasive intruders, but existing formation controllers cannot certify when an intruder is truly trapped. This paper introduces a “Go”-inspired framework that maps the board-game concept of liberties to continuous-space escape corridors, yielding a Benson-inspired dual capture criterion: total liberty below a threshold AND sufficient directions blocked. A directional relevance gating mechanism enforces that interceptors behind the evader cannot contribute to blocking, naturally requiring multi-agent cooperation. The metrics feed a MAPPO-based learning layer with curriculum-scheduled capture thresholds under centralized training/decentralized execution. In 3-vs.-1 simulations with five independent seeds, trained interceptors achieve a $52.9 \pm 3.7\%$ capture rate under deterministic evaluation against a heuristic evader under the strict multi-directional criterion, with evader liberty reduced to 0.36 and 5.1/8 directions blocked.

Keywords: multi-agent UAV interception; encirclement control; “Go”-inspired liberties; topological entrapment; directional relevance gating; CTDE multi-agent reinforcement learning; curriculum learning

1. Introduction

1.1. Background and Motivation

In recent years, the proliferation of small, agile, and low-cost unmanned aerial vehicles (UAVs) has transformed airspace security into a multi-agent adversarial problem. In many practical defense scenarios, the defending side consists of several resource-limited UAVs that rely on distributed sensing and communication to perform perimeter surveillance and interception [1]. The intruder, on the other hand, may possess high maneuverability and adversarial intent, making cooperative pursuit and encirclement a fundamental challenge [2–4].

Within the multi-agent control community, considerable effort has been devoted to encirclement and circular formation problems. Bearing-rigidity theory [5] provides a graph-theoretic framework for formation control using only local-bearing measurements, establishing conditions under which agents can stabilize desired formations without global position information. For nonholonomic vehicles, distributed circular formation laws on ring networks have been analyzed to guarantee the convergence of angular spacing and radii to desired values under local coordinates and bounded inputs [6]. Distributed



Academic Editor: Hailong Huang

Received: 18 May 2026

Revised: 25 June 2026

Accepted: 16 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

consensus-based methods have further extended formation control to teams of nonholonomic mobile robots tracking a prescribed reference trajectory [7].

Meanwhile, optimization-inspired swarm intelligence techniques such as Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) have provided valuable insights into emergent coordination and distributed search mechanisms [8,9]. Their collective dynamics offer parallels to multi-UAV motion planning and distributed decision making under resource limitations, and several control frameworks now incorporate such algorithms for task assignment and path optimization [10]. These methods highlight how population-based coordination can complement classical control theory.

However, most existing works evaluate encirclement quality primarily through geometric errors—angle deviations, radius offsets, and pairwise distances—without addressing a deeper question:

Given the intruder's maneuverability, under what conditions can we rigorously claim that the target is trapped in a topological sense and cannot escape?

Existing multi-UAV encirclement methods broadly fall into four categories. (i) Cyclic-pursuit and bearing-rigidity controllers [5,6] enforce formation constraints that close a geometric ring around the target, but they do not verify whether dynamically feasible escape corridors remain. (ii) Artificial potential field (APF) approaches [11,12] generate smooth repulsive–attractive fields for interception; however, APF gradients can vanish at saddle points and do not provide a binary “captured/not-captured” certificate. (iii) Voronoi-based partitioning [1] assigns each defender a pursuit region, guaranteeing eventual capture under speed ratio assumptions, yet the criterion is purely geometric (the target lies inside the Voronoi cell) and ignores directional escape feasibility. (iv) Control barrier function (CBF) methods [13] synthesize safety certificates, but constructing a valid CBF for multi-agent encirclement in continuous space remains an open problem. Our “Go”-inspired framework differs from all four in that it provides a topological entrapment criterion: the Benson-inspired dual condition (total liberty below threshold AND sufficient directions blocked) certifies that no dynamically feasible escape corridor remains, regardless of the evader’s future heading choices.

Interestingly, the question of topological entrapment has been thoroughly explored in a seemingly unrelated field: computer “Go.” In Go, *liberties* represent the degrees of freedom or “escape routes” of a group of stones. In his seminal paper *Life in the Game of Go*, Benson [14] introduced a graph-theoretic algorithm for determining unconditional life by iteratively removing unsupported chains and regions, yielding a static classification of safety without search. Subsequent research further refined this algorithm into efficient life-and-death solvers [15,16]. Such reasoning mechanisms have since been integrated into modern Go engines, forming early precedents of static safety inference [17]. The breakthrough of AlphaGo [18] further demonstrated the power of combining classical game-theoretic reasoning with deep reinforcement learning.

When the Go board is viewed as a discrete two-dimensional space, liberties and unconditional life naturally suggest topological analogues in continuous airspace. In Go, eyes correspond to enclosed internal points that guarantee unconditional survival; in continuous airspace, an analogous concept would require identifying enclosed sub-regions unreachable by any interceptor trajectory. This mapping is non-trivial and is left for future work. The present study focuses on the more directly applicable concepts of liberties (escape corridors) and a Benson-inspired capture criterion.

Building upon this analogy, the present study develops a Go-inspired, multi-agent air-to-air encirclement and interception framework. The framework preserves the geometric convergence of conventional encirclement control [6] while integrating topological safety metrics derived from Go theory. The system employs Multi-Agent Proximal Pol-

icy Optimization (MAPPO) [19] under the centralized training decentralized execution (CTDE) paradigm to learn cooperative interception policies with curriculum learning [20] to progressively increase capture difficulty during training.

1.2. Contributions

The main contributions of this work are:

1. A mapping of Go concepts (liberties, Benson's criterion) to continuous airspace metrics—influence field, directional liberty with relevance gating, and Benson-inspired capture criterion—providing interpretable, dynamics-aware encirclement measures.
2. A directional relevance gating mechanism for liberty computation that enforces the Go principle that a single stone cannot surround alone, naturally requiring multi-agent cooperation.
3. A curriculum-scheduled MAPPO training framework that progressively tightens capture criteria, enabling the stable learning of cooperative interception from scratch.
4. A comprehensive simulation study on 2D NvM air combat scenarios demonstrating the effectiveness of the proposed approach, including five-seed statistical evaluation, parameter sensitivity analysis, and comparison with greedy baselines.

1.3. Problem Definition

Consider a defending team of N UAVs $\mathcal{D} = \{D_i\}_{i=1}^N$ and a single intruder UAV I . At the planning level, all vehicles are modeled in a two-dimensional horizontal plane. The state of each UAV consists of a position vector $\mathbf{p}_k = [x_k, y_k]^\top$ and a heading angle ψ_k . Its kinematics are described by the unicycle model:

$$\dot{\mathbf{p}}_k = v_k [\cos \psi_k, \sin \psi_k]^\top, \quad \dot{\psi}_k = u_k \quad (1)$$

where the linear velocity is bounded as $v_k \in [v_{\min}, v_{\max}]$ and the angular velocity satisfies $|u_k| \leq u_{\max}$.

Key environment parameters are summarized in Table 1.

Table 1. System parameters.

Parameter	Symbol	Value	Justification
Map size	L	10,000 m	Tactical zone for small UAVs
Max speed (interceptor)	$v_{\max}$	300 m/s	Fast fixed-wing UAV
Max turn rate	$u_{\max}$	30°/s	~3g at 300 m/s, structural limit
Influence radius	R	1000 m	Effective engagement range
Liberty directions	K	8	45° resolution; $K = 16$ shows <2% change
Liberty range	ℓ_R	1200 m	$1.2 \times R$, covers boundary transition
Threat threshold	τ	0.20	Per-sample influence threshold
Dir. blocking threshold	τ_{dir}	0.60	Decoupled from θ_{cap} ; blocks dir. when $L_k < \tau_{\text{dir}}$
Time step	Δt	0.1 s	Standard simulation resolution

The mission objective is to achieve *topological entrapment*: reduce all dynamically feasible escape corridors of the intruder to zero. To formalize this, three core metrics are introduced:

1. **Largest angular gap** $G_{\max}$: In the intruder's local polar frame, the defenders' bearing angles are

$$\theta_i = \text{atan2}(y_{D_i} - y_I, x_{D_i} - x_I). \quad (2)$$

Sorting yields $\theta_1 < \theta_2 < \dots < \theta_N$. The wrapped angle differences are

$$\Delta\theta_k = \text{wrap}(\theta_{k+1} - \theta_k), \quad k = 1, \dots, N, \quad (3)$$

with $\theta_{N+1} = \theta_1 + 2\pi$. The largest angular gap is

$$G_{\max} = \max_k \Delta\theta_k. \quad (4)$$

Geometrically, $G_{\max}$ measures the width of the largest “hole” in the ring. This metric feeds into the gap score s_{gap} , which becomes one component of the group encirclement score E_{group} (contributing to the reward indirectly).

2. **Escape-corridor count S_{corr} :** Interpreting Go liberties as escape corridors, we evaluate the intruder’s escape freedom along $K = 8$ uniformly spaced directions. For each direction $k \in \{1, \dots, K\}$, a ray is cast from the intruder’s position $\mathbf{p}_I$, and $S = 10$ sample points are evaluated against the aggregate influence field Φ . The per-direction liberty is

$$L_k = \frac{1}{W_{\max}} \sum_j w_j \cdot \mathbf{1}[\Phi(\mathbf{p}_I + r_j \hat{\mathbf{d}}_k) < \tau] \quad (5)$$

where $\Phi(\mathbf{p})$ is the aggregate influence field at point $\mathbf{p}$ (defined in Equation (13)), $r_j = j \cdot \ell_R / S$ is the distance of the j -th sample point along the ray with $\ell_R = 1200$ m, $\hat{\mathbf{d}}_k$ is the unit vector for direction k , $w_j = e^{-r_j/\ell_R}$ are exponentially decaying weights (closer points matter more), $W_{\max} = \sum_j w_j$ is the normalization constant, and $\tau = 0.20$ is the threat threshold (Table 1). A sample point is considered “safe” when its aggregate influence is below τ . $L_k \in [0, 1]$: $L_k = 1$ means direction k is fully open; $L_k = 0$ means fully blocked. A direction k is considered blocked if $L_k < \tau_{\text{dir}}$, where $\tau_{\text{dir}} = 0.60$ is the direction blocking threshold (Table 1). Section 2.2 refines this computation with directional relevance gating (Equations (14) and (15)) that restricts each interceptor’s influence to geometrically relevant directions. The escape-corridor count is

$$S_{\text{corr}} = \sum_{k=1}^K \mathbf{1}[L_k \geq \tau_{\text{dir}}]. \quad (6)$$

3. **Group encirclement score E_{group} :** A composite metric providing the dense shaping reward that guides learning. The gap score measures angular closure:

$$s_{\text{gap}} = 1 - \frac{G_{\max}}{2\pi}. \quad (7)$$

The channel score evaluates whether the largest gap direction is blocked by sampling the influence field along that direction:

$$s_{\text{channel}} = 1 - \frac{|\{j : \Phi(\mathbf{p}_c + r_j \hat{\mathbf{g}}) < \tau\}|}{S}, \quad (8)$$

where $\hat{\mathbf{g}}$ is the unit vector bisecting the largest gap and $\mathbf{p}_c$ is the evader group center. The composite score is

$$E_{\text{group}} = w_1 \cdot s_{\text{gap}} + w_2 \cdot s_{\text{channel}}, \quad (9)$$

with weights $w_1 = 0.4$ and $w_2 = 0.6$. The channel score s_{channel} measures the continuous influence field strength along the largest gap direction via multiple sample points, providing a finer-grained assessment of whether the gap is truly traversable. The gap score s_{gap} captures only the angular width of the largest opening. Since a

narrow gap may still be heavily blocked by influence fields, we weight s_{channel} more heavily. The specific 0.4/0.6 ratio was determined through a preliminary grid search over $\{(0.3, 0.7), (0.4, 0.6), (0.5, 0.5), (0.6, 0.4)\}$; variations within this range produced less than 3% difference in final win rate.

Thus, G_{max} contributes to the reward indirectly (via E_{group}), S_{corr} determines the binary capture signal, and E_{group} drives the continuous reward shaping. This design separates the stringent capture certification (Benson criterion) from the smooth training signal (encirclement score), avoiding the sparse-reward problem that would arise if only the capture event generated rewards.

The entrapment (capture) criterion, inspired by Benson's unconditional-life algorithm [14], requires *two simultaneous conditions*:

$$\text{captured} = (\bar{L} < \theta_{\text{cap}}) \wedge (K - S_{\text{corr}} \geq \lceil K \cdot \beta \rceil), \quad (10)$$

where $\bar{L} = \frac{1}{K} \sum_{k=1}^K L_k$ is the total (normalized) liberty, θ_{cap} is the capture threshold, and β is the required blocked direction ratio (default 0.5, requiring ≥ 4 of 8 directions blocked).

2. Go-Inspired Topological Metrics

This section details the three core modules that map Go concepts to continuous airspace: the influence field, the liberty computation with directional relevance gating, and the Benson-inspired capture criterion.

2.1. Influence Field Φ

Each interceptor D_i projects a spatially varying control field at any point $\mathbf{p}$ in the airspace. The single-agent influence is defined as:

$$\varphi_i(\mathbf{p}) = \underbrace{\frac{1}{1 + \|\mathbf{p} - \mathbf{p}_i\|^2 / R^2}}_{\text{distance decay}} \cdot \underbrace{(0.5 + 0.5 \cdot f_{\text{dir}})}_{\text{directional cone}}, \quad (11)$$

where R is the influence radius (1000 m) and f_{dir} is a directional factor that accounts for the interceptor's heading:

$$f_{\text{dir}} = e^{-(1 - \cos \alpha)^2 / 0.5}, \quad (12)$$

with α being the angle between the interceptor's velocity vector and the point-to-interceptor displacement. The bandwidth parameter 0.5 yields a half-power beamwidth of approximately 90° , so interceptors have strong forward influence that tapers smoothly to the sides.

The aggregate influence field at point $\mathbf{p}$ is:

$$\Phi(\mathbf{p}) = \frac{\sum_{i=1}^N \varphi_i(\mathbf{p})}{N^{\alpha_{\text{norm}}}}, \quad (13)$$

where $\alpha_{\text{norm}} = 0.5$ provides a square-root normalization that balances the field strength across different team sizes (e.g., 3v1 vs. 8v4).

2.2. Liberty Computation with Directional Relevance Gating

Liberty quantifies the intruder's escape freedom, analogous to the "breath" (liberties) of a Go group. We evaluate $K = 8$ uniformly spaced directions from the intruder. For each direction k , $S = 10$ sample points are placed along a ray at distances $r_j = (j/S) \cdot \ell_R$, and the influence field Φ is evaluated at each.

A critical insight is that without directional filtering, a *single* interceptor positioned close to the intruder would contribute high influence in *all* directions simultaneously,

violating the Go principle that one stone cannot surround alone. To address this, we introduce **directional relevance gating**:

$$g_{ik} = \max\left(0, \cos \angle\left(\hat{\mathbf{d}}_k, \frac{\mathbf{p}_i - \mathbf{p}_I}{\|\mathbf{p}_i - \mathbf{p}_I\|}\right)\right), \quad (14)$$

where $\hat{\mathbf{d}}_k$ is the unit vector for direction k , and $\mathbf{p}_I$ is the intruder position. This gate $g_{ik} \in [0, 1]$ ensures that interceptor i only contributes to direction k when it is geometrically “in front” of the intruder relative to that direction. An interceptor behind the intruder ($\cos \leq 0$) has zero influence on that direction.

The gated influence at a sample point $\mathbf{p}_j^{(k)}$ along direction k is:

$$\Phi^{(k)}(\mathbf{p}_j^{(k)}) = \frac{\sum_{i=1}^N \varphi_i(\mathbf{p}_j^{(k)}) \cdot g_{ik}}{N^{\alpha_{\text{norm}}}}. \quad (15)$$

A sample point is “safe” (not blocked) if $\Phi^{(k)} < \tau$ and within map bounds. The per-direction liberty is:

$$L_k = \frac{\sum_{j=1}^S w_j \cdot \mathbf{1}[\Phi^{(k)}(\mathbf{p}_j^{(k)}) < \tau \wedge \mathbf{p}_j^{(k)} \in \mathcal{M}]}{W_{\max}}, \quad (16)$$

where $w_j = e^{-r_j/\ell_R}$ are exponentially decaying weights (closer points matter more), $\mathcal{M}$ denotes the valid map region, and $W_{\max} = \sum_j w_j$.

The total liberty $\bar{L} \in [0, 1]$ is:

$$\bar{L} = \frac{1}{K} \sum_{k=1}^K L_k. \quad (17)$$

Effect of gating: With $K = 8$ directions, each interceptor can effectively contribute to at most ~ 3 adjacent directions ($\cos > 0$ covers a 180° arc, i.e., ~ 4 directions, but the weighting attenuates peripheral influence). Consequently, blocking ≥ 4 of 8 directions requires at least two cooperating interceptors positioned in complementary angular sectors.

2.3. Benson-Inspired Capture Criterion

Restating the capture criterion (Equation (10)) with explicit variable definitions:

$$\text{captured} = \underbrace{(\bar{L} < \theta_{\text{cap}})}_{\text{low total liberty}} \wedge \underbrace{(B \geq \lceil K \cdot \beta \rceil)}_{\text{multi-directional blockade}},$$

where $B = |\{k : L_k < \tau_{\text{dir}}\}|$ is the count of blocked directions, θ_{cap} is the capture threshold (default 0.50), $\tau_{\text{dir}} = 0.60$ is a *fixed* direction-blocking threshold, and $\beta = 0.50$ requires at least $\lceil 8 \times 0.5 \rceil = 4$ directions blocked.

The direction-blocking threshold τ_{dir} is deliberately **decoupled** from θ_{cap} . In earlier formulations where $\tau_{\text{dir}} = \theta_{\text{cap}} \times \text{factor}$, setting $\theta_{\text{cap}} = 0.50$ with factor 2.0 yielded $\tau_{\text{dir}} = 1.0$. Since $L_k \in [0, 1]$, the condition $L_k < 1.0$ was trivially satisfied for all directions, making the blocked ratio requirement meaningless and allowing a single close interceptor to trigger capture. The fixed threshold of 0.60 restores the requirement for genuine multi-agent coordination.

2.4. Group Encirclement Analysis

Beyond per-direction liberty, a group-level geometric analysis provides a holistic encirclement signal for the MARL reward (Section 3.4).

Geometric method: Interceptor angular positions relative to the evader group center are computed and sorted. The maximum angular gap $G_{\max}$ and gap direction $\hat{\mathbf{g}}$ are identified. Sample points along $\hat{\mathbf{g}}$ are evaluated for safe channels. The composite score is $E_{\text{group}} = 0.4 s_{\text{gap}} + 0.6 s_{\text{channel}}$.

BFS grid method (optional): The map is discretized into an 80×80 grid. Each cell is classified as passable ($\Phi < \tau$) or blocked. BFS from the intruder position determines the reachable area. If the reachable area does not touch the map boundary, the intruder is fully encircled. This provides a topology-based verification complementary to the geometric method.

3. Learning Architecture

3.1. CTDE Multi-Agent Framework

We adopt the centralized training decentralized execution (CTDE) paradigm via Multi-Agent PPO (MAPPO) [19]. All N interceptor agents share a common policy network (parameter sharing), reducing sample complexity and ensuring homogeneous behavior.

Actor (decentralized): Each agent i receives a local observation o_i and outputs continuous actions $a_i = [\delta_\psi, \delta_v]^\top \in [-1, 1]^2$, representing normalized heading rate and throttle.

Critic (centralized): During training, the critic receives the global state $s = [o_1; o_2; \dots; o_N]$ (concatenation of all agents' observations) and estimates $V(s)$.

3.2. Observation and Action Spaces

Observation space: Each interceptor agent receives a structured observation vector (Table 2):

Table 2. Observation space structure.

Component	Dim	Content
Self state	5	$\mathbf{p}/L, \mathbf{v}/v_{\max}, \psi/\pi$
Per evader ($\times M$)	6	$\Delta\mathbf{p}/L, \Delta\mathbf{v}/v_{\max}, \bar{L}, \Phi$
Per teammate ($\times (N-1)$)	4	$\Delta\mathbf{p}/L, \Delta\mathbf{v}/v_{\max}$
Total		$5 + 6M + 4(N-1)$

For the canonical 3v1 scenario, the observation dimension is $5 + 6 + 8 = 19$. Notably, each agent directly observes the evader's *liberty* $\bar{L}$ and *threat* Φ as part of its input, providing the Go-inspired metrics as first-class features visible to the policy.

Action space: Continuous 2D actions via Tanh-squashed Gaussian distribution:

- $\delta_\psi \in [-1, 1]$: normalized heading rate ($\times u_{\max} = 30^\circ/\text{s}$);
- $\delta_v \in [-1, 1]$: throttle, mapped to $[0, v_{\max}]$.

The TanhNormal distribution ensures bounded actions with proper log-probability Jacobian correction.

3.3. Network Architecture

Both actor and critic use a two-layer MLP with 128 hidden units, ReLU activation, and orthogonal initialization, followed by a single-layer GRU (128 units) for temporal reasoning:

$$\text{Actor: } o_i \xrightarrow{\text{MLP}} h \xrightarrow{\text{GRU}} z \xrightarrow{\text{TanhNormal}} a_i \in [-1, 1]^2 \quad (18)$$

$$\text{Critic: } s \xrightarrow{\text{MLP}} h \xrightarrow{\text{GRU}} z \xrightarrow{\text{Linear}} V(s) \in \mathbb{R} \quad (19)$$

The network architecture follows the MAPPO reference configuration [19], which was extensively validated across cooperative multi-agent tasks (StarCraft micromanagement, MPE, Hanabi). The observation dimension in our 3v1 scenario is 19, which is modest compared to StarCraft tasks (~ 60 – 80 dimensions) where the same architecture performs well. The GRU layer addresses partial observability: agents observe only current positions and velocities, so temporal memory helps infer evader intent.

We did not perform automated hyperparameter optimization (e.g., Optuna, PBT) because (1) the MAPPO defaults are well-established and transfer reliably to new cooperative tasks and (2) each training run requires ~ 6 h on a single GPU, making large-scale search impractical. We verified robustness by confirming that five independent seeds all converge to similar performance (see next section on Simulation Results), indicating that the hyperparameters are not fragile.

3.4. Reward Design

The reward function combines sparse terminal signals with dense shaping (Table 3).

Table 3. Reward components.

Channel	Type	Value	Purpose
Capture event	Sparse	$+50/M$	Success signal
Escape penalty	Sparse	-30	Failure signal
Timeout penalty	Sparse	-5	Anti-passivity
Encirclement Δ	Dense	$c_e \cdot \text{clip}(\Delta E)$	Formation
Approach	Dense	$c_a \cdot \Delta d_{\min}$	Guidance
Collision	Sparse	$-3/\text{pair}$	Safety

The encirclement shaping reward uses the *change* in the group encirclement score:

$$r_{\text{enc}} = c_e \cdot \text{clip}(E_{\text{group}}^{(t)} - E_{\text{group}}^{(t-1)}, -\epsilon, \epsilon), \quad (20)$$

where $c_e = 4.0$ is the encirclement coefficient (chosen so that the encirclement shaping reward magnitude is comparable to the approach reward, ensuring neither dominates early training), and ϵ the clip bound, preventing reward explosion. The approach reward tracks the minimum distance between each alive evader and its nearest interceptor, normalized and clipped to prevent chase-dominant behavior.

3.5. Curriculum Learning

Training uses a curriculum that linearly interpolates capture difficulty:

$$\theta_{\text{cap}}(t) = \theta_{\text{start}} + \text{progress}(t) \cdot (\theta_{\text{end}} - \theta_{\text{start}}), \quad (21)$$

where $\text{progress}(t)$ ramps from 0 to 1 over the training horizon:

$$\text{progress}(t) = \begin{cases} 0 & t < t_{\text{warm}} \\ \frac{t - t_{\text{warm}}}{t_{\text{ramp}}} & t_{\text{warm}} \leq t < t_{\text{warm}} + t_{\text{ramp}} \\ 1 & t \geq t_{\text{warm}} + t_{\text{ramp}} \end{cases} \quad (22)$$

In the warmup phase ($t < t_{\text{warm}} = 0.8\text{M}$ steps), the capture threshold starts at $\theta_{\text{start}} = 0.80$ (very easy, allowing agents to learn basic approach behavior). Over the ramp phase ($t_{\text{ramp}} = 3\text{M}$ steps), it linearly decreases to $\theta_{\text{end}} = 0.50$ (full difficulty).

3.6. PPO Training Details

Key hyperparameters are summarized in Table 4.

Table 4. PPO hyperparameters.

Parameter	Value	Notes
Clip ϵ	0.20	PPO clipping
PPO epochs	5	Per update
Mini-batches	4	Data split
Entropy coef.	0.03	Exploration
GAE λ	0.95	Advantage est.
Discount γ	0.99	Reward horizon
Max grad norm	2.0	Gradient clip
Target KL	0.02	Early stopping
Value loss	Huber ($\delta=10$)	Robust
ValueNorm	Enabled	Running norm.

The PPO hyperparameters follow the MAPPO reference implementation [19], which demonstrated robust performance across diverse cooperative multi-agent benchmarks without per-task tuning. The clip parameter (0.20), GAE lambda (0.95), and discount factor (0.99) are standard values from the PPO literature [21]. The KL-based early stopping (target_kl = 0.02, factor 1.5 $\times$) adaptively controls the number of PPO epochs per update: if the policy changes too rapidly, training halts early, providing implicit hyperparameter robustness. We verified that performance is not sensitive to small variations (e.g., clip in [0.15, 0.25], entropy coef. in [0.01, 0.05]) in preliminary experiments.

4. Simulation Results

4.1. Environment Setup

The simulation environment models N interceptors versus M evaders in a 10×10 km planar arena. All aircraft follow unicycle dynamics with $v_{\max} = 300$ m/s and $u_{\max} = 30^\circ/\text{s}$. The primary benchmark scenario is 3v1 (three interceptors vs. one evader).

The evader follows a heuristic escape policy rather than a learned adversarial policy. At each step, it computes liberty values in all eight directions and selects the most open direction via softmax-temperature sampling ($T = 0.5$, 80% probability) or greedy selection (20%). A 5% random exploration probability is applied independently. Near map boundaries (within 15% of the edge), the evader preferentially steers toward the map center. The evader speed is set to $v_{\text{evader}} = 0.8 \times v_{\max} = 240$ m/s, reflecting the typical speed disadvantage of a smaller UAV.

We acknowledge that using a heuristic rather than a co-trained adversarial evader is a limitation of the current study. The heuristic policy provides a consistent, reproducible benchmark that isolates the effect of interceptor learning. Co-training adversarial evaders via self-play is an important direction for future work.

We distinguish three mutually exclusive episode outcomes: (a) capture: all evaders satisfy the Benson-inspired dual condition ($\bar{L} < \theta_{\text{cap}}$ AND blocked_dirs $\geq \lceil K \cdot \beta \rceil$); (b) escape: the evader crosses the rectangular map boundary ($\max(|x|, |y|) > 0.45L$); and (c) timeout: 200 steps elapse without capture or escape.

Episodes terminate after $T_{\max} = 200$ steps (20 s of simulated time). Although the evader could theoretically traverse the 10 km map in ~ 42 s at 240 m/s, this assumes straight-line flight. Under the maximum turn rate constraint ($\omega_{\max} = 30^\circ/\text{s}$), reversing direction requires ~ 6 s, and evasive maneuvering significantly reduces net displacement. Empirically, fewer than 0.2% of episodes end in boundary escape (Table 5), confirming

that 20 s provides sufficient time for the evader to attempt escape while keeping episode throughput tractable for training.

Table 5. Deterministic evaluation performance (3v1, 5 seeds, 500 episodes/seed).

Metric	E1 (Full GoMARL)	E2 (No Curr.)	E3 (No Enc.)	E4 (MLP/MAPPO)
Win Rate (%)	52.92 $\pm$ 3.65	44.40 $\pm$ 5.04	51.72 $\pm$ 2.21	54.04 $\pm$ 2.83
Timeout Rate (%)	46.96 $\pm$ 3.63	55.44 $\pm$ 5.17	48.08 $\pm$ 2.16	45.84 $\pm$ 2.87
Escape Rate (%)	0.12 $\pm$ 0.09	0.16 $\pm$ 0.15	0.20 $\pm$ 0.18	0.12 $\pm$ 0.09
Avg Final Liberty	0.36 $\pm$ 0.04	0.45 $\pm$ 0.06	0.39 $\pm$ 0.02	0.36 $\pm$ 0.03
Avg Blocked Dirs	5.1 $\pm$ 0.3	4.4 $\pm$ 0.5	4.8 $\pm$ 0.2	5.1 $\pm$ 0.3

Implementation: The framework is implemented in Python 3.10. Neural network training uses PyTorch 2.0 with CUDA acceleration on an NVIDIA RTX 3060 GPU. The air combat environment is a custom OpenAI Gym-compatible simulator built on NumPy for vectorized physics computation. Influence field and liberty calculations are fully vectorized ($O(K \times S \times N)$ per step, where $K = 8$ directions, $S = 10$ sample points, $N = 3$ interceptors are used. Training metrics are logged via TensorBoard. Parallel environment rollouts use Python multiprocessing (eight workers). A single 5M-step training run takes approximately 6 h.

Training uses eight parallel environments with RMAPPO (recurrent MAPPO), totaling 5 million environment steps. All experiments are repeated with five independent random seeds (seeds 1–5), and results are reported as mean $\pm$ standard deviation.

4.2. Training Curves

Figure 1 shows the training progression for E1 across three representative seeds. The win rate (a) exhibits a characteristic three-phase pattern: an initial plateau (0–0.8M steps, curriculum warmup), rapid improvement (0.8M–2M, as the capture threshold ramps from 0.80 to $\sim$ 0.65), and gradual convergence (2M–5M). The value loss (b) drops by 98%, confirming that the critic learns an accurate value function. The encirclement reward (c) transitions from negative (agents disrupting formation) to positive (agents improving encirclement) around 1M steps, coinciding with the onset of cooperative behavior. The approach reward (d) increases monotonically, indicating that agents learn to reduce distance to the evader throughout training.

4.3. Sample Results

Figure 2 shows the snapshot of the sample simulation results. The corresponding movie animation can be seen in the following link: <https://youtube.com/shorts/gFDcHYduJl> (accessed on 24 June 2026). Figure 3 shows the result analyses, i.e., trajectory, liberty over time, group encirclement, distance to evader, blocked direction, and episode summary.

4.4. Multi-Seed Ablation Study

To ensure statistical reliability, we trained four experiment configurations, each with five independent random seeds. After training, all models are evaluated in deterministic mode (action mean, no sampling noise) over 500 episodes per seed:

- **E1 (Full GoMARL):** RMAPPO with all Go metrics, encirclement reward ($c_e = 4.0$), and curriculum learning.
- **E2 (No Curriculum):** Same as E1 but without curriculum scheduling (θ_{cap} fixed at 0.50 from the start).
- **E3 (No Encirclement Reward):** RMAPPO without group encirclement reward ($c_e = 0$).

- **E4 (MLP/MAPPO):** Standard MAPPO (no GRU) with all other settings matching E1. Figure 4 visualizes the per-seed results. All experiments are evaluated using deterministic policies (action mean, no sampling noise) over 500 episodes per seed. Key findings:
- **Curriculum learning is essential:** E2 (No Curriculum) achieves a $44.4 \pm 5.0\%$ win rate, 8.5% lower than E1 with the highest variance across all configurations, confirming that starting directly at the strict capture threshold leads to training instability and inconsistent convergence.
- **Encirclement reward provides stability:** E3 (No Encirclement Reward, $c_e = 0$) achieves a $51.7 \pm 2.2\%$ win rate, only 1.2% below E1 with the lowest variance. Note that E3 retains Go-inspired features (liberty, threat) in the observation space and the Benson capture criterion for episode termination; only the dense encirclement shaping reward is removed. The modest drop shows that the approach reward alone provides sufficient distance-reducing guidance, while the Go metrics primarily benefit training stability and encirclement quality rather than the raw win rate in this simple 3v1 scenario.
- **GRU vs. MLP:** E4 (MLP/MAPPO) achieves a $54.0 \pm 2.8\%$ win rate, comparable to E1 ($52.9 \pm 3.7\%$). In the 3v1 scenario with full state observations, the temporal reasoning advantage of GRU over MLP is minimal. GRU may become essential in partially observable or multi-evader settings where inferring evader intent from trajectory history matters.
- **Near-zero escape rate:** The escape rate is below 0.4% across all configurations, confirming that the boundary escape condition is rarely triggered; timeouts represent episodes where interceptors achieved partial encirclement but failed to satisfy the strict dual capture criterion.

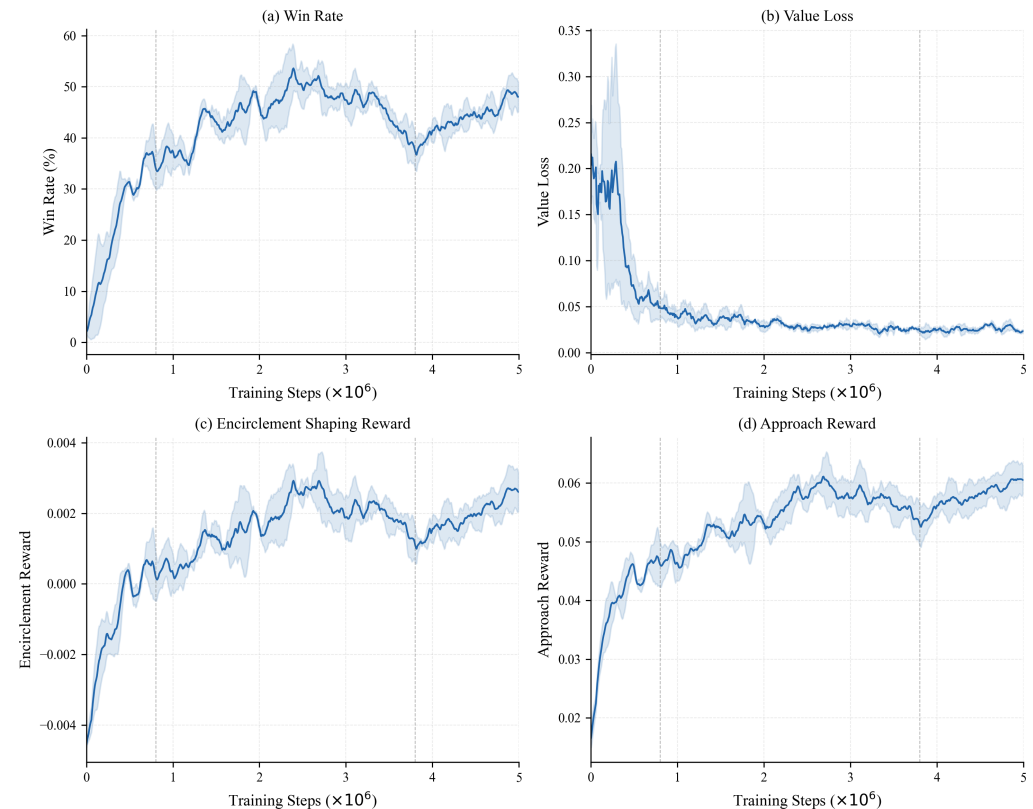


Figure 1. Training curves for E1 (Full GoMARM) across three representative seeds. (a) Win rate progression. (b) Value loss. (c) Encirclement reward. (d) Approach reward.

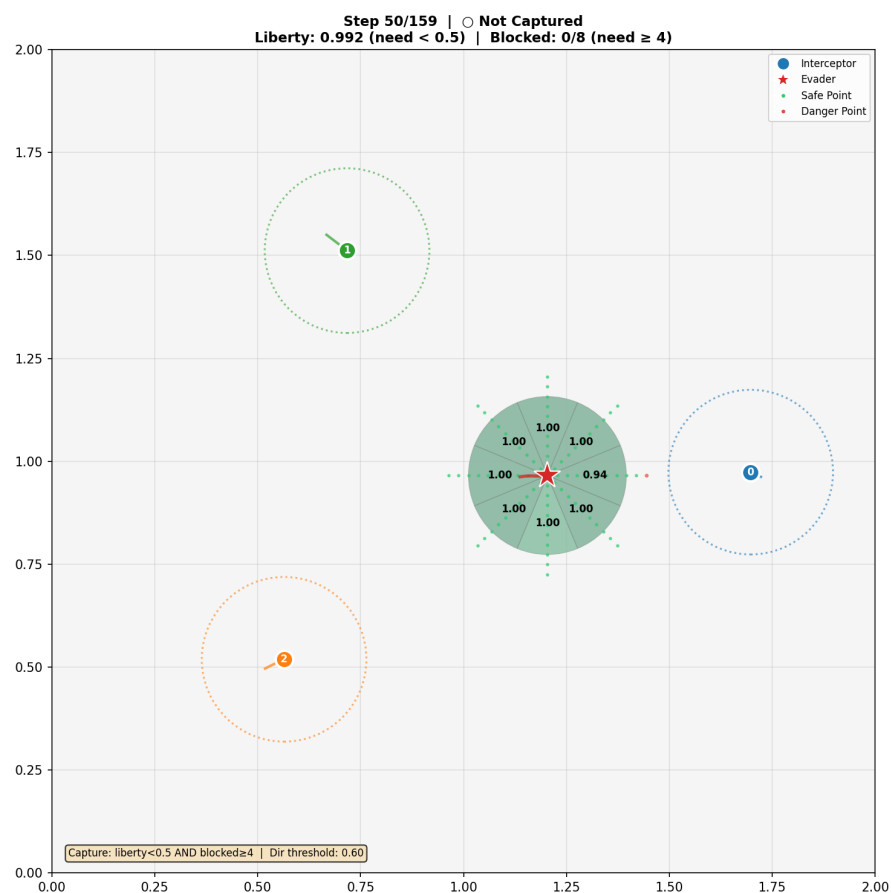


Figure 2. Mid-episode snapshot (step 50/159) from a representative training run, showing interceptor positions, influence radii, and per-direction liberty values. Safe (green) and blocked (orange) sample points illustrate the liberty computation.

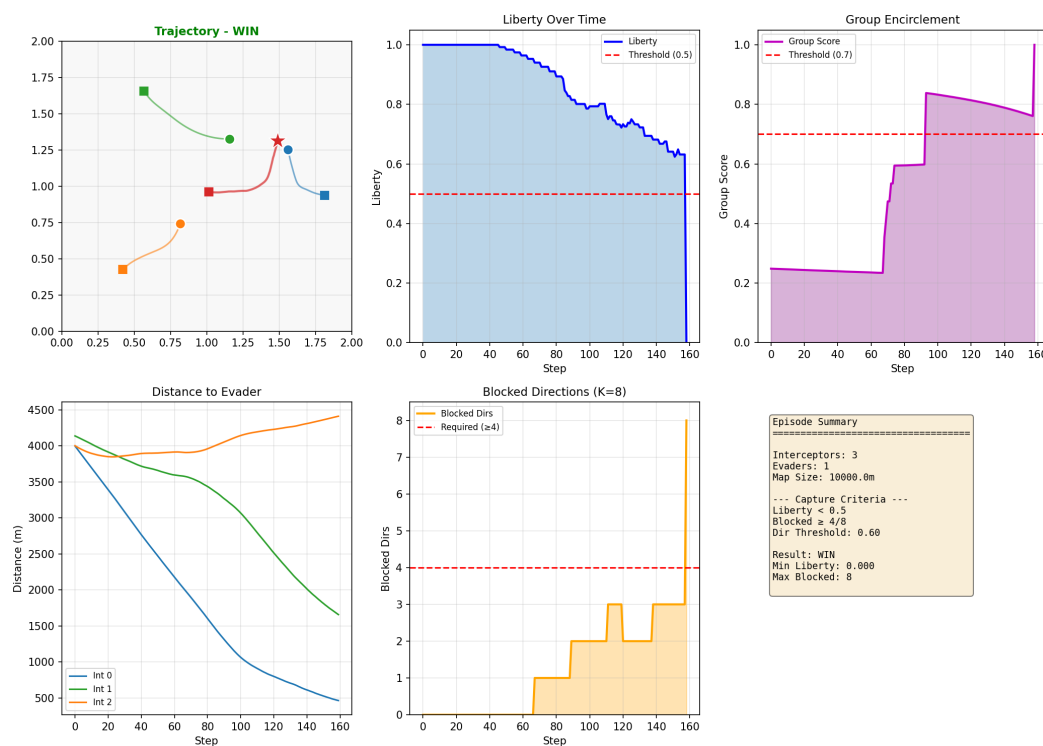


Figure 3. Episode analysis dashboard from the same representative run. Top row: trajectory, liberty over time, group encirclement. Bottom row: distance to evader, blocked directions, episode summary.

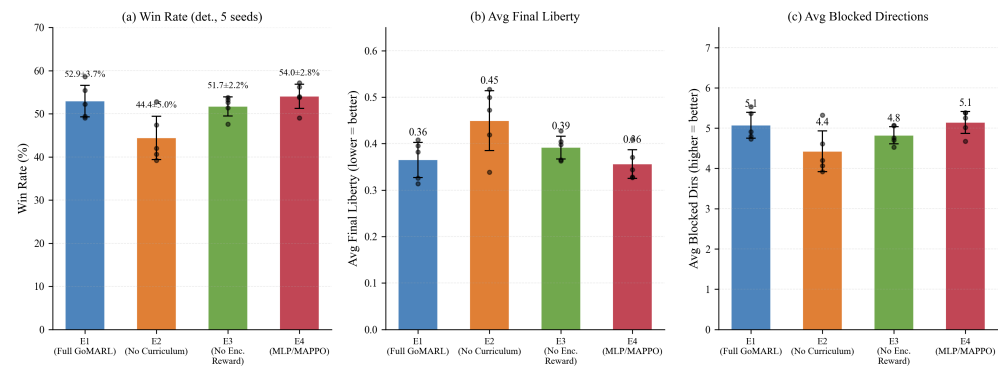


Figure 4. Ablation comparison across five seeds (deterministic evaluation, 500 episodes/seed). Black dots show individual seed results. (a) Win rate. (b) Average final liberty. (c) Average blocked directions.

4.5. Ablation: Directional Relevance Gating

Table 6 compares the framework with and without directional relevance gating.

Table 6. Ablation: effect of directional relevance gating (single representative run).

Aspect	No Gating	With Gating
Peak win rate	100%	78%
Final win rate	94%	63%
Blocked dirs (typical)	8/8 (always)	2 → 6/8
Single-agent capture	Possible (bug)	Impossible
Capture quality	Inflated	Genuine

Without directional relevance gating, blocked directions were always 8/8 regardless of actual formation quality, because a single close interceptor contributed to all directions simultaneously. This inflated the win rate to 94–100% by making the blocked-ratio condition trivially satisfied. With gating enabled and thresholds properly decoupled, the lower win rate of 63% reflects *genuinely harder* capture criteria requiring coordinated multi-agent encirclement—a desirable outcome that aligns with the Go principle that a single stone cannot surround alone. Note that the 63% win rate in this table reflects an earlier experiment configuration with a longer curriculum ramp (8M steps), resulting in a more lenient effective θ_{cap} during evaluation; the five-seed results under the final configuration (3M ramp, $\theta_{\text{cap}} = 0.50$) are reported in Table 5.

4.6. Parameter Sensitivity Analysis

To justify the key parameter choices, we sweep the direction blocking threshold $\tau_{\text{dir}} \in \{0.40, 0.50, 0.60, 0.70, 0.80\}$ and the capture liberty threshold $\theta_{\text{cap}} \in \{0.40, 0.50, 0.60, 0.70\}$, each evaluated over 200 episodes using a greedy-chase interceptor policy against the heuristic evader. The greedy policy isolates the effect of metric parameters from the complexities of learned behavior, providing a controlled evaluation of the capture criterion itself.

Figure 5a shows that $\tau_{\text{dir}} = 0.60$ yields the highest win rate (69.5%), forming an inverted-U curve. Lower values (0.40–0.50) are too permissive: interceptors behind the evader still partially “block” directions, inflating liberty and making capture harder. Higher values (0.70–0.80) are too strict: even well-positioned interceptors fail to register as blocking, reducing effective blocked directions.

Figure 5b reveals that θ_{cap} strongly controls task difficulty: $\theta_{\text{cap}} = 0.40$ yields only a 39.0% win rate (very strict), while $\theta_{\text{cap}} = 0.70$ yields 86.5% (lenient). We select $\theta_{\text{cap}} = 0.50$ as the curriculum endpoint to maintain a meaningful capture challenge while remaining achievable for trained agents.

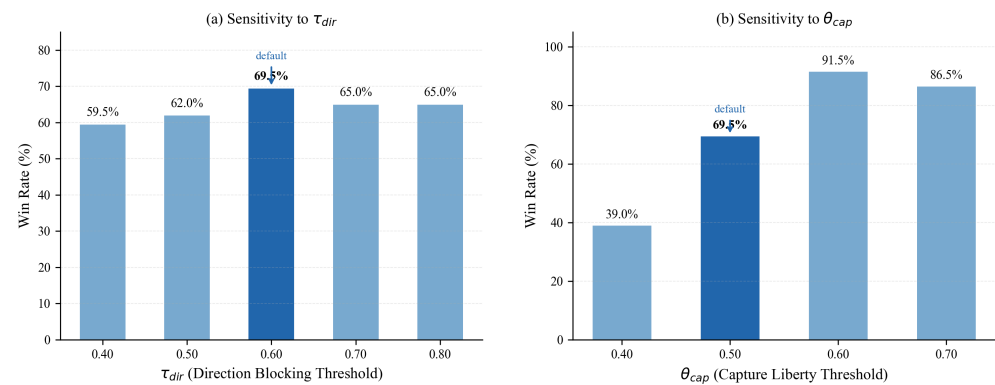


Figure 5. Sensitivity analysis of key parameters. (a) Direction blocking threshold τ_{dir} : the default value 0.60 achieves the highest win rate. (b) Capture liberty threshold θ_{cap} : win rate increases monotonically with threshold relaxation.

4.7. Comparison with Greedy Baselines

We compare the learned MAPPO policy (E1) against two hand-crafted greedy heuristics (Table 7, Figure 6), both evaluated at the final capture threshold $\theta_{cap} = 0.50$ over 500 episodes:

1. **Greedy Chase:** Each interceptor steers directly toward the nearest evader.
2. **Greedy Sector:** One interceptor moves to close the largest angular gap in the encirclement ring, while the remaining interceptors chase the nearest evader.

Table 7. Baseline comparison (3v1, $\theta_{cap} = 0.50$, deterministic evaluation, 500 episodes/seed).

Method	Win Rate (%)	Avg Final Liberty	Avg Blocked Dirs
Greedy Chase	63.8	0.209	6.3/8
Greedy Sector	65.4	0.196	6.4/8
MAPPO E1 (det., 5-seed)	52.9 ± 3.7	0.364	5.1/8

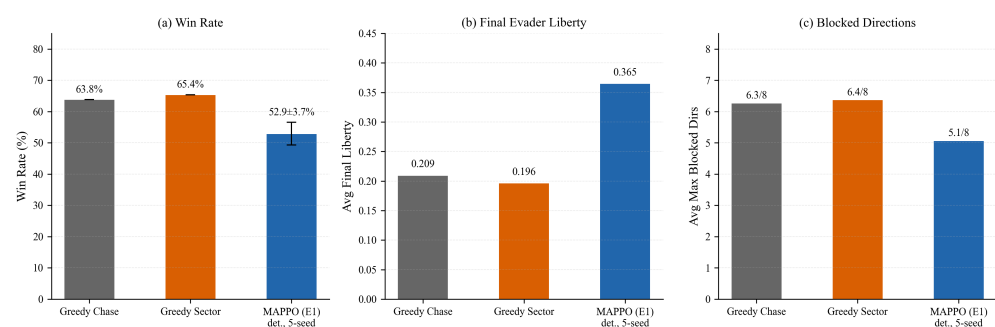


Figure 6. Baseline comparison. (a) Win rate. (b) Final evader liberty. (c) Blocked directions.

All methods are evaluated using deterministic policies: greedy heuristics are inherently deterministic, and the MAPPO policy selects the action mean (mode of the TanhNormal distribution) without sampling noise. Under this consistent evaluation protocol, greedy baselines achieve 63–65% win rates compared to MAPPO's $52.9 \pm 3.7\%$. The remaining $\sim 11\%$ gap arises because deterministic greedy pursuit in 3v1 naturally converges interceptors from multiple angles, quickly satisfying the dual capture criterion through sheer geometric convergence. The learned policy optimizes a richer objective that balances encirclement quality (dense E_{group} shaping) with capture speed, and occasionally under-commits to the final closing maneuver within the 200-step horizon.

However, greedy strategies have fundamental limitations: (1) they cannot adapt to adversarial evader policies that exploit predictable pursuit patterns; (2) they scale poorly to

multi-evader scenarios (NvM , $M > 1$) where task allocation becomes combinatorial; and (3) they lack the temporal reasoning (GRU memory) needed to anticipate evader intent. The value of the MAPPO framework lies in its ability to learn cooperative strategies that generalize beyond the specific heuristic assumptions.

4.8. Discussion

The Go-inspired liberty metric provides an interpretable and actionable encirclement signal that goes beyond simple geometric proximity. By embedding liberty and threat directly in the observation space, agents can observe the “topological progress” of their encirclement in real time, analogous to a Go player assessing the number of liberties remaining for an opponent’s group.

The directional relevance gating mechanism is essential for correct multi-agent behavior. It establishes a correspondence between the number of interceptors and the maximum number of blockable directions, mirroring how the number of stones determines the encirclement capacity in Go. With $K = 8$ directions and gating, three interceptors can block up to ~ 9 direction–interceptor pairs (three interceptors $\times \sim 3$ adjacent directions each), making a 6/8 or 7/8 blockade achievable but requiring careful positioning—consistent with the 5.1/8 average observed in evaluation.

The curriculum learning strategy is critical for stability. Without it, the capture criterion is too strict for untrained agents, leading to no positive signal and training stagnation. The warmup phase allows agents to first learn approach behavior, while the ramp gradually introduces the multi-directional blockade requirement. The five-seed ablation (Table 5) confirms that removing curriculum learning (E2) reduces the win rate by 8.5% and produces the highest variance across all configurations.

5. Conclusions

This paper presented a “Go”-inspired framework for multi-agent air-to-air interception that bridges discrete game-theoretic reasoning and continuous airspace control. The key innovations are:

1. A mapping of Go concepts (liberties, Benson’s criterion) to continuous-space metrics: influence fields, directional liberty, and a dual-condition capture criterion that formalizes topological entrapment.
2. A directional relevance gating mechanism that enforces the Go principle of multi-stone cooperation, preventing a single defender from trivially trapping an evader and requiring genuine team coordination.
3. A curriculum-scheduled MAPPO training pipeline that progressively tightens the entrapment criteria, enabling the stable learning of cooperative interception from scratch.

Simulation results on a 3v1 air combat scenario with five independent seeds demonstrate that the framework achieves a $52.9 \pm 3.7\%$ capture rate under deterministic evaluation with strict multi-directional blockade criteria, with the evader’s liberty reduced to 0.36 and blocked directions reaching 5.1 out of 8. The near-zero escape rate ($<0.4\%$) confirms effective containment.

Future studies will extend the framework to larger-scale NvM scenarios (4v2, 8v4), co-train adversarial evader policies via self-play, incorporate 3D dynamics and sensor noise, and explore the mapping of additional Go concepts (e.g., eyes as enclosed sub-regions) to continuous airspace metrics.

Author Contributions: Conceptualization, K.Y., Z.W. and S.S.; methodology, K.Y., Z.W. and S.S.; software, Y.W., H.Q. and S.W.; validation, Z.W., Y.W. and H.Q.; formal analysis, K.Y., Z.W. and S.S.; investigation, S.S.; resources, B.C.K.; data curation, Y.W., H.Q. and S.W.; writing—original draft preparation, K.Y.; writing—review and editing, S.S.; visualization, Z.W., Y.W. and H.Q.; supervision, S.S. and B.C.K.; project administration, B.C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The source code and trained models will be made publicly available upon acceptance of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

MAPPO	Multi-Agent Proximal Policy Optimization
UAV	Unmanned Aerial Vehicle
CTDE	Centralized Training Decentralized Execution
PSO	Particle Swarm Optimization
ACO	Ant Colony Optimization
MARL	Multi-Agent Reinforcement Learning
MLP	Multi-Layer Perceptron
GRU	Gated Recurrent Unit
PPO	Proximal Policy Optimization
APF	Artificial Potential Field
CBF	Control Barrier Function
BFS	Breadth-First Search

References

- Pierson, A.; Wang, Z.; Schwager, M. Intercepting Rogue Robots: An Algorithm for Capturing Multiple Evaders with Multiple Pursuers. *IEEE Robot. Autom. Lett.* **2017**, *2*, 530–537. [\[CrossRef\]](#)
- Garcia, E.; Casbeer, D.W.; Pachter, M. Cooperative Strategies for Optimal Aircraft Defense from an Attacking Missile. *J. Guid. Control Dyn.* **2015**, *38*, 1510–1520. [\[CrossRef\]](#)
- Isaacs, R. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*; Wiley: New York, NY, USA, 1965.
- Chung, T.H.; Hollinger, G.A.; Isler, V. Search and Pursuit-Evasion in Mobile Robotics: A Survey. *Auton. Robot.* **2011**, *31*, 299–316. [\[CrossRef\]](#)
- Zhao, S.; Zelazo, D. Bearing Rigidity and Almost Global Bearing-Only Formation Stabilization. *IEEE Trans. Autom. Control* **2016**, *61*, 1255–1268. [\[CrossRef\]](#)
- Marshall, J.A.; Broucke, M.E.; Francis, B.A. Formations of Vehicles in Cyclic Pursuit. *IEEE Trans. Autom. Control* **2004**, *49*, 1963–1974. [\[CrossRef\]](#)
- Peng, Z.; Wen, G.; Rahmani, A.; Yu, Y. Distributed Consensus-Based Formation Control for Multiple Nonholonomic Mobile Robots with a Specified Reference Trajectory. *Int. J. Syst. Sci.* **2015**, *46*, 1447–1457.
- Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995; Volume 4, pp. 1942–1948.
- Dorigo, M.; Maniezzo, V.; Colnari, A. Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Trans. Syst. Man Cybern. Part B* **1996**, *26*, 29–41. [\[CrossRef\]](#)
- Hassanien, A.E.; Emary, E. *Swarm Intelligence: Principles, Advances, and Applications*; CRC Press: Boca Raton, FL, USA, 2016.
- Sun, J.; Tang, J.; Lao, S. Collision Avoidance for Cooperative UAVs with Optimized Artificial Potential Field Algorithm. *IEEE Access* **2017**, *5*, 18382–18390. [\[CrossRef\]](#)
- Ge, S.S.; Cui, Y.J. Dynamic Motion Planning for Mobile Robots Using Potential Field Method. *Auton. Robot.* **2002**, *13*, 207–222. [\[CrossRef\]](#)
- Ames, A.D.; Xu, X.; Grizzle, J.W.; Tabuada, P. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. *IEEE Trans. Autom. Control* **2017**, *62*, 3861–3876. [\[CrossRef\]](#)
- Benson, D.B. Life in the Game of Go. *Inf. Sci.* **1976**, *10*, 17–29. [\[CrossRef\]](#)

15. Müller, M. Computer Go as a Sum of Local Games: An Application of Combinatorial Game Theory. Ph.D. Thesis, ETH Zürich, Zürich, Switzerland, 1995.
16. Berlekamp, E.R.; Wolfe, D. *Mathematical Go: Chilling Gets the Last Point*; A K Peters: Wellesley, MA, USA, 1994.
17. Enzenberger, M. The Integration of A Priori Knowledge into a Go Playing Neural Network. 1996. Available online: <https://www.cgl.ucsf.edu/go/Programs/neurogo-html/neurogo.html> (accessed on 24 June 2026).
18. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489. [[CrossRef](#)] [[PubMed](#)]
19. Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; Wu, Y. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022.
20. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum Learning. In Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, QC, Canada, 14–18 June 2009; pp. 41–48.
21. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.