

Article

Graph Attention Network-Enhanced Multi-Agent Proximal Policy Optimization for Cooperative Guidance in Attack–Defense Confrontation

Yuanli Cai *  and Junchao Zhao 

School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
zjames@stu.xjtu.edu.cn

* Correspondence: ylicai@mail.xjtu.edu.cn

Abstract

A graph attention network-enhanced multi-agent proximal policy optimization (GAT-MAPPO) framework is proposed for cooperative guidance in adversarial engagement scenarios. A dynamic heterogeneous interaction graph is formulated over interceptors and targets at every decision epoch. Through a multi-head graph attention encoder, relational features capturing both inter-interceptor cooperation and target threat dynamics are adaptively aggregated. These graph-enriched observations are processed by a Centralized-Training, Decentralized-Execution (CTDE) MAPPO architecture, guided by a hierarchical reward function that mandates miss distance minimization, simultaneity of arrival consensus, multi-directional encirclement, and smooth control effort. Furthermore, the integration of a three-stage curriculum learning strategy allows for robust cooperative policy derivation across transitions from rectilinear to highly adaptive evasion patterns, eliminating the need for explicit rule engineering. Extensive Monte Carlo simulations confirm GAT-MAPPO's superior performance: achieving > 95% interception success rate in 4-vs.-4 scenarios and reducing mean simultaneity error by 41.4% compared to the MAPPO baseline. Comprehensive ablation and sensitivity studies validate the critical roles played by graph attention encoding, reward hierarchy design, and progressive curriculum staging.

Keywords: cooperative guidance; multi-agent reinforcement learning; graph attention network; adversarial engagement; proximal policy optimization; formation control

1. Introduction

In aerospace guidance and autonomous multi-agent control, engagement paradigms are gradually shifting from single-interceptor operation to multi-interceptor cooperative guidance [1]. The advantage of cooperative engagement is that multiple interceptors can approach maneuvering targets from different directions while coordinating their arrival times, thereby increasing overall interception probability [2]. Realizing this advantage, however, requires addressing two intertwined technical challenges that have historically been studied in isolation.

The first challenge involves cooperative guidance. Each interceptor must continuously coordinate its time-to-go and approach angle with teammates to ensure simultaneous arrival at the target cluster while forming a multi-directional encirclement formation [3]. Early work augmented proportional navigation using impact-time error feedback [1], followed by studies integrating impact-angle constraints [3,4]. Zhao and Zhou [2] advanced



Academic Editor: Gokhan Inalhan

Received: 17 June 2026

Revised: 5 July 2026

Accepted: 8 July 2026

Published: 10 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.

this by formulating cooperative timing as a graph-theoretic consensus problem. More recent advancements include prescribed-performance guidance laws on both time and angle domains within Lyapunov frameworks [5], fully actuated spatio-temporal modeling [6], and integrated detection/guidance optimization routines in general engagements [7]. While these techniques provide rigorous stability proofs, they universally presuppose that the target's maneuver pattern is either known a priori or sufficiently bounded [8].

The second critical challenge stems from the adversarial nature of the target. Targets are not passive; they actively execute reactive evasive maneuvers and can adapt their strategies based on observed interceptor formations [8]. While theoretical works analyze error-dynamics optimality [8] or optimize midcourse trajectories via sequential convex programming [9], these models still rely on predictable target motion predictions. The fundamental tension, therefore, lies between the cooperative constraints, which demand global information exchange and joint optimization, and the realities of adversarial environments, where such global state information is incomplete and non-stationary.

Deep Reinforcement Learning (DRL) provides a powerful paradigm to address this challenge. In multi-agent contexts, the Centralized-Training, Decentralized-Execution (CTDE) framework is standard for cooperation. MADDPG [10] introduced CTDE for continuous action spaces using coupled centralized critics and decentralized actors. For scalability, QMIX [11] decomposed the global Q-function using monotonic value decomposition. MAPPO [12] achieved strong performance due to combining a central value function with the stability of PPO [13]. Subsequent works have pursued this direction: Cui et al. [14] adapted MAPPO for multi-interceptor guidance under uncontrollable velocities, and Chi et al. [15] utilized swarm intelligence principles for UAV adversarial confrontation. These efforts establish DRL as a viable methodology but still lack an explicit mechanism to optimally model the continuously evolving relational topology inherent in cooperative engagement scenarios.

Notwithstanding these advancements, applying MARL directly to multi-interceptor cooperative guidance presents a critical bottleneck: the effective modeling of dynamic interaction relationships between interceptors and targets. Conventional MARL frameworks typically process agent observations via simple vector concatenation into a global state. This flattened representation discards vital topological information, such as spatial neighbor proximity or relative threat ranking (“who is closest to whom” or “which target poses the greatest threat”). Graph Neural Networks (GNNs) provide an inherent structure for modeling such relational dependencies. GCNs [16] extended convolution to graphs, while GATs [17] enhance this by using learnable attention weights. This feature is crucial: the weight assigned to the closest target must substantially exceed that of distant targets—a differentiation impossible with uniform GCN aggregation. Li et al. [18] successfully integrated GNNs into MADDPG for UAV confrontation, demonstrating state-of-the-art performance (97.4% effectiveness). However, existing research largely restricts GNN deployment to discrete-action, high-level tactical processes (e.g., target assignment or waypoints), leaving the deep integration of GATs into continuous guidance command generation within a MAPPO framework—especially for multi-interceptor cooperative engagement scenarios—as an unexplored frontier.

Although these works have made important progress in their respective domains, deeply integrating graph attention mechanisms into the continuous guidance command generation pipeline within a MAPPO framework, specifically targeting multi-interceptor adversarial engagement scenarios, remains an unexplored direction. Existing combinations of GNN and MARL have been confined to discrete action spaces for high-level tactical decisions such as target assignment and waypoint planning, whereas multi-interceptor cooperative guidance requires generating continuous overload commands that must satisfy

strict flight dynamics constraints and cooperative performance requirements. Furthermore, handling the non-stationarity of the adversarial training environment to gradually build robust policies is a problem that has not been adequately addressed.

To address this research gap, GAT-MAPPO, a graph attention network-enhanced multi-agent proximal policy optimization framework for cooperative guidance (see Figure 1), is proposed in this paper. The main contributions of this work are summarized as follows.

First, a dynamic heterogeneous interaction graph construction method and a multi-head graph attention observation encoder for multi-interceptor adversarial engagement scenarios are proposed. The encoder unifies inter-interceptor coordination and interceptor-to-target relevance relationships into a single graph structure and adaptively extracts topology-aware features at each decision step, replacing the conventional state concatenation approach.

Second, a hierarchical cooperative reward function that decouples miss distance, simultaneous arrival error, interception angle diversity, and overload smoothness into independent reward components is designed. A dynamic weight annealing mechanism is employed to mitigate the inherent conflicts among these competing objectives during training.

Third, a three-stage curriculum adversarial training strategy that progressively transitions from stationary targets through known-pattern maneuvering targets to RL-driven adaptive evasion targets is proposed, ensuring stable training convergence while improving the learned policy's performance against simulated adaptive evasion within the numerical engagement environment.

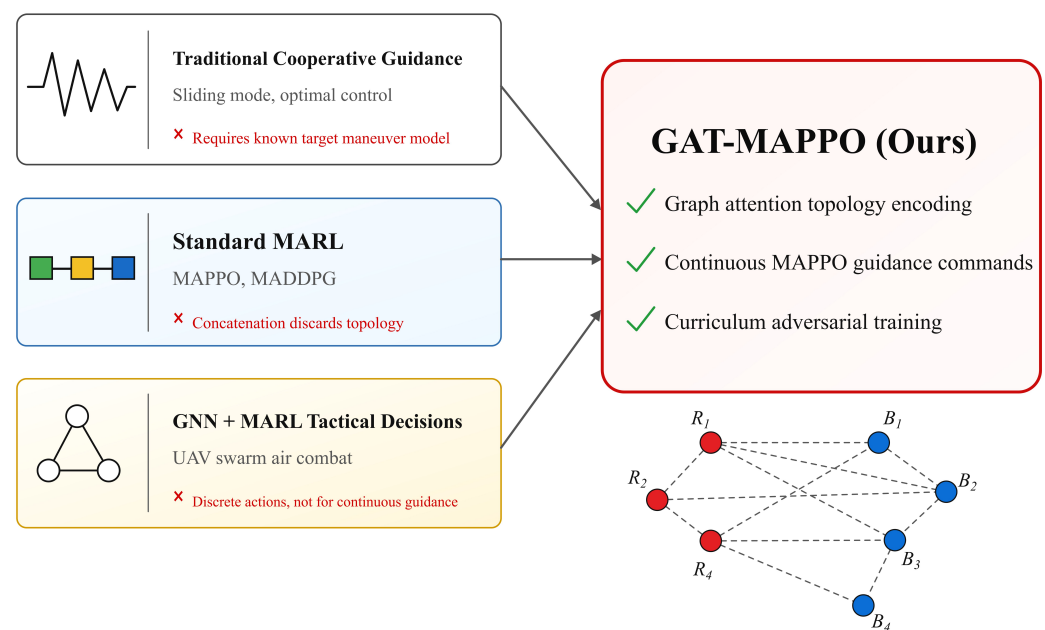


Figure 1. Comparison of existing paradigms and positioning of the proposed method. **Left** column: traditional cooperative guidance laws (model-dependent); standard MARL (topology-agnostic); and GNN + MARL for high-level decisions (discrete actions only). **Right:** GAT-MAPPO integrates graph attention topology encoding, continuous MAPPO guidance, and curriculum adversarial training. **Bottom:** a 4-vs-4 engagement scenario with dynamic graph edges.

2. Related Work

2.1. Cooperative Guidance Laws

The development of cooperative guidance laws spans several distinct phases. Initially, research focused on impact-time control methods, where proportional navigation was augmented with time-deviation feedback to synchronize multiple interceptors [1]. This concept

naturally extended to scenarios demanding joint control over both impact angle and impact time [3]. The methodology then shifted toward distributed cooperative models: Zhao and Zhou [2] pioneered formulating temporal synchronization as a consensus problem on a communication graph, requiring only local neighbor-to-neighbor information. Later works featured adaptive fast terminal sliding mode control to balance angle constraints with finite-time convergence [4]. Current state-of-the-art methods focus on enhancing performance guarantees and applicability: prescribed-performance controls constrain both transient and steady-state behavior through Lyapunov analysis [5]; spatio-temporally unified modeling treats the system as fully actuated [6]; joint optimization addresses detection and guidance concurrently [7], among others; and multi-angle search strategies optimize coverage for maneuvering targets [19]. Despite significant mathematical advancements, a persistent limitation across these models is their reliance on—or implicit assumption of—known or bounded target maneuver profiles.

2.2. Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) facilitates discovering optimal cooperative strategies by treating the environment interaction as a black box, thus bypassing the need for prior adversarial knowledge. Key paradigms include Centralized-Training, Decentralized-Execution (CTDE): MADDPG [10] pioneered this for continuous actions using centralized critics and decentralized actors. For resource efficiency, VDN [20] and QMIX [11] decompose the global Q-function additively or monotonically to ensure optimal scaling across large teams. MAPPO [12] has achieved state-of-the-art performance in cooperative benchmarks by stabilizing PPO with parameter sharing and a centralized value function. More recent critical applications include Cui et al. [14], who applied MAPPO to multi-interceptor guidance, thus validating the continuous control capability. This established DRL as a leading approach, but still, integrating relational graph structures is a missing piece for robust physical systems modeling in this domain.

2.3. Graph Neural Networks for Multi-Agent Interaction Modeling

Graph Neural Networks (GNNs) provide an essential framework for encoding structured interactions among multiple agents. Initial approaches include GCN [16], which extends spectral convolution to graph-structured data for tasks such as node classification. A significant advance is the Graph Attention Network (GAT) [17], replacing fixed kernels with learnable attention weights. This capability is paramount in aerospace applications: during terminal interception, the weight assigned to the nearest target must significantly exceed that of distant targets—a discrimination not possible with uniform GCN aggregation. Iqbal and Sha [21] demonstrated attention's advantage in multi-agent critics. More recently, Li et al. [18] utilized GNN to significantly enhance MADDPG critics for UAV confrontation. Critically, current work confines GNN deployment primarily to discrete-action, high-level tactical decision-making. Integrating GAT into the continuous guidance command generation ($\mathbf{a}_i$) within the MAPPO actor–critic architecture for multi-interceptor cooperative engagement—the core focus of this paper—remains a significant and unaddressed technical gap.

3. Problem Formulation

3.1. Multi-Body Adversarial Engagement Scenario

Consider a three-dimensional engagement scenario in which N interceptors cooperatively engage M maneuvering targets. For notational convenience, interceptor-related states are denoted by the superscript R and target-related states by the superscript B . The state of the i -th interceptor at time t is $\mathbf{s}_i^R(t) = [x_i, y_i, z_i, V_i, \theta_i, \psi_i]^\top$, where (x_i, y_i, z_i) denotes the

position in an inertial frame, V_i is the speed magnitude, and θ_i and ψ_i are the flight-path angle and heading angle. The 3D kinematics are governed by:

$$\dot{x}_i = V_i \cos \theta_i \cos \psi_i, \quad \dot{y}_i = V_i \sin \theta_i, \quad \dot{z}_i = -V_i \cos \theta_i \sin \psi_i, \quad (1)$$

and the guidance update equations are given by:

$$\dot{\theta}_i = \frac{a_{yi} - g \cos \theta_i}{V_i}, \quad \dot{\psi}_i = \frac{-a_{zi}}{V_i \cos \theta_i}, \quad (2)$$

where a_{yi} and a_{zi} are the longitudinal and lateral normal overloads, and g is gravitational acceleration. The control constraint is $\sqrt{a_{yi}^2 + a_{zi}^2} \leq a_{\max}$. Target kinematics follow an identical structure; however, target maneuver commands are generated independently by evasion strategies (rule-based or RL-driven), meaning the interceptor team only receives delayed, noisy estimates of the target state.

3.2. Observation and Action Space Design

Each interceptor i generates a local observation $\mathbf{o}_i$ comprising three distinct information layers. The self-state layer includes position, velocity, flight-path angles, and the residual available overload capacity $a_{\text{rem},i} = a_{\max} - \|\mathbf{a}_i(t)\|$. The target-perception layer provides relative position, relative velocity, and Line-of-Sight (LOS) angle rate for all detectable targets within range d_{detect} , normalized by maximum sensor capacity. Similarly, the teammate-communication layer includes relative position, relative velocity, and estimated time-to-go (t_{go}) for friendly interceptors within d_{comm} . Crucially, the dimensionality of the perception and communication layers is inherently dynamic: they vary as targets enter/leave detection cones or teammates move out of communication range. This variable-length input structure cannot be processed by standard fixed-dimension Multi-Layer Perceptrons (MLPs), which provides the core motivation for graph-structured observation encoding.

The action space $\mathcal{A}_i = [-a_{\max}, a_{\max}]^2$ is two-dimensional and continuous. At each step, the actor network outputs the mean (μ) and log-standard-deviation ($\log(\sigma)$) of a Gaussian distribution over $[a_{yi}, a_{zi}]$. The resulting action is passed through a tanh activation function and rescaled back to $[-a_{\max}, a_{\max}]$. This mechanism ensures that the physical overload constraint is satisfied directly at the network output, eliminating the need for additional penalties in the reward function.

3.3. Partially Observable Markov Game

The problem is modeled as a partially observable Markov game $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, P, \{r_i\}, \gamma \rangle$, where $\mathcal{N} = \{1, \dots, N\}$ is the set of Red interceptors and $\mathcal{S}$ is the global state space. $\mathcal{O}_i \subset \mathcal{S}$ represents the local observation of the i -th interceptor. The state transition function P is defined by kinematic Equations (1) and (2), r_i is the hierarchical reward detailed in Section 4.3, and $\gamma \in (0, 1)$ is the discount factor. The episode terminates upon target interception/escape or when any interceptor exceeds $T_{\max}$. Letting $\pi_i : \mathcal{O}_i \rightarrow \Delta(\mathcal{A}_i)$ denote the stochastic policy of the i -th interceptor, the objective function for policy optimization is defined as:

$$\max_{\{\pi_i\}} \mathbb{E} \left[\sum_{t=0}^T \gamma^t \sum_{i=1}^N r_i \left(s_t, \mathbf{a}_t^{(1)}, \dots, \mathbf{a}_t^{(N)} \right) \right]. \quad (3)$$

4. GAT-MAPPO Cooperative Guidance Framework

Figure 2 illustrates the overall architecture of the proposed GAT-MAPPO framework. In a multi-interceptor-versus-multi-target adversarial engagement scenario, the following pipeline is executed at each decision step: (1) a dynamic heterogeneous interaction graph $\mathcal{G}_t$ is constructed to capture both inter-interceptor cooperation and interceptor-to-target threat relationships in the current engagement geometry (Section 4.1); (2) a multi-head graph attention observation encoder compresses the graph-structured information into a fixed-length enriched observation vector for each interceptor (Section 4.2); (3) within a CTDE MAPPO architecture, each interceptor's decentralized actor network takes the enriched observation as input and independently outputs an overload command, while a shared centralized critic network estimates the joint state value from the concatenation of all enriched observations (Section 4.5). The training process is jointly guided by a hierarchical cooperative reward function (Section 4.3) and a three-stage curriculum adversarial training strategy (Section 4.4).

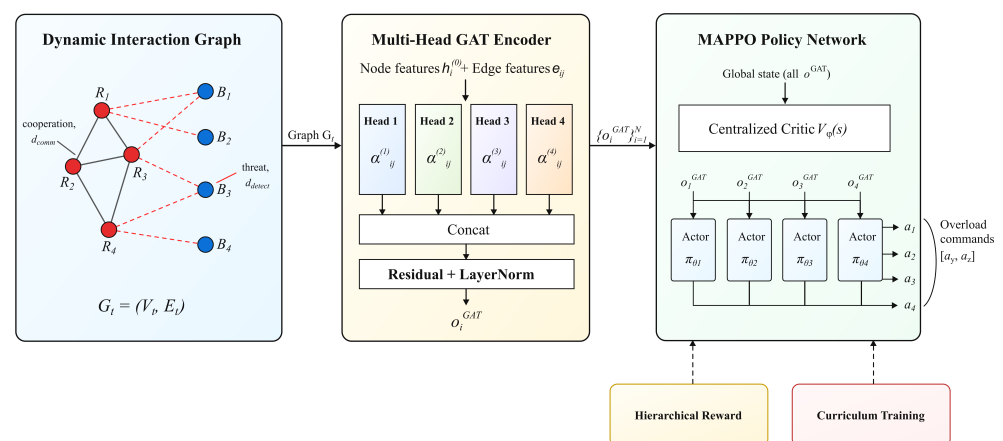


Figure 2. Overall architecture of the GAT-MAPPO cooperative guidance framework. **Left:** dynamic interaction graph construction with N Red interceptor nodes and M Blue target nodes (gray solid edges for interceptor–interceptor cooperation, red dashed edges for interceptor–target threat). **Center:** multi-head GAT observation encoder with K -head attention, residual connections, and layer normalization. **Right:** MAPPO policy network with a centralized critic and N decentralized actors. **Bottom:** hierarchical reward function and curriculum adversarial training.

4.1. Dynamic Interaction Graph Construction

At each decision step t , a heterogeneous dynamic interaction graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ is constructed to encode the current engagement topology. The node set $\mathcal{V}_t = \mathcal{V}^R \cup \mathcal{V}^B$ contains all interceptor and target nodes. Each interceptor node $i \in \mathcal{V}^R$ is associated with a feature vector $\mathbf{h}_i^{(0)} = [\mathbf{p}_i, \mathbf{v}_i, a_{\text{rem},i}] \in \mathbb{R}^{d_R}$ encoding its position, velocity, and residual overload capacity. Each target node $j \in \mathcal{V}^B$ carries $\mathbf{h}_j^{(0)} = [\hat{\mathbf{p}}_j, \hat{\mathbf{v}}_j, \hat{\mathbf{m}}_j] \in \mathbb{R}^{d_B}$, where the hat notation indicates estimated quantities that incorporate sensor delay and noise. Because $d_R \neq d_B$ in general, all node features are projected to a common dimension d through a shared linear layer $\mathbf{W}_{\text{proj}}$ before entering the attention layers.

The edge set $\mathcal{E}_t$ comprises two distinct types of connections. Interceptor–interceptor cooperation edges are defined as $\mathcal{E}_t^{RR} = \{(i, j) \mid i, j \in \mathcal{V}^R, \|\mathbf{p}_i - \mathbf{p}_j\| \leq d_{\text{comm}}\}$, connecting pairs within communication range to reflect coordination potential. Similarly, interceptor–target threat edges are $\mathcal{E}_t^{RB} = \{(i, j) \mid i \in \mathcal{V}^R, j \in \mathcal{V}^B, \|\mathbf{p}_i - \hat{\mathbf{p}}_j\| \leq d_{\text{detect}}\}$, reflecting immediate threat awareness. Target-to-target edges are intentionally excluded because targets operate independently; introducing such edges would erroneously inject spurious correlations into the graph structure. Each edge (i, j) is characterized by a feature

vector $\mathbf{e}_{ij} = [\Delta \mathbf{p}_{ij}/d_{\max}, \Delta \mathbf{v}_{ij}/V_{\max}, \dot{q}_{ij}]$, encoding normalized relative position, relative velocity, and LOS angle rate. This distance-threshold rule enables a dynamic graph topology: the graph is sparse during early phases but densifies naturally as interceptors converge toward the target area.

4.2. Multi-Head Graph Attention Observation Encoder

On the dynamic interaction graph, we employ a multi-head GAT mechanism [17] (Figure 3) to adaptively aggregate neighborhood information for each node. Crucially, our formulation enhances standard GAT by incorporating edge features directly into the attention coefficient computation, allowing geometry to modulate the aggregation weights. For any node i and its neighbor $j \in \mathcal{N}_i$, the k -th head calculates the raw attention score as

$$e_{ij}^{(k)} = \text{LeakyReLU}\left(\mathbf{a}^{(k)\top} \left[\mathbf{W}^{(k)} \mathbf{h}_i \parallel \mathbf{W}^{(k)} \mathbf{h}_j \parallel \mathbf{W}_e^{(k)} \mathbf{e}_{ij} \right]\right), \quad (4)$$

where $\mathbf{W}^{(k)} \in \mathbb{R}^{d' \times d}$ and $\mathbf{W}_e^{(k)} \in \mathbb{R}^{d' \times d_e}$ are the learnable projection matrices, $\mathbf{a}^{(k)} \in \mathbb{R}^{3d'}$ is the shared attention coefficient vector, and $\parallel$ denotes concatenation. The normalized raw score yields the weight $\alpha_{ij}^{(k)} = \text{softmax}_j(e_{ij}^{(k)})$. The k -th head aggregates output is then:

$$\mathbf{h}_i^{(k)} = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(k)} \mathbf{W}^{(k)} \mathbf{h}_j \right), \quad (5)$$

where σ is the ELU activation. The final node representation is obtained by concatenation across K heads, followed by a residual connection and layer normalization:

$$\mathbf{h}_i' = \text{LN} \left(\left[\mathbf{h}_i^{(1)} \parallel \dots \parallel \mathbf{h}_i^{(K)} \right] + \mathbf{W}_{\text{res}} \mathbf{h}_i \right). \quad (6)$$

After L layers of encoding, the final graph-enriched observation for interceptor i is $\mathbf{o}_i^{\text{GAT}} = [\mathbf{h}_i^{(L)} \parallel \bar{\alpha}_i]$, where $\bar{\alpha}_i$ is a summary (mean and max) of attention weights across all heads, providing a concise topological snapshot to the MAPPO module. At a physical level, this encoder achieves two functions: first, information aggregation via cooperation edges allows decentralized nodes to perceive teammates' intent at each decision step; second, relevance assessment through attention weights naturally gauges relative priority. This continuous dynamic reweighting of influence inherently mirrors the crucial progression from global mid-course coordination to terminal spatial precision without requiring explicit control logic.

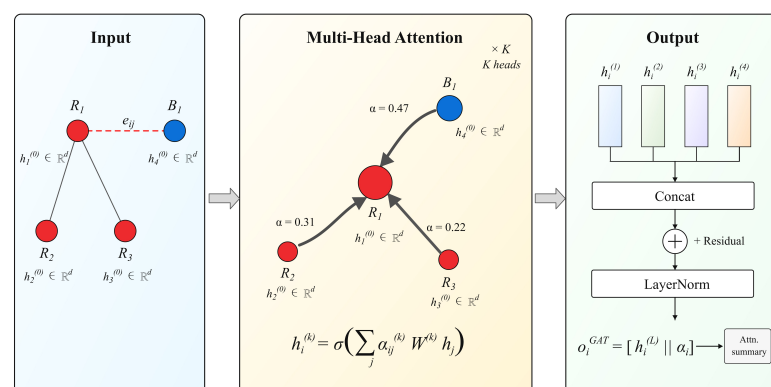


Figure 3. Internal structure of the multi-head graph attention observation encoder. **Left:** dynamic interaction graph with Red interceptor nodes and Blue target nodes. **Center:** K -head attention computation incorporating edge features, followed by concatenation, ELU activation, residual connection, and layer normalization. **Right:** output graph-enriched observation vector $\mathbf{o}_i^{\text{GAT}}$.

4.3. Hierarchical Cooperative Reward Function

Multi-interceptor cooperative guidance must simultaneously satisfy several competing performance objectives: small miss distance, synchronized arrival time, diversified interception angles, and smooth overload commands. Combining these into a single scalar reward with fixed weights is prone to gradient interference and training oscillation. To address this, the total reward is decomposed into four independently defined components:

$$r_i(t) = w_1 r_{\text{miss}}(t) + w_2 r_{\text{sync}}(t) + w_3 r_{\text{angle}}(t) + w_4 r_{\text{smooth}}(t). \quad (7)$$

The miss-distance reward $r_{\text{miss}} = -d_i(t)/d_i(0) + \beta \cdot \mathbb{I}[d_i(t_f) < d_{\text{int}}]$ encourages continuous distance reduction and provides a bonus $\beta = 10$ upon successful interception, where $d_i(0)$ normalizes for initial condition differences. The simultaneous-arrival reward $r_{\text{sync}} = -\lambda_s \cdot \sigma(\{t_{go,i}\}_{i=1}^N)$ penalizes the standard deviation of estimated times-to-go, with $\lambda_s = 2.0$. The angle-diversity reward $r_{\text{angle}} = \lambda_a \cdot |\mathcal{P}|^{-1} \sum_{(i,j) \in \mathcal{P}} |\Delta\phi_{ij}(t)|$ encourages angular spread among interceptors targeting the same object, with $\lambda_a = 0.5$. The overload-smoothness reward $r_{\text{smooth}} = -\lambda_u \|\mathbf{a}_i(t) - \mathbf{a}_i(t-1)\|^2$ suppresses command chattering, with $\lambda_u = 0.1$.

The weights $\{w_k\}$ are dynamically annealed during training. Letting $\eta \in [0, 1]$ denote the normalized training progress, the weights are set as $w_1(\eta) = 0.3 + 0.4\eta$, $w_2(\eta) = 0.35 - 0.15\eta$, $w_3(\eta) = 0.25 - 0.15\eta$, and $w_4 = 0.1$ (constant). In early training, the cooperation-related weights (w_2, w_3) dominate so that agents quickly develop global coordination awareness. As training progresses, the guidance-precision weight (w_1) gradually increases, shifting the optimization focus toward individual strike effectiveness. This annealing scheme effectively mitigates gradient conflicts between cooperative and precision objectives.

4.4. Curriculum Adversarial Training

Training MARL policies from random initialization directly against highly agile adversaries typically leads to convergence failure, because the initial random policy receives almost no positive reward signal in a strongly adversarial environment, resulting in extremely noisy gradient estimates. Inspired by curriculum learning [22], which incrementally increases task difficulty during training, a three-stage progressive training strategy is designed (Figure 4). This approach addresses the convergence challenge that persists across different MARL paradigms: counterfactual policy gradient methods such as COMA [23] address multi-agent credit assignment but do not resolve the reward sparsity inherent in initial random policies facing strong adversaries; hierarchical decision-making frameworks for UAV swarm confrontation [24] and mean-field reinforcement learning for large-scale engagements [25] likewise resort to structured or staged training to overcome initial policy instability.

Stage I: Basic cooperative learning. Targets execute rectilinear constant-velocity flight. This simplified environment allows the policy network to rapidly acquire fundamental trajectory tracking and temporal coordination skills, and provides the GAT encoder with meaningful initial parameters for topology extraction. Training proceeds from random initialization for $E_1 = 2000$ episodes. The stage transition criterion requires that the interception success rate exceeds 95% and the simultaneous-arrival error falls below 0.5 s across 100 randomly initialized evaluation runs.

Stage II: Maneuvering adversarial learning. Targets randomly select from a pre-defined evasion maneuver library containing five basic patterns: sinusoidal weaving (frequency 0.5–2 Hz, amplitude 5–15 g), cosine barrel rolls, step turn maneuvers, spiral descent, and random combinations thereof. Training resumes from the Stage I checkpoint for $E_2 = 3000$ additional episodes. The cooperation reward weights are gradually reduced

according to the annealing schedule, while the precision weights increase. The transition criterion requires maintaining at least 88% success rate against the most challenging combined maneuver pattern.

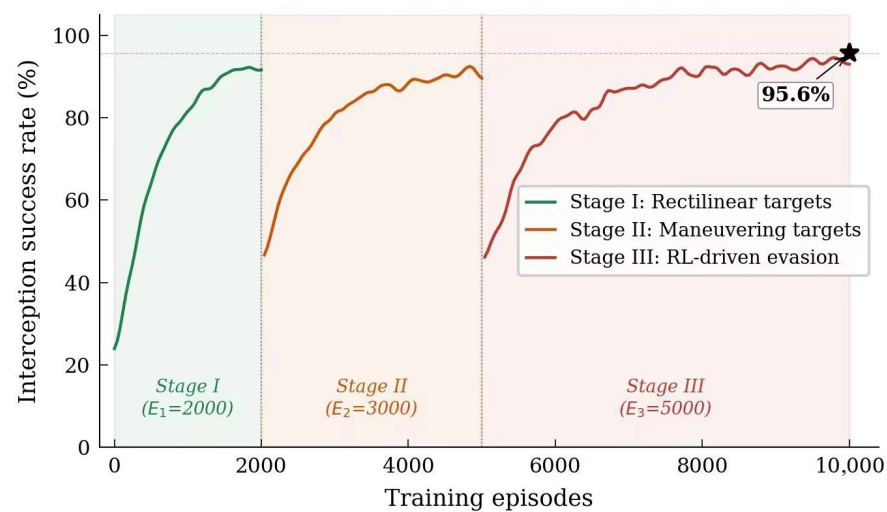


Figure 4. Three-stage curriculum adversarial training strategy. Stage I ($E_1 = 2000$ episodes): rectilinear targets, learning basic cooperative guidance. Stage II ($E_2 = 3000$ episodes): predefined maneuver library, learning adversarial coordination. Stage III ($E_3 = 5000$ episodes): RL-driven adaptive evasion, interceptor–evader self-play. The training curves illustrate the interception success rate versus training episodes for the three target types.

Stage III: Self-play adversarial training. Targets are controlled by an independently trained RL agent (Blue policy) whose objective is to maximize escape probability. The Blue policy adopts a single-agent PPO architecture with a 3-layer MLP (hidden dimension 128, Tanh activations). Its observation consists of the target’s own state and the estimated positions and velocities of all detected interceptors within a 20 km sensing radius. The Blue action space is $\mathcal{A}^B = [-a_{\max}^B, a_{\max}^B]^2$ with $a_{\max}^B = 20$ g, and the evasion reward is defined as $r^B = \min_i d_i(t_f)/d_{\max} - \lambda_e \|\mathbf{a}^B(t)\|^2$, balancing escape distance maximization with energy efficiency. Red and Blue train in alternation: every $\Delta E = 200$ episodes, one side is frozen while the other updates. The Blue strategy pool retains the five most recent checkpoints, and at each training iteration the opponent is sampled uniformly from this pool to prevent overfitting to a single adversary. Training resumes from the Stage II checkpoint for $E_3 = 5000$ episodes. We emphasize that the evader policy is used only to generate challenging simulated maneuver patterns during curriculum training; it is not intended to represent a validated real target-control system.

4.5. MAPPO Policy Architecture and Optimization

Under the CTDE paradigm, each interceptor i is equipped with an independent actor network $\pi_{\theta_i}(\mathbf{a}_i|\mathbf{o}_i^{\text{GAT}})$ that maps the graph-enriched observation to a Gaussian distribution over the overload command $\mathbf{a}_i = [a_{yi}, a_{zi}]$. All interceptors share a single centralized critic network $V_{\phi}(\mathbf{s})$ whose input is the concatenation of all interceptors’ graph-enriched features. Both actor and critic networks are three-layer MLPs with hidden dimension 256 and Tanh activations. The policy update employs the clipped PPO objective [13]:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}[\min(\rho_t \hat{A}_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) \hat{A}_t)], \quad (8)$$

where $\rho_t = \pi_{\theta}(\mathbf{a}_t|\mathbf{o}_t)/\pi_{\theta_{\text{old}}}(\mathbf{a}_t|\mathbf{o}_t)$ is the importance sampling ratio, $\hat{A}_t$ is the generalized advantage estimate (GAE) with $\lambda_{\text{GAE}} = 0.95$, and $\epsilon = 0.2$. The GAT encoder parameters

are optimized jointly with the actor and critic in an end-to-end fashion; gradients flow back through the graph-enriched observations to the encoder weights, allowing the learned attention coefficients to be directly shaped by the policy objective. Compared to off-policy approaches such as DDPG [26] and SAC [27], which achieve sample efficiency through experience replay, the on-policy PPO framework with clipping provides more stable gradient estimates in multi-agent cooperative settings where multiple policies are updated concurrently. This choice aligns with findings in phase-adaptive reinforcement learning for missile guidance [28], which likewise favors on-policy methods for their training stability in control-critical aerospace domains.

5. Simulation Experiments

5.1. Experimental Setup

The simulation environment is built upon the three-dimensional kinematics model described in Section 3, numerically integrated using a fourth-order Runge–Kutta method with a time step of $\Delta t = 0.01$ s. The decision interval for the RL policy is $T_{\text{step}} = 0.1$ s (i.e., one RL step spans 10 integration steps). The key physical parameters are: interceptor speed $V_R = 900$ m/s, target speed $V_B = 600$ m/s, maximum overload $a_{\text{max}} = 30$ g, intercept radius $d_{\text{int}} = 5$ m, communication range $d_{\text{comm}} = 20$ km, and detection range $d_{\text{detect}} = 30$ km. The maximum episode duration is $T_{\text{max}} = 60$ s. Initial interceptor-to-target distances are uniformly sampled from $[15.0, 25.0]$ km, and initial flight-path angles are uniformly sampled within the feasible envelope. To model sensing imperfections, zero-mean Gaussian noise is added to the target state observations: position noise $\sigma_p = 50$ m, velocity noise $\sigma_v = 5$ m/s, and a communication delay of one decision step (0.1 s) is applied to teammate state information. All observations are normalized to $[-1, 1]$ prior to network input.

The GAT encoder uses $L = 2$ layers with $K = 4$ attention heads and a hidden dimension of 128, with residual connections followed by layer normalization. This configuration was selected through preliminary experiments over $L \in \{1, 2, 3, 4\}$ and $K \in \{2, 4, 8\}$. Increasing from $L = 1$ to $L = 2$ yielded a 3.1 percentage point improvement in SR, whereas $L = 3$ showed marginal improvement ($<0.5\%$) and $L = 4$ caused a 1.8% performance drop accompanied by increased feature cosine similarity across nodes (from 0.31 to 0.67), indicating the onset of oversmoothing. For attention heads, $K = 4$ achieved the best trade-off: $K = 2$ underperformed by 1.4% SR, while $K = 8$ increased computational cost by 38% without meaningful accuracy gain ($<0.3\%$). The residual connections in Equation (6) further mitigate oversmoothing by preserving lower-level features. The actor and critic MLPs each have 3 layers with hidden dimension 256 and Tanh activations. The learning rate is 3×10^{-4} with linear decay, the GAE parameter is $\lambda_{\text{GAE}} = 0.95$, the discount factor is $\gamma = 0.99$, and each update uses 10 mini-batches iterated over 5 epochs. Training is conducted on 2 NVIDIA RTX 4090 GPUs with parallel environment sampling; the three curriculum stages require approximately 36 h in total.

Five baselines were selected for comparison: (1) Augmented Proportional Navigation (APN) [8], a classical guidance benchmark; (2) Cooperative Sliding Mode Guidance (CSMG) [4], a cooperative control baseline; (3) MADDPG [10]; (4) MAPPO [12]; and (5) GCN-MAPPO, a variant replacing GAT with GCN [16] in the proposed framework. All MARL baselines use identical network capacities and training hyperparameters, differing only in observation encoding, to ensure a fair comparison. We note that value-decomposition methods such as QMIX [11] and VDN [20] are designed for cooperative tasks with discrete action spaces. Since the cooperative guidance problem studied here requires continuous overload command generation, these methods are not directly applicable without significant architectural modification (e.g., action discretization), which would in-

introduce quantization errors and compromise guidance precision. Therefore, MADDPG and MAPPO were selected as the primary MARL baselines, as they natively support continuous action spaces and represent state-of-the-art policy gradient and actor–critic approaches, respectively. Five evaluation metrics are reported: interception success rate (SR), average miss distance (AMD), simultaneous-arrival error (SAE, the standard deviation of individual arrival times), angle diversity (AD, the standard deviation of interception angles; higher is better), and average overload rate of change (AOR, reflecting command smoothness). Each configuration is evaluated over 500 independent Monte Carlo runs, and mean $\pm$ standard deviation values are reported.

5.2. Main Comparison in the Four-versus-Four Scenario

Table 1 presents the quantitative comparison in the 4-vs.-4 engagement with RL-driven target evasion. All metrics are reported as mean $\pm$ standard deviation over 500 independent Monte Carlo runs. GAT-MAPPO achieves the best performance across all five metrics. Compared to MAPPO, it improves the success rate from $86.3 \pm 2.8\%$ to $95.6 \pm 1.6\%$ (an absolute gain of 9.3 percentage points), reduces the simultaneous-arrival error by 41.4% (from 0.58 ± 0.14 s to 0.34 ± 0.07 s), and increases the angle diversity by 20.3% (from $40.3 \pm 3.7^\circ$ to $48.5 \pm 2.9^\circ$). The non-overlapping confidence intervals confirm the statistical significance of these improvements. Moreover, GAT-MAPPO consistently exhibits the smallest standard deviations among all methods, indicating that the graph attention mechanism not only enhances average performance but also improves stability across diverse engagement geometries. The underlying mechanism is that the GAT encoder, through cooperation edges, allows each interceptor's observation to implicitly encode the positions and flight intentions of its teammates, enabling implicit information sharing within decentralized decision-making.

Table 1. Performance comparison in the 4-vs.-4 scenario with RL-driven target evasion (500 Monte Carlo runs; mean $\pm$ std reported).

Method	Category	SR (%)	AMD (m)	SAE (s)	AD ($^\circ$)	AOR (g/s)
APN	Classical	64.1 ± 4.7	8.15 ± 2.63	1.73 ± 0.42	22.6 ± 5.1	15.2 ± 3.8
CSMG	Cooperative	73.9 ± 3.9	6.04 ± 1.87	0.89 ± 0.23	31.7 ± 4.6	19.5 ± 4.2
MADDPG	MARL	81.5 ± 3.4	4.37 ± 1.52	0.76 ± 0.19	34.2 ± 4.3	12.4 ± 2.9
MAPPO	MARL	86.3 ± 2.8	3.19 ± 1.08	0.58 ± 0.14	40.3 ± 3.7	10.7 ± 2.4
GCN-MAPPO	GNN + MARL	92.8 ± 2.1	2.51 ± 0.74	0.41 ± 0.09	43.8 ± 3.2	9.3 ± 1.8
GAT-MAPPO	GNN + MARL	95.6 ± 1.6	1.72 ± 0.53	0.34 ± 0.07	48.5 ± 2.9	8.1 ± 1.5

Compared to GCN-MAPPO, GAT-MAPPO achieves a 2.8 percentage point higher success rate and a 31.5% lower miss distance. The key difference lies in the aggregation strategy: GCN applies uniform-weight spectral convolution over all neighbors, whereas GAT's attention weights enable each node to dynamically focus on the most relevant neighbors given the current situation. This differentiated aggregation is particularly valuable during the terminal phase when interceptor-to-target distances are shrinking rapidly.

Among classical methods, APN achieves only 64.1% success rate, as the closed-loop gain of proportional navigation is insufficient against RL-driven high-agility evasion. CSMG maintains a 73.9% success rate and 0.89 s SAE through its sliding-mode structure, but its AOR of 19.5 g/s is the highest among all methods, reflecting the inherent high-frequency chattering of sliding-mode control. MADDPG (81.5%) underperforms MAPPO (86.3%), consistent with the finding of Yu et al. [12] that PPO's clipping mechanism provides more stable training in continuous control than DDPG's deterministic policy gradient.

5.3. Scalability Analysis

Table 2 illustrates the performance degradation trend as the engagement scale expands from 3-vs.-2 to 6-vs.-6. All baseline methods show a decline in success rate with increased scale. Notably, GAT-MAPPO exhibits optimal performance retention: from 4-vs.-4 to 6-vs.-6, it drops only 10.2 percentage points, compared to steeper drops exhibited by MAPPO ($\Delta = 11.4\%$) and MADDPG ($\Delta = 13.4\%$). This superior scalability is fundamentally linked to the graph attention's local neighborhood aggregation property: its computational complexity scales linearly with the number of edges, $|\mathcal{E}|$, rather than quadratically with the number of nodes, $|\mathcal{V}|^2$. Furthermore, when numerical advantage exists (e.g., 3-vs.-2), performance gaps are compressed; the method's superiority is most pronounced as the interceptor–target ratio approaches parity, underscoring its relevance for critical high-density engagement scenarios.

Table 2. Interception success rate (%) across different engagement scales (RL-driven evasion).

Method	3-vs.-2	4-vs.-4	5-vs.-5	6-vs.-6
APN	78.5	64.1	52.8	47.3
CSMG	84.7	73.9	65.2	56.8
MADDPG	90.6	81.5	73.6	68.1
MAPPO	94.3	86.3	81.4	74.9
GCN-MAPPO	97.1	92.8	87.3	82.6
GAT-MAPPO	98.0	95.6	92.0	85.4

5.4. Qualitative Visualization

Figure 5 offers a direct trajectory comparison in 4-vs.-4 scenarios (top-down view). The visual evidence clearly corroborates the superior encirclement formation achieved by GAT-MAPPO via graph attention encoding. MAPPO results show decentralized interception patterns, often failing to achieve spatial enclosure and leaving targets vulnerable. Although GCN-MAPPO improves timing coordination, one target still successfully evades. In contrast, GAT-MAPPO demonstrates the four interceptors proactively adjusting trajectories using modeled cooperation edges, converging on the target cluster from distinct bearings with near-perfect temporal simultaneity, resulting in universal interception success. The visualization confirms that sophisticated local information exchange is paramount for robust, comprehensive collaborative targeting.

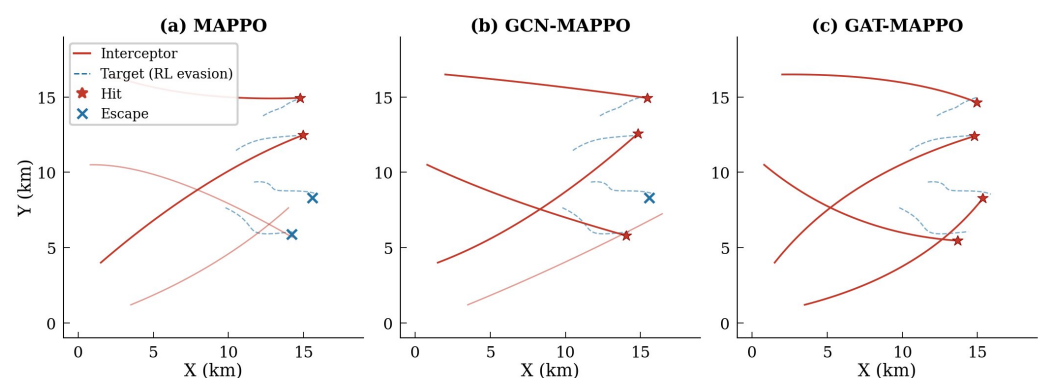


Figure 5. Representative 4-vs.-4 interception trajectories (top-down projection). (a) MAPPO: dispersed arrival times, two targets escape. (b) GCN-MAPPO: improved synchronization, one target escapes. (c) GAT-MAPPO: near-simultaneous multi-directional encirclement, all targets intercepted. Red solid lines denote interceptor trajectories; blue dashed lines denote target trajectories with RL evasion; stars mark successful hits; crosses mark escapes.

The visualization in Figure 6 reveals the strategic depth modeled by the GAT encoder through attention weights. During initial phases (distance > 15 km), attention weights are distributed relatively uniformly across both cooperation and threat edges due to low information distinctiveness. In the mid-course phase (8–15 km), cooperation-edge attention rapidly increases, signifying intensive coordination in time-to-go estimation required for arrival synchronization. When distances fall below 5 km, the focus dramatically shifts: target attention surges while teammate attention decreases, confirming an effective self-prioritization from global cooperative planning to localized terminal precision—a transition achieved entirely through data-driven feature weighting, bypassing manual switching logic.

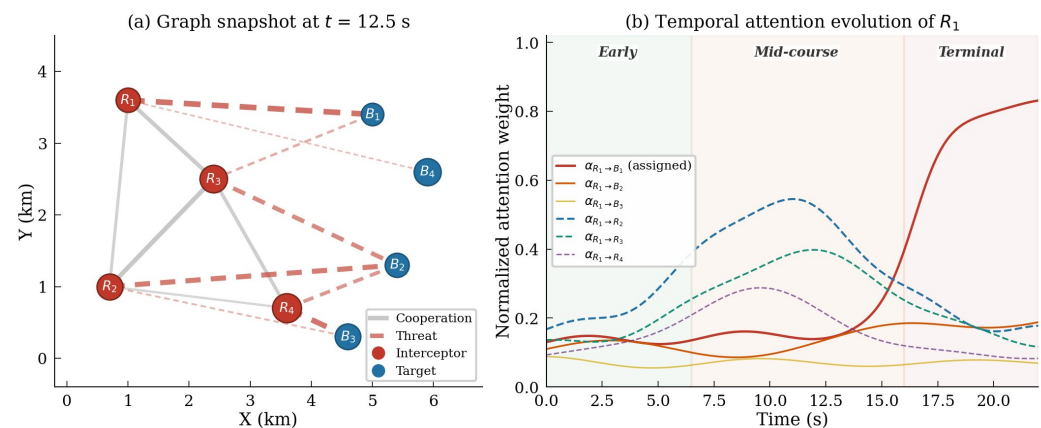


Figure 6. Attention weight visualization and cooperative situational analysis. (a) Dynamic interaction graph snapshot at a mid-course time step: node sizes are proportional to feature norms; edge widths and colors reflect attention coefficients α_{ij} . (b) Temporal evolution of interceptor R_1 's attention weights toward four targets and three teammates: target attention surges as closing distance decreases, while teammate attention peaks mid-course then diminishes, embodying the “mid-course cooperation, terminal precision” principle.

5.5. Ablation Studies

To isolate the contribution of each framework component, three groups of ablation experiments are conducted in the 4-vs.-4 scenario with RL-driven evasion.

5.5.1. Effect of Observation Encoding

Table 3 compares the efficiency of five observation encoding methodologies under a constrained MAPPO policy architecture. Standard MLP concatenation performs poorly, establishing that topological information is essential for cooperative guidance. Both GCN and GraphSAGE enhance performance by incorporating topology, but their aggregation methods (uniform spectral convolution; mean/max sampling) lack differentiation in neighbor importance. Notably, removing edge features from GAT (“GAT w/o edge features”) reduces the success rate from 95.6% to 93.1% and the angle diversity from 48.5° to 44.7° . This 2.5 percentage point drop demonstrates that edge features—encoding relative geometry, closing velocity, and LOS rate—provide critical contextual information that modulates the attention weights beyond what node features alone can capture. With edge features, the full GAT achieves the most marked gain in AD ($43.8^\circ \rightarrow 48.5^\circ$), confirming that accurate neighborhood prioritization informed by relational geometry is critical for modeling complex formation dynamics. GraphSAGE performs marginally below GCN, potentially due to the stochastic variance introduced by its sampling mechanism on smaller graph instances.

Table 3. Ablation study 1: effect of observation encoding method.

Encoder	SR (%)	AMD (m)	SAE (s)	AD (°)
MLP (concatenation)	86.3	3.19	0.58	40.3
GCN	92.8	2.51	0.41	43.8
GraphSAGE	89.7	2.83	0.49	42.1
GAT (w/o edge features)	93.1	2.24	0.40	44.7
GAT (proposed)	95.6	1.72	0.34	48.5

5.5.2. Effect of Reward Components

Table 4 systematically validates the contribution of each component reward. The miss-distance objective alone leads to individualistic pursuit (SAE = 1.41 s, AD = 24.8°). Incorporating r_{sync} significantly reduces SAE ($\rightarrow$ 0.47 s), yet this only marginally improves AD (30.1°), confirming that temporal and angular formation are separable, independently optimized objectives. The introduction of the angle-diversity reward yields a substantial leap in AD (30.1° $\rightarrow$ 47.2°), demonstrating its efficacy in driving multi-directional encirclement strategies. Finally, integrating r_{smooth} boosts the success rate (from 93.8% to 95.6%) without compromising other metrics, confirming that smooth path commands minimize residual energy waste and pointing errors at terminal range.

Table 4. Ablation study 2: incremental contribution of reward components.

Reward Configuration	SR (%)	AMD (m)	SAE (s)	AD (°)
r_{miss} only	84.3	2.71	1.41	24.8
+ r_{sync}	89.7	2.16	0.47	30.1
+ r_{angle}	93.8	1.93	0.38	47.2
+ r_{smooth} (full)	95.6	1.72	0.34	48.5

5.5.3. Effect of Curriculum Training

Table 5 underscores the vital necessity of a structured curriculum. Training exclusively against adaptive adversaries (Stage III only) leads to divergence (SR 79.4%, failure to converge within 10,000 episodes) because initial random policies receive negligible positive reward signals in such a hard environment. Omitting Stage II slows convergence by 52.8% and leaves the policy underdeveloped, as indicated by the performance drop when tested against reactive evasion patterns. The gap between only Stages I $\rightarrow$ II ($\sim$ 91.5%) and full curriculum (95.6%) critically highlights the unique value of self-play (Stage III), which ensures full adversarial generalization capability necessary for robust performance against diverse opponents. The complete three-stage approach achieves convergence to 95.6% in a manageable $\sim$ 5300 episodes, optimizing skill acquisition efficiency while maximizing final performance.

Table 5. Ablation study 3: effect of curriculum training strategy.

Training Strategy	SR (%)	SAE (s)	Convergence (Episodes)
Stage III only	79.4	0.56	>10,000 (not converged)
Stage I $\rightarrow$ III	88.6	0.45	$\sim$ 8100
Stage I $\rightarrow$ II	91.5	0.38	$\sim$ 4500
Stage I $\rightarrow$ II $\rightarrow$ III (full)	95.6	0.34	$\sim$5300

5.6. Computational Overhead Analysis

The GAT forward pass is analyzed for computational overhead under the adopted numerical simulation setting. In the 4-vs.-4 scenario, the GAT encoder overhead is 0.8 ms

per step (2 layers, 4 heads) on a single NVIDIA RTX 4090 GPU, compared to 0.3 ms for standard MAPPO. The total inference latency of 1.1 ms remains well below the 100 ms RL decision interval (T_{step}), indicating that the graph-encoding overhead is limited relative to the adopted simulation integration step. In an expansionary 6-vs.-6 scenario, the GAT overhead increases proportionally to 1.4 ms, maintaining feasibility within the simulation framework. We note that these latency measurements are obtained on a desktop GPU platform; deployment on embedded or flight-qualified hardware would require additional profiling and potential model compression (e.g., quantization, pruning), which is beyond the scope of the current algorithmic study. The total offline training cost is 36 h (vs. ≈ 22 h for MAPPO); this additional training cost is acceptable for offline algorithm development, but it should not be interpreted as evidence of real-platform deployment readiness.

5.7. Sensitivity and Robustness Analysis

To address the important question of robustness under degraded conditions within the numerical simulation framework, we conduct four groups of sensitivity experiments in the 4-vs.-4 scenario. All experiments use the fully trained GAT-MAPPO policy and are evaluated over 500 Monte Carlo runs per configuration.

5.7.1. Graph Construction Threshold Sensitivity

Table 6 examines the sensitivity to graph construction thresholds. The performance is relatively stable across reasonable parameter ranges: reducing d_{comm} from 20 km to 15 km decreases SR by only 1.4 percentage points (at $d_{\text{detect}} = 30$ km), while expanding d_{detect} beyond 30 km yields diminishing returns. The most significant factor is the communication range d_{comm} , as it directly affects the availability of cooperation edges for temporal coordination. The default values ($d_{\text{comm}} = 20$ km, $d_{\text{detect}} = 30$ km) achieve the best performance and are consistent with typical sensor specifications.

Table 6. Sensitivity to graph construction thresholds d_{comm} and d_{detect} (4-vs.-4, 500 runs).

d_{comm} (km)	d_{detect} (km)	SR (%)	AMD (m)	SAE (s)	AD (°)
15	25	93.1	2.08	0.39	45.9
15	30	94.2	1.91	0.37	46.8
20	25	94.8	1.86	0.36	47.3
20	30	95.6	1.72	0.34	48.5
20	35	95.3	1.78	0.35	48.1
25	30	95.0	1.83	0.36	47.6
25	35	94.7	1.89	0.37	47.2

5.7.2. Sensor Noise and Communication Delay Sensitivity

Table 7 reports the degradation profile under increasing sensing and communication impairments. Under moderate noise ($\sigma_p \leq 100$ m), the success rate remains above 94%, indicating that the graph attention mechanism exhibits inherent noise resilience through its weighted aggregation, which downweights unreliable observations. Performance degrades more noticeably at $\sigma_p = 300$ m, where the SR drops to 88.2%, suggesting that severe localization errors fundamentally compromise the quality of the interaction graph. Communication delay up to 0.3 s (3 decision steps) reduces SR by only 1.5 percentage points, as the GAT encoder partially compensates by relying on threat edges that use direct sensor measurements. Edge dropout experiments simulate communication packet loss by randomly removing a fraction of cooperation edges at each step. The policy remains robust up to 20% dropout (SR > 93%), but degrades more rapidly beyond this level as temporal coordination information becomes insufficient for simultaneous-arrival planning.

Table 7. Sensitivity to sensor noise level and communication delay (4-vs.-4, 500 runs).

Condition	SR (%)	AMD (m)	SAE (s)	AD (°)
Position noise σ_p (velocity noise σ_v scaled proportionally)				
$\sigma_p = 0$ m (ideal)	96.8	1.54	0.31	49.7
$\sigma_p = 50$ m (nominal)	95.6	1.72	0.34	48.5
$\sigma_p = 100$ m	94.3	1.95	0.38	47.1
$\sigma_p = 200$ m	91.7	2.43	0.46	44.6
$\sigma_p = 300$ m	88.2	3.06	0.55	42.3
Communication delay ($\sigma_p = 50$ m fixed)				
0 steps (0 s)	96.2	1.61	0.32	49.1
1 step (0.1 s, nominal)	95.6	1.72	0.34	48.5
3 steps (0.3 s)	94.1	1.98	0.40	46.7
5 steps (0.5 s)	92.3	2.31	0.47	44.8
Edge dropout rate ($\sigma_p = 50$ m, delay = 1 step)				
0% (nominal)	95.6	1.72	0.34	48.5
10%	94.7	1.88	0.37	47.3
20%	93.2	2.13	0.41	45.6
30%	90.8	2.52	0.48	43.1

5.7.3. Reward Weight Sensitivity

Table 8 investigates the sensitivity to initial reward weight configurations. The proposed annealed configuration (0.3, 0.35, 0.25, 0.1) achieves the best overall balance. The “Precision-first” variant prioritizes miss distance and achieves a slightly lower AMD (1.68 m) but sacrifices cooperation quality (SAE increases to 0.43 s). Conversely, “Cooperation-first” achieves superior SAE (0.31 s) and AD (49.1°) at the cost of reduced individual guidance precision. The “w/o annealing” row demonstrates that dynamic weight annealing provides a 1.8 percentage point improvement over fixed weights, validating the curriculum-aligned reward scheduling. These results confirm that while the proposed weights are near-optimal, the framework is not critically sensitive to moderate weight perturbations.

Table 8. Sensitivity to initial reward weights (4-vs.-4, 500 runs).

Weight Configuration	SR (%)	AMD (m)	SAE (s)	AD (°)
Proposed: (0.3, 0.35, 0.25, 0.1)	95.6	1.72	0.34	48.5
Equal: (0.25, 0.25, 0.25, 0.25)	93.4	1.95	0.39	46.7
Precision-first: (0.5, 0.2, 0.2, 0.1)	94.1	1.68	0.43	43.9
Cooperation-first: (0.2, 0.4, 0.3, 0.1)	92.7	2.14	0.31	49.1
w/o annealing (fixed proposed)	93.8	1.91	0.37	47.0

5.7.4. Robustness to Actuator Dynamics

To investigate the effect of actuator imperfections that are not modeled in the nominal three-dimensional kinematics framework, we introduce a first-order actuator lag model $\dot{a}_{\text{actual}} = (a_{\text{cmd}} - a_{\text{actual}})/\tau$ with time constant τ , and impose a command-rate limit $|\dot{a}_{\text{cmd}}| \leq \dot{a}_{\text{max}}$.

Table 9 evaluates robustness under actuator dynamics not present during training. With moderate lag ($\tau = 0.05$ s) and rate limits ($\dot{a}_{\text{max}} = 100$ g/s), the SR decreases by only 2.1 percentage points to 93.5%, indicating that the overload-smoothness reward (r_{smooth}) implicitly encourages command profiles that are compatible with realistic actuator bandwidth. Under more severe constraints ($\tau = 0.10$ s, $\dot{a}_{\text{max}} = 60$ g/s), the performance degradation becomes more pronounced (SR = 89.6%), which is expected as the commanded overloads can no longer be faithfully tracked. These results suggest that incorporating

actuator dynamics into the training environment—through domain randomization over τ —would further improve robustness and constitutes a promising direction for future work.

Table 9. Robustness to actuator lag and command-rate limits (4-vs.-4, 500 runs).

τ (s)	$\dot{a}_{\max}$ (g/s)	SR (%)	AMD (m)	SAE (s)	AD (°)
0 (ideal)	∞	95.6	1.72	0.34	48.5
0.02	150	94.8	1.83	0.36	47.9
0.05	100	93.5	2.07	0.39	46.4
0.08	80	91.9	2.38	0.43	44.7
0.10	60	89.6	2.81	0.49	42.8

6. Conclusions

In this paper, GAT-MAPPO was proposed as a cooperative guidance framework that deeply integrates graph attention networks into multi-agent proximal policy optimization to address the key challenges of multi-interceptor cooperative guidance in adversarial engagement scenarios. The dynamic heterogeneous interaction graph unifies inter-interceptor coordination and interceptor-to-target relevance relationships into a single topological representation. The multi-head graph attention encoder adaptively aggregates neighborhood information through learnable attention weights, overcoming the fundamental limitation of state concatenation that discards structural information. The hierarchical reward function decouples four competing performance objectives and employs dynamic weight annealing to mitigate gradient conflicts. The three-stage curriculum adversarial training strategy ensures stable convergence while progressively improving policy performance against increasingly challenging simulated evasion patterns.

Extensive Monte Carlo simulations validated the effectiveness of GAT-MAPPO. In the 4-vs.-4 scenario, the framework achieved a 95.6% interception success rate, a 41.4% reduction in simultaneous-arrival error, and a 20.3% improvement in angle diversity compared to the MAPPO baseline. Scalability experiments from 3-vs.-2 to 6-vs.-6 confirmed graceful performance degradation, with GAT-MAPPO maintaining 85.4% success rate even in the balanced 6-vs.-6 setting. Ablation studies verified the essential contributions of graph attention encoding (including the marginal benefit of edge features), reward hierarchy, and curriculum staging. Sensitivity analyses demonstrated robust performance across varying sensor noise levels, communication delays, edge dropout rates, graph construction thresholds, and reward weight configurations. Robustness tests with actuator lag models further confirmed the framework's tolerance to unmodeled dynamics under moderate conditions.

It should be noted that the present work is an algorithmic study validated in a numerical simulation environment based on three-dimensional engagement kinematics. The current validation does not include full six-degree-of-freedom (6-DOF) aerodynamic simulation, high-fidelity autopilot/actuator modeling, hardware-in-the-loop testing, or real-platform experiments. While the sensitivity analyses in Section 5.7 partially characterize robustness margins under degraded conditions, the reported results should be interpreted as evidence of the effectiveness of the proposed graph-enhanced MARL framework under the tested numerical simulation settings, rather than as proof of operational robustness or deployment readiness.

Three directions will be pursued in future work. First, under communication-constrained conditions, intermittent or bandwidth-limited inter-interceptor communication links will challenge graph construction, motivating the development of robust sparse graph attention mechanisms with missing-edge tolerance. Second, extending the framework to high-fidelity 6-DOF simulation environments with autopilot/actuator dynamics, state-

estimation degradation, and hardware-in-the-loop validation will be pursued to bridge the gap between algorithmic validation and practical deployment. Third, in larger-scale engagements involving tens of interceptors, hierarchical graph attention or graph pooling operations will be explored to maintain expressive power while controlling computational complexity.

Author Contributions: Conceptualization, Y.C. and J.Z.; methodology, J.Z.; software, J.Z.; validation, Y.C. and J.Z.; formal analysis, J.Z.; investigation, J.Z.; writing—original draft, J.Z.; writing—review and editing, Y.C.; visualization, J.Z.; supervision, Y.C.; project administration, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The simulation code and data supporting the findings of this study will be made available upon reasonable request to the corresponding author.

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

References

1. Jeon, I.S.; Lee, J.I.; Tahk, M.J. Impact-Time-Control Guidance Law for Anti-Ship Missiles. *IEEE Trans. Control Syst. Technol.* **2006**, *14*, 260–266. [\[CrossRef\]](#)
2. Zhao, S.Y.; Zhou, R. Cooperative Guidance for Multimissile Salvo Attack. *Chin. J. Aeronaut.* **2008**, *21*, 533–539. [\[CrossRef\]](#)
3. Zhang, Y.A.; Ma, G.X.; Liu, A.L. Guidance Law with Impact Time and Impact Angle Constraints. *Chin. J. Aeronaut.* **2013**, *26*, 960–966. [\[CrossRef\]](#)
4. Song, J.H.; Song, S.M.; Zhou, H.B. Adaptive Nonsingular Fast Terminal Sliding Mode Guidance Law with Impact Angle Constraints. *Int. J. Control Autom. Syst.* **2016**, *14*, 99–114. [\[CrossRef\]](#)
5. Yang, F.; Ye, J.; Xue, X.; Luo, R.; Shao, L. Design of Terminal Guidance Law for Cooperative Multiple Vehicles Based on Prescribed Performance Control. *Aerospace* **2025**, *12*, 898. [\[CrossRef\]](#)
6. Liu, J.; Guo, Y.; Wang, S.; Zhang, S.; Sang, J.; Zhang, C. A Spatio-Temporally Cooperative Guidance Law for Highly Maneuverable Target. *Aerospace* **2025**, *12*, 1079. [\[CrossRef\]](#)
7. Zhang, C.; Wang, X.; Li, H.; Chen, J.; Liu, S.; Zhao, Y. Integrated Design of Cooperative Detection and Guidance Considering Equal Numbers of Aircraft on Both Sides. *Aerospace* **2025**, *12*, 1112. [\[CrossRef\]](#)
8. He, S.; Lee, C.H. Optimality of Error Dynamics in Missile Guidance Problems. *J. Guid. Control Dyn.* **2018**, *41*, 1522–1540. [\[CrossRef\]](#)
9. Zhang, J.; Li, J.; Shao, L.; Ye, J.; He, Y. Midcourse Guidance via Variable-Discrete-Scale Sequential Convex Programming. *Aerospace* **2025**, *12*, 952. [\[CrossRef\]](#)
10. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Long Beach, CA, USA, 2017; Volume 30.
11. Rashid, T.; Samvelyan, M.; Schroeder de Witt, C.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *J. Mach. Learn. Res.* **2020**, *21*, 178.
12. Yu, C.; Velu, A.; Vinitzky, E.; Gao, J.; Wang, Y.; Baez, A.; Liang, H.; Wu, Y. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Adv. Neural Inform. Process. Syst.* **2022**, *35*, 24611–24624. [\[CrossRef\]](#)
13. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
14. Cui, H.; Zhang, K.; Tan, M.; Wang, J. A Deep Reinforcement Learning-Based Cooperative Guidance Strategy Under Uncontrollable Velocity Conditions. *Aerospace* **2025**, *12*, 411. [\[CrossRef\]](#)
15. Chi, P.; Wei, J.H.; Wu, K.; Di, B.; Wang, Y.X. A Bio-Inspired Decision-Making Method of UAV Swarm for Attack-Defense Confrontation via Multi-Agent Reinforcement Learning. *Biomimetics* **2023**, *8*, 222. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
17. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.
18. Li, S.; Zhao, B.; Liu, H.; Wang, Z.; Zhang, H. Graph Neural Network-Enhanced Multi-Agent Reinforcement Learning for Intelligent UAV Confrontation. *Aerospace* **2025**, *12*, 687. [\[CrossRef\]](#)
19. Li, J.; Liu, P.; Zhang, H.; Li, C.; Yu, H.; Yu, X. Guidance Laws for Multi-Agent Cooperative Interception from Multiple Angles Against Maneuvering Target. *Aerospace* **2025**, *12*, 531. [\[CrossRef\]](#)

20. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-Decomposition Networks for Cooperative Multi-Agent Learning. In Proceedings of the AAMAS, Stockholm, Sweden, 10–15 July 2018; pp. 2085–2087.
21. Iqbal, S.; Sha, F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In Proceedings of the ICML, Long Beach, CA, USA, 9–15 June 2019; pp. 2961–2970.
22. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum Learning. In Proceedings of the ICML, Montreal, QC, Canada, 14–18 June 2009; pp. 41–48.
23. Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; Whiteson, S. Counterfactual Multi-Agent Policy Gradients. In Proceedings of the AAAI, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
24. Chang, G.; Ren, S.; Zhang, S.; Zhang, X. Hierarchical Decision Making-Based Intelligent Game Confrontation on UAV Swarm. *Aerospace* **2025**, *12*, 1033. [[CrossRef](#)]
25. Wang, B.; Li, S.; Gao, X.; Xie, T. Weighted Mean Field Reinforcement Learning for Large-Scale UAV Swarm Confrontation. *Appl. Intell.* **2023**, *53*, 5274–5289. [[CrossRef](#)]
26. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. In Proceedings of the ICLR, San Juan, Puerto Rico, 2–4 May 2016.
27. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the ICML, Stockholm, Sweden, 10–15 July 2018; pp. 1861–1870.
28. Tan, C.; Wang, J.; Cai, H.; Hu, S.; Zhang, B.; Zhu, W. Phase-Adaptive Reinforcement Learning for Self-Tuning PID Control of Cruise Missiles. *Aerospace* **2025**, *12*, 849. [[CrossRef](#)]

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.