# MSMAR-RL: Multi-Step Masked-Attention Recovery Reinforcement Learning for Safe Maneuver Decision in High-Speed Pursuit-Evasion Game

**Yang Zhao** [1,2,3] *, **Wenzhe Zhao**[1] , **Xuelong Li**[4]
[1]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai
[3]National Key Laboratory of Air-based Information Perception and Fusion, LuoYang , China
[4]Institute of Artificial Intelligence (TeleAI), China Telecom, China
izhaoyang@nwpu.edu.cn, zhaowenzhe@mail.nwpu.edu.cn, xuelong_li@ieee.org

## Abstract

Ensuring the safety of high-speed agent in dynamic adversarial environments, such as pursuit-evasion games with target-purchase and obstacle-avoidance, is a significant challenge. Existing reinforcement learning methods often fail to balance safety and reward under strict safety constraints and diverse environmental conditions. To address these limitations, this paper proposes a novel zero-constraint-violation recovery RL framework tailored for high-speed uav pursuit-evasion combat games. The framework includes three key innovations. (1) An extendable multi-step reach-avoid theory: we provide a zero-constraint-violation safety guarantee for multi-strategy reinforcement learning and enabling early danger detection in high speed game. (2) A masked-attention recovery strategy: we introduce a padding-mask attention architecture to handle spatiotemporal variations in dynamic obstacles with varying threat levels. (3) Experimental validation: we validate the framework in obstacle-rich pursuit-evasion scenarios, demonstrating its superiority through comparison with other algorithm and ablation studies. Our approach also shows potential for extension to other rapid-motion tasks and more complex hazardous scenarios. Details and code could be found at https://msmar-rl.github.io.

## 1 Introduction

Reinforcement Learning (RL) has made significant progress in various fields in recent years, particularly in games [Vinyals *et al.*, 2019], robotic control [He *et al.*, 2024], and autonomous driving [Wu *et al.*, 2024]. Among these applications, high-speed aircrafts pursuit games, which involve dynamic adversarial environments where UAVs must simultaneously pursue targets, avoid obstacles, and execute tasks, present one of the most challenging scenarios.

In recent years, reinforcement learning methods have achieved significant success in UAV control and game theory. In [Zhao *et al.*, 2024a; Guo *et al.*, 2022], the researchers used different RL algorithm to successfully win in the games without considering safe conditions. Meanwhile, safe reinforcement learning methods have been applied in path planning or low-dimension combat scenarios. Despite these achievements, ensuring the safety of high-speed aircraft in adversarial environments, especially during pursuit-evasion and obstacle avoidance tasks, remains a critical challenge.

In high-speed UAV game scenarios, UAVs face the following key challenges:

1. Safety Criticality and Physical Limitations: Due to strong inertia and limited maneuverability of high-speed agent, obstacle avoidance decisions must be made earlier to strictly ensure zero violation of safety.
2. Dynamic Input and Hazard Heterogeneity: The number of hazardous factors varies dynamically over time, and the severity of risks posed by multiple obstacles differs spatially. Consequently, there is a critical need to address these spatiotemporal challenges by handling dynamic inputs while effectively mining and utilizing information on varying hazard levels.

Traditional safe reinforcement learning have weakness in addressing the key challenges listed above. Firstly, they depend too much on reward shaping to balance safety and reward in complex scenarios, which may cause conservative or risky decisions [Liang *et al.*, 2018; Li *et al.*, 2023]. Secondly, in terms of the extreme danger and limited maneuver in flight, a zero-constraint-violation safety theory is urgently needed [Ying *et al.*, 2022]. Thirdly, facing with complexity of hazards in both temporal and spatial variations, algorithms need to possess strong adaptive and generalization capabilities.

To address the challenges and limitations of existing methods, this paper introduces Recovery Reinforcement Learning to high-speed pursuit-evasion game for the first time and makes the following three key contributions:

- A Extendable Multi-Step Reach-Avoid Theory for Zero-Constraint-Violation : We extend the reach-avoid theory to Recovery Reinforment Learning with multi-strategy, which provides a theoretical foundation for quantifying safety boundaries in UAV game scenarios. Based on

*Corresponding Author

this theory, a multi-step safety-discriminant value is proposed for earlier danger detection.

- Masked-Attention Recovery Strategy: Faced with dangerous observations changing with time and space, we introduce a recovery strategy based on padding-mask attention mechanism. This mechanism not only enable agent to tackle different numbers of obstacle avoidance by padding-mask method, but it also helps select critical features, improving training efficiency and recovery capabilities.
- Experimental Validation: We validate the framework in obstacle-rich pursuit-evasion scenarios, demonstrating its superiority over state-of-the-art methods through comparative and ablation studies. The results show zero-constraint-violation performance and highlight the framework's potential for extension to rapid-motion tasks and complex hazardous scenarios.

## 2 Related Work

### 2.1 UAV Pursuit Game

The target of UAV pursuit-evasion games is to position adversarial UAVs within own firing range while avoiding being targeted. Various methods, including expert systems, game theory, and reinforcement learning, have been applied to this domain and have achieved significant success [Pope *et al.*, 2022]. Among these, RL-based methods have shown the best performance due to their ability to learn adaptive strategies in complex and dynamic environments.

Recent studies have also made progress in incorporating safety considerations into UAV adversarial games [Vinod *et al.*, 2022; Yue *et al.*, 2023]. These works use techniques such as barrier functions and constrained RL to ensure collision avoidance and safe operation. However, they often lack rigorous theoretical safety guarantees and are limited to simple, low-dimensional scenarios. These limitations hinder their applicability to real-world, high-speed UAV combat scenarios.

### 2.2 Recovery Reinforcement Learning

Recovery RL is a reinforcement learning approach designed to help agents recovery from hazardous or suboptimal states during task execution. It employs a hierarchical policy architecture: a main policy for task completion and a recovery policy for safety [Thananjeyan *et al.*, 2021]. To detect risks and trigger recovery actions, methods like safety constraints, model predictive control (MPC), and uncertainty estimation are commonly used [Zhao *et al.*, 2024b] .

Recovery RL is applied in autonomous driving and robotics, enabling agents to handle emergencies like collisions or sensor failures [Zhao *et al.*, 2023]. However, challenges remain including lack of theoretical safety guarantees and formal verification. Additionally, most applications are in simple environments, limiting its effectiveness in complex, real-world scenarios like urban traffic or multi-robot systems.

### 2.3 Hamilton-Jacobi Reachability Analysis

Hamilton-Jacobi (HJ) reachability analysis is a rigorous tool that verifies the safety and liveness of a dynamic system [Ganai *et al.*, 2024]. For safety analysis, HJ reachability can provide the set of initial states from which the system may be forced into the failure set despite best-case efforts. This verification method provides guarantees on the safety properties of a system and generalizes to various challenging problem settings. These include problems with nonlinear dynamics, reach-avoid problems with time-varying goals or constraints and so on.

The Reach-Avoid value, derived from HJ reachability analysis, has demonstrated success in safety-critical applications such as autonomous driving and robotics [Akshay *et al.*, 2024; Hsu *et al.*, 2021]. However, it faces several challenges: (1) solving the HJ equations is computationally expensive, especially in high-dimensional spaces; (2) the safety thresholds often require empirical tuning, which limits its practicality; and (3) its application in high-speed, maneuver-constrained scenarios, such as UAV combat, remains underexplored.

## 3 Preliminaries

This section introduces the foundational knowledge and theoretical framework required for our proposed method, covering four main aspects: (1) Constrained Markov Decision Process (CMDP). (2) Reach-Avoid theory.

### 3.1 Constrained Markov Decision Process

In high-speed UAV games, the goal of reinforcement learning is not only to maximize cumulative rewards but also to satisfy safety constraints. This problem can be formalized as a Constrained Markov Decision Process (CMDP), defined by a tuple (S,A,P,R,C), where S is the state space,A is the action space, P(s|s,a) is the state transition probability,R(s,a) is the reward function,C(s,a) is the cost function used to quantify safety constraints.

The cumulative rewards could be presented as the equation below.

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

The objective of CMDP is to find a policy $\pi$ that maximizes the cumulative reward while satisfying the safety constraint.$J_c(\pi)$ is the cumulative constraints similar to reward, and $J_c(\pi)$ should satisfy the constraint $d_i$ each.

$$\begin{aligned} \pi_{k+1} = \arg\max_{\pi \in \Pi_k} J(\pi) \\ \text{s.t. } J_c(\pi) \le d_i \quad i = 1, \ldots, m. \end{aligned} \tag{2}$$

### 3.2 Reach Avoid Theory

To quantify safety of game, we introduce Reach-Avoid theory, which provides a theoretical guidance for getting to target and avoiding the obstacles at the same time. The key part is to defines the Failure Set, Safe Set, Target Set and Reach-Avoid Set.

(1) Failure Set and Safe Set

Failure Set represents the set of states where the UAV enters an unsafe condition, typically including collisions with

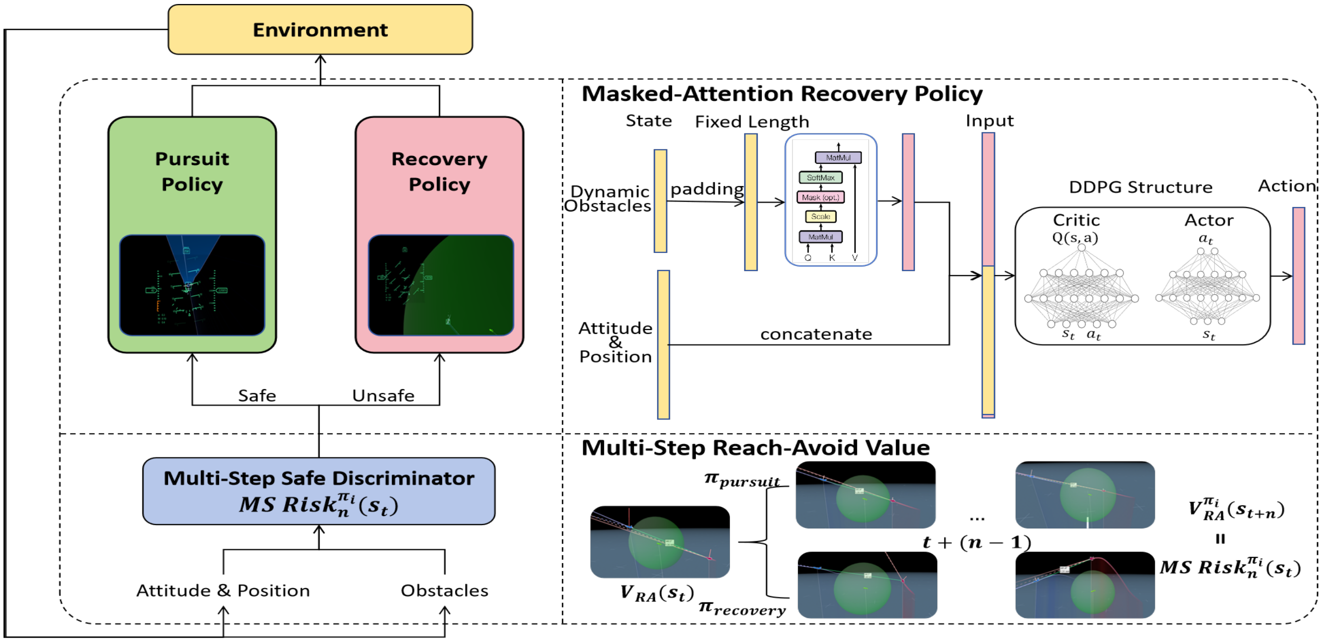


Figure 1: System Framework

obstacles or exceeding mission boundaries. It can be presented by the zero-sublevel set of a Lipschitz-continuous function $g(s)$ as:

$$s \in Failure\ Set \Leftrightarrow g(s){>}0. \tag{3}$$

On the contrary, the safe set is the complement set of the failure set, which means agent could remain safe all the time. It could be presented as followed.

$$s \in Safe\ Set \Leftrightarrow g(s) \leq 0. \tag{4}$$

(2) Target Set

Target Set represents the set of states coming from trajectory leading to the target. It can also be presented by the zero-sublevel set of a Lipschitz-continuous function $l(s)$ as

$$s \in Target\ Set \Leftrightarrow l(s) \leq 0. \tag{5}$$

(3) Reach-Avoid Set

The reach-avoid value quantifies the probability that the UAV, starting from state s, can reach the target set while avoiding the Failure Set:

$$\begin{aligned} RA^{\pi}(Target; Safe) := \{ & s_t \in Safe\ Set \mid g^{\pi}_{s_t}(0) = s_t, \\ & \forall t' \in [0, T-t], g^{\pi}_{s_t}(t') \notin Failure\ Set\}. \end{aligned} \tag{6}$$

## 4 Methodology

The overview of our system framework is shown in Figure 1 . The basic structure is recovery reinforcement learning architecture, which contains one safe critic discriminator and two policy structure. The safe critic is used to judge the safe condition of the near states, with a multi-step reach avoid value learned by the agent. The two policy are responsible for selecting actions for maximum reward or keeping in safe states.

When the agent get observation from the environment, the safety discriminator determine the current state (safe or unsafe) using a state-safe value $MSRisk_{\pi_i}(s_t)$. Based on the value, the UAV selects either the target policy (for task completion) or the recovery policy (for avoiding entering a dangerous state) and interacts with the environment. After the interaction, the UAV receives new observations, and the process repeats, forming a closed-loop control that ensures both safety and task efficiency in dynamic environments.

### 4.1 Extendable Multi-Step Reach-Avoid Theory

Reach avoid theory could provide a safe guidance for a deterministic Markov Decision Process (MDP). However, the current theory just provide a safe guidance for single policy, which is not suitable for recovery reinforcement learning.

This part firstly provide an illustration of safe guidance for recovery reinforcement learning, which could be extended to multi-policy reinforcement learning as well. Based on the theory, we further offer a multi-step method of reach-avoid value, which could significantly improve the safety in high-speed combat games.

**Extendable Reach-Avoid Theory**

As we mentioned in 3.2, the key point of reach avoid theory is to find a set in which the state could lead agents get to the target without offending constraints. In continuous environment, finding all states is impossible. So researchers [Hsu *et al.*, 2021] defined reach-avoid value in a finite time T for a deterministic control policy $\pi : S \rightarrow A$ :

$$V^{\pi}(s,T) = \min_{t\in[0\ldots T]} \max \left\{ l(s_t), \max_{\tau\in[0\ldots T]} g(s_{\tau}) \right\}, \tag{7}$$

where $s$ represents state, $t$ and $\tau$ represents separate moment from 0 to T. Besides, the $g(s)$ is failure set function in Equation 3 and the $l(s)$ is target set function in Equation 5.

From the equation above, if the reach-avoid value $V^{\pi}(s,T) \geq 0$ , it means that within the time 0 to T, there exists one state when $g(s) \geq 0$, which indicates that the agent violate the constraints. In contrast, if $V^{\pi}(s,T) \leq 0$ through out the whole T, it means that the agent get to the target without any violation of danger within the whole process, thus provide a zero-constraint-violation theoretical guidance.

To extend this theory to recovery reinforcement learning, we use $V^{\pi_i}(s,T_i)$ to present multi-strategy safe value.

As $V^{\pi}(s)$ shows, the reach-avoid value is associated with the strategy that agent uses. In recovery reinforcement learning, agent would switch policy to avoid get into dangerous states, which means that we can not use Equation 7 to compute data-based reach-avoid value. In other words, the reach avoid value of the safe policy $V^{\pi_{safe}}(s,T)$ and recovery policy $V^{\pi_{recovery}}(s,T)$ are not equal, which could be presented as:

$$V^{\pi_i}(s,T) \neq V^{\pi_j}(s,T),\ i \neq j. \tag{8}$$

To provide zero-constraints-violation guidance for recovery reinforcement learning, we extend the reach-avoid theory from single deterministic strategy to more than two deterministic strategies and demonstrate safety of the whole process. The concrete demonstration could be found in Appendix A. And the extended reach avoid theory could be represented as:

$$\begin{aligned} RA^{\pi_i}(Target; Safe) := \{s_t \in Safe\ Set \mid g^{\pi_i}_{s_t}(0) = s_t, \\ \forall t' \in [t_i+1, t_{i+1}-t], g^{\pi_i}_{s_t}(t') \notin Failure\ Set\}, \end{aligned} \tag{9}$$

where $g^{\pi_i}(s)$ is the i-th safe function in Equation 3, and $t_i$ is the moment when $\pi_i$ is switching to be used. Each $g(s)$ is a lipschitz-continuous function and the value of $g^i(s)$ and $g^{i+1}(s)$ at the switching point should be equal.

Having extended the reach-avoid theory to multi-strategy scenarios, we could farther get the Bellman formulation of the value $V^{\pi_i}(s,T_i)$. In Equation 10, number i represents the different policy. $l(s)$ is a lipschitz-continuous function, and $g^{\pi_i}(s)$ satisfy the constraint above.

$$V^{\pi_i}(s,T_i) = \min_{t \in [0 \ldots T_i]} \max\left\{ l(s_t), \max_{\tau \in [0 \ldots T_i]} g^{\pi_i}(s_{\tau_i}) \right\}. \tag{10}$$

### Multi-Step Reach-Avoid Value

Now that different reach avoid value $V^{\pi_i}(s,T_i)$ is connected with its own policy, the training process should be separate as well. Firstly, different tuples should be divided into their own experience pool for training. The discounted reach-avoid Bellman formulation suitable for learning with temporal difference learning could be written below.

$$\begin{aligned} V^{\pi_i}(s) &= (1-\gamma)\max\{l(s), g(s)\} \\ &+ \gamma \max\left\{ \min\left\{ l(s), \min_{a \in \mathcal{A}} V^{\pi_i}(s') \right\}, g(s) \right\}, \end{aligned} \tag{11}$$

where s' is the next state produced by the MDP upon taking action a from state s.

However, as mentioned in introduction, caused by physical limitations of high-speed vehicles, high-speed UAVs require earlier obstacle avoidance decisions to prevent collisions due to delayed reactions. Traditional usage of the reach-avoid values usually sets a threshold between the safe set and the failure set, which comes from experience and is hard to quantify. Besides, it is always too close to the failure set for high-speed aircrafts to avoid obstacles and too late to take recovery policy, probably leading to a crash for the aircrafts.

Considering the inertial and limited maneuverability, we introduce a multi-step reach-avoid value calculation method. As shown above, $V^{\pi_i}(s,T_i)$ have relations with states and policy, thus we use a function Multi-Step Risk(MS Risk) to estimate the reach-avoid value in several steps. If $MS\ Risk^{\pi_i}_n(s_t)$ qualifies the unsafe threshold, which means that in n steps later the agent would get into a dangerous state without switching policy, the agent will take the recovery policy to avoid this danger. The connection between$MS\ Risk^{\pi_i}_n(s_t)$ and $V^{\pi_i}(s_{t+n})$ could be demonstrated in appendix B.

$$MS\ Risk^{\pi_i}_n(s_t) = V^{\pi_i}(s_{t+n}). \tag{12}$$

As for the concrete number of n, it is a trade-off between accuracy and computational resource.

## 4.2 Masked-Attention Recovery Strategy

In high-speed UAV adversarial games, the spatiotemporal distribution of obstacles is highly random and complex. Spatially, obstacles may appear sparsely or densely within certain area. Temporally, some obstacles may currently be within the observation space but disappear after the UAV passes, while others may enter the observation space as the UAV moves closer, leading to dynamic input of policy network. This spatiotemporal uncertainty poses significant challenges to the UAV's recovery strategy.

To address the spatiotemporal uncertainty, we propose an masked-attention recovery strategy. This strategy tackles the challenges in two aspects:

### Handling Temporal Dynamic Observation

Since the number of obstacles in the observation space varies over time, traditional fixed-dimensional input methods struggle to handle such variable-length inputs. To address this, we use a padding-mask method to convert variable-length inputs into fixed-length inputs. The specific steps are as follows:

Padding and Masking: Let the maximum number of obstacles be $N_{max}$. For each time step, if the current number of obstacles $n{<}N_{max}$, the remaining positions are padded with zero vectors, and a binary mask M is generated

$$M_i = \begin{cases} 1, & \text{if the } i\text{-th obstacle exists,} \\ 0, & \text{if the } i\text{-th obstacle is a padding value.} \end{cases} \tag{13}$$

Finishing masking, a shared neural network $f(x)$ encodes each obstacle feature into a feature embedding $h_i = f(x_i)$.

$$\alpha_i = \frac{\exp(e_i) \cdot M_i}{\sum_{j=1}^{N} \exp(e_j) \cdot M_j}. \tag{14}$$

A learnable query vector q computes the attention weights for each feature embedding $e_i = q \top h_i$, and the mask M is used to ignore padding values.

The weighted feature embeddings are aggregated into a fixed-length global feature representation z.This approach ensures that the output z has a consistent dimension regardless of the number of input obstacles n, enabling direct input to the subsequent reinforcement learning network.

$$z = \sum_{i=1}^{N} \alpha_i h_i. \tag{15}$$

**Handling Spatial Distribution Uncertainty**

In environments with multiple nearby obstacles, the UAV needs to prioritize avoidance based on the urgency of each obstacle. We use the attention mechanism to assign an urgency weight to each obstacle, enabling targeted avoidance. With the help of the attention framework, the agent could avoid the obstacles based on the weighted urgency.

To improve training effectiveness, we adopt a curriculum learning approach motivated by [Yu *et al.*, 2023], gradually transitioning from simple to complex scenarios. Specifically, training starts with sparse obstacles and low-speed scenarios, and the density of obstacles and UAV speed are gradually increased. This progressive training method enables the network to master obstacle avoidance capabilities from simple to complex scenarios.

# 5 Experience

## 5.1 Training Sets

The training of our proposed method consists of three parts: Pursuit Network, Multi-Step Safety Discriminator, and Recovery Network. The details of each part are as follows.

**Pursuit Policy Reward**

The Pursuit Network is designed to enable the UAV to pursue adversarial targets effectively. It is implemented using the Deep Deterministic Policy Gradient (DDPG) algorithm, which combines Actor and Critic networks to learn a policy that maximizes cumulative rewards.

The reward function is carefully designed to guide the UAV toward the target, consisting of 3 components, including distance reward, angle reward and boundary reward which references to paper [Zhao *et al.*, 2024a]. Distance reward encourages the agent get close to the opposite agent, which take use of the matrix game theory. Angle reward consists the angle between own UAV and opposite UAV, aiming to make the UAV heading to the opposite UAV. Boundary reward prevents the UAV from flying out of bounds by imposing a penalty when it flys too high or too low. The whole reward could be presented as followed:

$$R = \omega_1 R_{distance} + \omega_2 R_{angle} + \omega_3 R_{boundary}, \tag{16}$$

where $\omega_1, \omega_2$, and $\omega_3$ are weight coefficients. This reward design ensures that the UAV could have great performance at gaining advantages against the opposite agent in the game.

**Safety Discriminator Loss**

The Multi-Step Safety Discriminator is designed to assess the risk of state-action pairs under the Pursuit policy, enabling the UAV to identify and avoid unsafe states. Motivated by [Ren *et al.*, 2023], it is implemented as a multi-layer perceptron (MLP) and trained using supervised learning. The loss function is defined as the mean squared error between the predicted and true risk values:

$$L_{discriminator} = \frac{1}{N} \sum_{i=1}^{N} (V_{RA}(s_{i+n}) - MS\hat{R}isk_{RA}(s_i))^2, \tag{17}$$

where $V_{RA}(s_{i+n})$ is the ground truth risk value in n steps later, and $MS\hat{R}isk_{RA}(s_i)$ is the predicted value.

According to the equation 10, $V^{\pi_i}(s)$ is dependent on the safe set function $g(s)$ and target set function $l(s)$. In the high-speed UAV combat game, we first provide failure set functions $g(s)$ $V^{\pi_i}(s)$:

$$g(s) = \begin{cases} 0 \quad , & d_{obstacle} \geq d_{safe}, \\ -\frac{(d_{safe} - d_{obstacle})}{d_{safe}}, & d_{obstacle} < d_{safe}, \end{cases} \tag{18}$$

where d is the distance between uav and the obstacle, and $d_{safe}$ represents the safe boundary of agent, which means that if $d < d_{safe}$, the agent get into the failure set. We set $d_{safe}$ as 1000m in the recovery task. Meanwhile, the function $l(s)$ of target set:

$$l(s) = \tanh \log \frac{d_{target}}{d_{success}}, \tag{19}$$

where $d_{target}$ represents the distance between the agent and target agent, and $d_{success}$ means that the agent is close enough to the target and get the task accomplished. In the training process, we set $d_{success}$ as 300m.

**Recovery Policy Reward**

The recovery network is designed to enable the UAV to recovery from unsafe states. Similar to [Qu *et al.*, 2024], the training objective is to minimize risk, with the loss function defined as the expected cumulative risk cost over time:

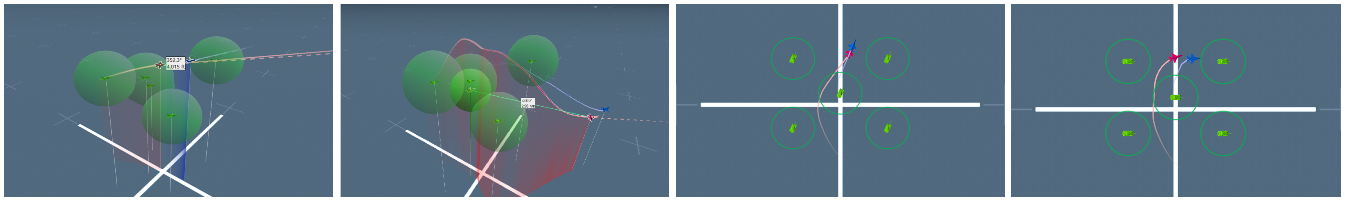
$$R_{recovery} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t)\right], \tag{20}$$

where $C(s_t, a_t)$ is the risk cost function, and $\gamma$ is the discount factor. To satisfy the reach avoid guidance [Kim *et al.*, 2023], we use $g(s)$ from Equation 18 as the $C(s_t, a_t)$, and $\gamma$ as 0.95.

## 5.2 Baselines

For the experimental settings, we consider four algorithm as the baselines below:

1. Our MSMAS-RL Method: Our recovery RL with multi-step safety discriminator and masked-attention recovery policy.
2. Pursuit Only: The pursuit network for target chasing.
3. Negative Rewards: Incorporates safety constraints as negative rewards into pursuit policy.[Qu *et al.*, 2023]
4. Primal-Dual Method: DDPG with constraints using Lagrangain optimization. [Brunke *et al.*, 2022]



(a) Pursuit-Only (b) MSMAR-RL (c) Pursuit-Only (d) MSMAR-RL

Figure 2: Performance Test. (a) The agent uses only pursuit policy. It could tail the target uav without avoiding obstacles. (b) The agent uses our MSMAR-RL method and carefully avoids all the obstacles while chasing the target uav. (c) The agent uses pursuit policy only. It could tail the target uav without avoiding the cylindrical obstacles. (d) The agent uses our MSMAR-RL method and greatly avoids all the cylindrical obstacles while chasing the target uav.

## 5.3 Evaluation Metrics

In high-speed UAV adversarial games, the objectives of the UAV include the following two aspects:

1) Completing the Adversarial Task: Successfully locking onto the adversarial UAV and completing the mission.

2) Obstacle Avoidance and Safe Flight: During the pursuit, the UAV must avoid obstacles to prevent collisions.

Based on these objectives, we designed the following three evaluation metrics to comprehensively assess the performance of the agent. More details could be found in Appendix C.

1. Success Rate: The task success rate measures the percentage of missions in which the UAV successfully locks onto the adversarial UAV and completes the task.

$$Rate_{success} = \frac{num_{success}}{num_{total}}. \tag{21}$$

2. Safety Rate: The safety rate measures the average probability of violating safety rules (colliding with obstacles) while completing the mission.

$$Rate_{safe} = 1 - \frac{NumStep_{danger}}{NumStep_{total}}. \tag{22}$$

Besides, to evaluate the zero-constraint-violation ability of our algorithm, we introduce a zero-danger-rate:

$$Rate_{zero\ danger} = \frac{NumEpisode_{safe}}{NumEpisode_{total}}, \tag{23}$$

where the episode means single whole game.

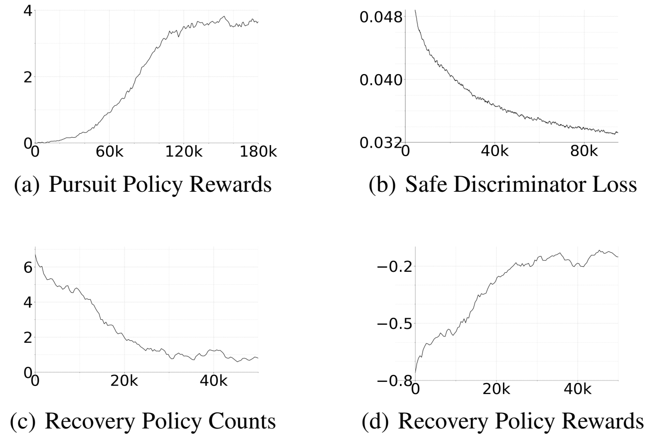
## 5.4 Training Analysis

The training process was divided into three stages:

1. Pursuit Task: The maximum steps per episode were set to 200, with a total training duration of 15,000 steps.

2. Safety Discriminator Training: The training duration was 10,000 steps.

3. Recovery Policy Training: The policy was trained until it converged to a low constraint violation.



(a) Pursuit Policy Rewards (b) Safe Discriminator Loss

(c) Recovery Policy Counts (d) Recovery Policy Rewards

Figure 3: Training Record. Horizontal axis represents training steps.

As showed in Figure 3(a), the pursuit task network which aims to fetch the opposite uav, gradually converged to a high reward. In Table 2, the pursuit policy could get a high success rate. The second training stage in Figure 3(b), the safe discriminator experienced a decrease of loss, showing a great performance in predicting dangerous value in several steps later. Within the training stage of recovery policy in Figure 3(c), 3(d), the reward gradually climbed up and danger counts decreased, showing the safe ability of agent.

## 5.5 Testing Analysis

Firstly, we test the target-pursuit obstacle-avoidance ability in a random environment compared to the training environment. On the one hand, we use random initial seed to put our uav, opposite uav and obstacles in a random place and attitude within a certain range. On the other hand, when comparing the ability of baselines algorithm, we set the same random seed to let the experience repeatable.

In Figure 2(a), the agent use the pursuit-only policy to complete the task. The agent could quickly get to the target without avoiding the obstacles, which could lead to horrible crash. Compared to this method, our MSMAR-RL algorithm could avoid the obstacles all time long and finally catch the opposite agent showing in Figure 2(b). Figure 2(c) and 2(d) use cylindrical obstacles to simulate buildings.

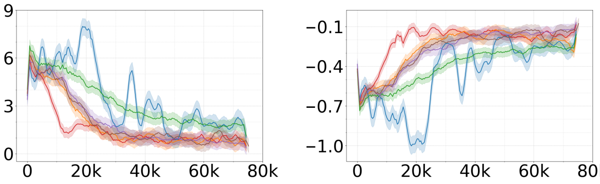
| Multi-step Module | Attention Module | Success Rate | Safe Rate | Zero-Danger-Rate |
|---|---|---|---|---|
| ✓ | ✓ | **61.91%** | **99.02%** | 80.95% |
| ✓ | | 57.14% | 98.93% | **85.71%** |
| | ✓ | 47.62% | 97.56% | 38.10% |
| | | 42.86% | 97.94% | 33.33% |

Table 1: Ablation Test

The performance of baselines algorithm are shown in Table 2. Our method gets the highest success rate and safe rate among the baselines at the cost of success time. The pursuit-only method is good at chasing the target, but always offend the constraints. Negative reward method put too much attention on avoiding obstacles, leading a totally loss of chasing ability. Traditional primal dual optimization method performs well in safe target, but gets a decrease in success rate.

| Method | Success Rate | Safe Rate | Success Time |
|---|---|---|---|
| MSMAR-RL | **61.91%** | **99.02%** | 87s |
| Pursuit-Only | 61.91% | 86.33% | **49s** |
| Negative Reward | 0 | 100% | ∞ |
| Primal Dual | 47.62% | 97.85% | 124s |

Table 2: Baselines Comparison



(a) Dangerous Counts

(b) Recovery Rewards

Figure 4: Ablation Experience Curves. In the figure, red curve is our MSMAR-RL method. Orange curve is the ablation method without attention module. Blue curve is the ablation method without multi-steps discriminator module. Green curve is the method without these two modules. Purple and brown curves use LSTM and RNN to replace attention module. Horizontal axis represents training steps.

## 5.6 Ablation Studies

The chapter would set experience to evaluate the function of multi-step safe discriminator and the attention part.

### Multi-Step Safe Discriminator

In this experiment, we removed the multi-step safety estimation module and used only a single-step safety discriminator.

In Table 1 above, with the similar success rate, our MSMAR-RL method have a much better performance in zero-constraint-violation rate than the method without multi-step safe discriminator, which is critically essential for flight safe scenario. And as shown in Figure 4, method with multi-step module could lead to a faster and better performance in training process. Besides, MSMAR-RL performes a bit worse in zero-constraint violation than that without attention module, this may be a cost of dynamic input of attention modules. These results demonstrate that the multi-step safety discriminator, through multi-step prediction of potential hazardous states, enables the UAV to take avoidance actions in advance, thereby significantly reducing constraints violation.

### Attention-based Mechanism

In the experiment without the Attention mechanism, we removed the Attention mechanism from the recovery network and used fixed-dimensional inputs.

Compared to the method without attention, MSMAR-RL has higher convergence rate than that without attention mechanism(red and orange curve). However, without a correct safe discriminator, the attention module would lead to a concussion in training process(blue curve in Figure 4). This may occur because that the attention module provide better learning ability to the agent, but exploration would lead to danger caused by the maneuverability limitation of the agent. With the multi-steps module, the concussion is much less, which demonstrates the greatness of the multi-step safe discriminator. Besides, compared with Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) (brown and purple curve), our attention mechanism converges faster, demonstrating the rapid adaptability to dangerous situations.

## 6 Conclusion

This paper proposes a zero-constraint-violation recovery reinforcement learning framework for high-speed UAVs in dynamic adversarial environments, addressing the challenge of balancing safety and task efficiency under strict constraints. The framework introduces three key innovations: (1) a dual-policy approach that separates target pursuit and obstacle avoidance, simplifying training in multi-objective scenarios; (2) an extendable multi-step reach-avoid strategy, providing safety guarantees and enabling early danger detection; and (3) an attention-based recovery strategy, which enhances adaptability to dynamic obstacles and improves training efficiency. Experimental results demonstrate the framework's effectiveness in achieving robust and safe decision-making for high-speed UAVs in complex gaming environments, surpassing existing methods in task completion rate, safety, and efficiency.

Future research directions include extending the framework to more complex scenarios. Additionally, exploring multi-UAV safe collaboration and recovery strategies will be crucial for advancing the field. Further optimization of computational efficiency and the development of adaptive learning mechanisms will also be essential to enhance real-time performance and adaptability in unpredictable environments. This study lays the foundation for safer and more efficient UAV control strategies in adversarial settings.

## Ethical Statement

This work is strictly for civilian applications, such as unauthorized drone interception in no-fly zones or public safety scenarios. The proposed method is not designed for, and should not be applied to, any military or offensive purposes.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2022ZD0160402), the National Natural Science Foundation of China (No. 62201471), and Aeronautical Science Fund (No. ASFC-20240001053006).

## Contribution Statement

Yang Zhao and Wenzhe Zhao make equal contributions to this work and are co-first authors. Yang Zhao is the corresponding author of this work.

## References

[Akshay *et al.*, 2024] S Akshay, Krishnendu Chatterjee, Tobias Meggendorfer, and ore Žikelić. Certified policy verification and synthesis for mdps under distributional reach-avoidance properties. *arXiv preprint arXiv:2405.04015*, 2024.

[Brunke *et al.*, 2022] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

[Ganai *et al.*, 2024] Milan Ganai, Sicun Gao, and Sylvia Herbert. Hamilton-jacobi reachability in reinforcement learning: A survey. *IEEE Open Journal of Control Systems*, 2024.

[Guo *et al.*, 2022] Junxiao Guo, Zihan Wang, Jun Lan, Bingchen Dong, Ran Li, Qiming Yang, and Jiandong Zhang. Maneuver decision of uav in air combat based on deterministic policy gradient. In *2022 IEEE 17th International Conference on Control & Automation (ICCA)*, pages 243–248. IEEE, 2022.

[He *et al.*, 2024] Tairan He, Chong Zhang, Wenli Xiao, Guanqi He, Changliu Liu, and Guanya Shi. Agile but safe: Learning collision-free high-speed legged locomotion. *arXiv preprint arXiv:2401.17583*, 2024.

[Hsu *et al.*, 2021] Kai Chieh Hsu, Vicenç Rubies-Royo, Claire J Tomlin, and Jaime F Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *17th Robotics: Science and Systems, RSS 2021*. MIT Press Journals, 2021.

[Kim *et al.*, 2023] Chan Kim, Jaekyung Cho, Christophe Bobda, Seung-Woo Seo, and Seong-Woo Kim. Sero: self-supervised reinforcement learning for recovery from out-of-distribution situations. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3884–3892, 2023.

[Li *et al.*, 2023] Yan Li, Xuejun Zhang, Yuanjun Zhu, and Ziang Gao. A uav path planning method in three-dimensional urban airspace based on safe reinforcement learning. In *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, pages 1–7. IEEE, 2023.

[Liang *et al.*, 2018] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

[Pope *et al.*, 2022] Adrian P Pope, Jaime S Ide, Daria Mićović, Henry Diaz, Jason C Twedt, Kevin Alcedo, Thayne T Walker, David Rosenbluth, Lee Ritholtz, and Daniel Javorsek. Hierarchical reinforcement learning for air combat at darpa's alphadogfight trials. *IEEE Transactions on Artificial Intelligence*, 4(6):1371–1385, 2022.

[Qu *et al.*, 2023] Xiuqing Qu, Wenhao Gan, Dalei Song, and Liqin Zhou. Pursuit-evasion game strategy of usv based on deep reinforcement learning in complex multi-obstacle environment. *Ocean Engineering*, 273:114016, 2023.

[Qu *et al.*, 2024] Yang Qu, Jinming Ma, and Feng Wu. Safety constrained multi-agent reinforcement learning for active voltage control. *arXiv preprint arXiv:2405.08443*, 2024.

[Ren *et al.*, 2023] Dejin Ren, Wanli Lu, Jidong Lv, Lijun Zhang, and Bai Xue. Model predictive control with reach-avoid analysis. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5437–5445, 2023.

[Thananjeyan *et al.*, 2021] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.

[Vinod *et al.*, 2022] Abraham P Vinod, Sleiman Safaoui, Ankush Chakrabarty, Rien Quirynen, Nobuyuki Yoshikawa, and Stefano Di Cairano. Safe multi-agent motion planning via filtered reinforcement learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7270–7276. IEEE, 2022.

[Vinyals *et al.*, 2019] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

[Wu *et al.*, 2024] Jingda Wu, Chao Huang, Hailong Huang, Chen Lv, Yuntong Wang, and Fei-Yue Wang. Recent advances in reinforcement learning-based autonomous driving behavior planning: A survey. *Transportation Research Part C: Emerging Technologies*, 164:104654, 2024.

[Ying *et al.*, 2022] Chengyang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. *arXiv preprint arXiv:2206.04436*, 2022.

[Yu *et al.*, 2023] Shuodian Yu, Junqi Jin, Li Ma, Xiaofeng Gao, Xiaopeng Wu, Haiyang Xu, and Jian Xu. Curriculum multi-level learning for imbalanced live-stream recommendation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2406–2414, 2023.

[Yue *et al.*, 2023] Longfei Yue, Rennong Yang, Ying Zhang, and Jialiang Zuo. Research on reinforcement learning-based safe decision-making methodology for multiple unmanned aerial vehicles. *Frontiers in Neurorobotics*, 16:1105480, 2023.

[Zhao *et al.*, 2023] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2023.

[Zhao *et al.*, 2024a] Yang Zhao, Zidong Nie, Kangsheng Dong, Qinghua Huang, and Xuelong Li. Autonomous decision making for uav cooperative pursuit-evasion game with reinforcement learning. *arXiv preprint arXiv:2411.02983*, 2024.

[Zhao *et al.*, 2024b] Yaya Zhao, Kaiqi Zhao, Zhiqian Chen, Yuanyuan Zhang, Yalei Du, and Xiaoling Lu. A graph-based representation framework for trajectory recovery via spatiotemporal interval-informed seq2seq. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 2588–2597, 2024.