Explainable Deep Reinforcement Learning for Multi-UAV Collision Avoidance in a Dynamic Airspace

Ruth R. Bahre[1,*] and Radu F. Babiceanu[1]
[*] *Corresponding Author: ruth.bahre@wmich.edu*

[1]*Department of Electrical and Computer Engineering*
Western Michigan University
Kalamazoo, MI, USA

***Abstract*—The rapid growth of Advanced Air Mobility (AAM) operations is creating unprecedented demands on low-altitude airspace. Unlike traditional aviation where centralized air traffic control manages flight paths, Uncrewed Aerial Vehicle (UAV) operations require autonomous, real-time collision avoidance systems that can safely navigate dynamic scenarios without prior coordination between operators. This research develops a scalable Explainable AI framework for real-time multi-UAV collision avoidance using deep reinforcement learning in a 3D simulation environment. Three agents (a learning Hero UAV, a companion Friend UAV, and an Adversary UAV) are trained under a three-phase Proximal Policy Optimization curriculum. Phase 1 establishes a successful goal-seeking baseline; Phase 2 trains a credible adversary achieving a 75.4% collision rate against the undefended Hero; Phase 3 fine-tunes the Hero against the trained Adversary under two scenarios: a single-attack and a persistent three-pass threat. In addition to the adversarial interaction, the Hero must maintain safe distance from the cooperative Friend teammate to not induce secondary conflicts. Post-training, ten interpretability methods, including gradient saliency, policy entropy analysis, Kullback-Leibler divergence from baseline, and Spearman-ranked feature attribution consistency, were used to quantify Hero's decision transparency. Results demonstrate a 62% improvement in goal-reach rate over the undefended baseline (90.6% and 86.6% for Scenarios 1 and 2 respectively), sub-millisecond inference latency (mean of 0.056–0.064 ms), and near-perfect attribution consistency (Spearman correlation coefficient larger than 0.995), validating both safety performance and explainability for autonomous UAV operations.**

Keywords — Unmanned Air Vehicles, Collision Avoidance, Deep Reinforcement Learning, Explainable AI

I. Introduction

Uncrewed Aerial Vehicles (UAVs) are increasingly deployed in complex, dynamic environments where multiple agents must operate safely within shared low-altitude airspace. As Beyond Visual Line of Sight (BVLOS) operations expand under Advanced Air Mobility (AAM) frameworks, developing reliable real-time collision avoidance system for autonomous UAVs has become a national research priority [1]. Traffic densities in urban air corridors are projected to increase exponentially over the coming years. Companies operating package delivery drones, infrastructure inspection fleets, and emergency response AAMs will soon share the same airspace corridors [2].

Current air traffic operations approaches rely on either rigid rule-based separation logic that cannot adapt to complex multi-agent interactions, or black-box machine learning systems that lack the transparency required for regulatory certification and public trust, creating a bottleneck for scaling UAV operations in shared airspace. Traditional rule-based collision avoidance systems such as Traffic Alert and Collision Avoidance System (TCAS) were designed for cooperative, higher-altitude piloted aircraft and do not transfer well to the low-altitude, non-cooperative multi-UAV environment, where agents lack transponders, flight plans are unknown, and traffic densities are far higher [3]. Black-box Deep Reinforcement Learning (DRL) approaches achieve strong empirical performance but cannot provide the transparent reasoning required for regulatory certification under the Federal Aviation Administration Uncrewed Aircraft Systems Traffic Management (UTM) frameworks [4].

Explainable Artificial Intelligence (XAI) offers a pathway to bridge this gap by pairing high-performance machine learning decision-making with interpretable attribution of the features driving each action. This paper presents a scenario-based multi-UAV collision avoidance framework that combines Proximal Policy Optimization (PPO) Reinforcement Learning with integrated XAI analysis in a 3D simulation environment. The proposed framework supports decentralized detect-and-avoid decision-making for autonomous UAVs operating in shared airspace. This work may support future certification efforts for AI-based collision avoidance by showing that the learned policy is both effective and interpretable.

II. Related Work

A. Multi-UAV Collision Avoidance using DRL

Reference [5] presented a Deep Reinforcement Learning-based collision avoidance framework for UAVs operating in dynamic environments, demonstrating that DRL agents can learn robust avoidance maneuvers from raw sensor inputs without explicit path planning. The work established a strong empirical baseline for learning-based UAV separation but did not address multi-agent interactions or provide interpretability of the learned policy. Similarly, the research depicted in [6] applied a Dueling Double Deep Q-Network augmented with Self-Attention Models (D3QN-SAM) to single-UAV obstacle avoidance, using attention weight visualization on depth image patches as an interpretability mechanism. While their approach

demonstrated improved generalization to novel environments, the attention maps operate over visual features and do not translate to the structured situational variables. Employing the same approach, D3QN with attention-based RL, the work in [7] builds a simulation environment with different encounter and traffic density scenarios and was able to prove the improvement of the approach compared to other learning and non-learning implementations found in the literature.

B. Interpretable RL for Safety-Critical Systems

Reference [8] provides a comprehensive survey of interpretable reinforcement learning, distinguishing between intrinsic interpretability (white-box model architectures) and post-hoc explainability methods applied to trained black-box policies. The survey argues that DRL is not yet mature enough for high-stakes deployment in domains such as autonomous driving or aviation without interpretability guarantees. The paper also identifies gradient-based saliency and feature attribution consistency as among the most practically relevant post-hoc techniques. Another comprehensive survey that covers deep learning for air traffic management is given in [9]. The paper covers a decade of AI solutions in aviation traffic and identifies the types of explainable models to interpret the behavior behind the AI algorithm as well as detect potential errors and unwanted behavior. This is obtained through the use of descriptive, predictive, and prescriptive XAI solutions.

The need for understanding the reasoning behind black-box AI algorithms work is critical for their acceptance in the safety-critical systems domain. Air traffic operations can be trusted only when there is transparency of the entire development and deployment of statistical learning models, which includes data explainability, model explainability, post-hoc explainability, and assessment of explanations. For more information on XAI, the readers can refer to [10]; for information on PPO algorithms, the readers can refer to [11]; for approaches to interpreting model predictions, the readers can refer to [12]; while for specific RL research applied to UAV simulation environments, a good resource is found in [13].

The survey work in [8] and conceptual modeling in [10] directly motivates the XAI evaluation methodology adopted in the current paper. Rather than redesigning the policy architecture for intrinsic interpretability, we apply a suite of post-hoc methods to a high-performing PPO policy and quantify the stability and trustworthiness of the resulting attributions. This current paper extends the paradigm to the UAV collision avoidance domain and introduces formal metrics including Spearman attribution consistency and decision transparency score that operationalize the evaluation criteria for safety-critical deployment.

III. Modeling Methodology

The framework uses a three-dimensional simulation environment as a bounded airspace (x: −10 to 30 m, y: −10 to 30 m, z: 0 to 20 m) populated by three UAV agents with distinct roles. The Hero UAV is the learning agent, trained to navigate from initial position at the origin to a fixed goal. The Friend UAV is a companion UAV flying alongside Hero with a +3 m lateral offset. To prevent collision between the Hero and Friend UAVs, the framework uses formation-offset policy combined with reward shaping rather than inter-agent communication. This design allows the agents to operate independently while reducing the likelihood of secondary conflicts. The Adversary UAV is an independently trained non-cooperative threat whose main goal is to collide with Hero UAV. This three-agent design reflects a realistic scenario where a primary UAV is navigating a shared flight corridor alongside an allied UAV while contending with an uncoordinated or a conflicting interceptor.

A. Observation and Action Spaces

The Hero UAV observes a 36-dimensional state vector including its own position, velocity, and orientation; the adversary's relative position, velocity, bearing, time-to-collision (TTC), and risk score; the friend's relative state; goal-relative position; directional clearances in six spatial sectors; and binary threat flags. This observation space enables the Hero to reason about geometric risk without requiring any communication with external agents. The Adversary observes an 18-dimensional vector which includes the Hero's relative state, bearing, closure rate, and its own position and velocity. Both agents share a six-action discrete action space: forward-climb, forward-level, forward-descend, strafe-left, strafe-right, and backward, with a fixed move delta of 0.75 m per step.

B. Policy Architecture

All agents use Proximal Policy Optimization (PPO) with a shared multi-layer perceptron (MLP) architecture. The policy network is a three-layer MLP mapping observations to action logits through hidden layers of 256 and 128 units with softmax output, while the value network uses an identical architecture projecting to a scalar. Key hyperparameters are clip epsilon ε = 0.2, discount factor γ = 0.99, entropy bonus coefficient 0.05, and gradient norm clip of 0.5. These choices prioritize stable convergence in the non-stationary multi-agent training setting.

C. Three-Phase Training Curriculum

a) Phase 1 – Goal Seeking

Using Deep Reinforcement Learning, the Hero trains without any adversary present, learning a point-to-point navigation in the order of takeoff, cruise and landing. Hero begins at (0,0,0) and ascends to a cruising altitude of z = 10.5 m during steps 1-14. The cruise segment (steps 16–46) maintains this altitude as the Hero proceeds toward the goal, passing through (9.0,9.0,10.5) at step 31, which lies near the midpoint of the encounter region. After clearing this region, Hero descends and lands at (20.3,20.3,0) during steps 51–63. The Friend UAV mirrors this trajectory with a constant 3 m offset at all times, terminating at (20.3,23.3,0). This ensures that the Hero's navigation does not induce conflicts with cooperative agents.

The Hero UAV is equipped with a 360° camera (50 m range), LiDAR (100 m range), and noisy GPS to build a 36-dimensional observation space including V2V friend measurements. These sensors provide the situational awareness to maintain stable navigation.

The reward function uses dense shaping to guide the hero toward its goal through horizontal distance progress (+10 per meter), proximity incentives (1/(distance+1)), and a time penalty (-0.05 per step). Goal achievement yields +200 base reward plus a speed bonus (+5 remaining steps), while episodes

exceeding the 200-step limit incur a -20 penalty. During this phase, Hero and Friend were trained for 1,000 episodes at a learning rate of 3×10^{-4}.

b) Phase 2 – Adversary

In this phase, the Adversary UAV trains to collide with Hero UAV mid-cruise using reinforcement learning. While Hero is on its flight path towards its goal, Adversary starts at position (20, 0, 10) and must intercept Hero around the collision target zone at (10,10,10). The adversary receives a 18-dimensional observation space (position, velocity, relative position/velocity to hero, distance, bearing, closing speed, and time-to-collision), controlled 6-dimensional movement actions (toward hero with climb/level/descend, lateral left/right, or away), and is rewarded with a base +200 collision bonus (+100 extra if in cruise zone), dense shaping rewards for closing distance (±10 per meter), proximity bonuses (1/(distance+1)), and penalties for hero reaching goal (-50) or timeout (-20). Adversary was trained for 1,000 episodes at a learning rate of 3×10^{-4}. During this phase, both Hero and Friend UAVs were not in active DRL training.

c) Phase 3 – Collision Avoidance

This phase represents the foundational collision avoidance scenario where Hero learns to detect and evade Adversary while maintaining its primary objective of reaching the goal. Two different scenarios were analyzed: single attack and a persistent adversary attack.

Scenario 1: Single-Attack Adversary

Hero fine-tunes its training from Phase 1 against the Phase 2 trained Adversary. The adversary deactivates after one close pass within 4 m. Episode length is capped at 300 steps, with evasion detection triggering at 25 m range. The reward function adds: +500 for reaching the goal, −300 for collision, $+10\times\Delta$ horizontal distance per step, +0.5 survival reward per step, and separation shaping when the adversary is within 10 m. This phase was run for 7000 episodes with a lower learning rate of 1×10^{-4} for stable fine-tuning. The model showed effective training results starting at episode 2000 but plateaued afterwards. The training pipeline uses batched PPO updates every 10 episodes to ensure stable gradient calculations and to monitor key performance indicators like collision rate, evasion success rate, timeout rate, and minimum distance maintained to the adversary throughout each episode.

Scenario 2: Persistent Adversary

This scenario is similar to Scenario 1, except the adversary attacks repeatedly up to three times per episode. This persistent threat requires Hero to maintain sustained evasion capabilities rather than recovering from a single Adversary encounter. This scenario was also run for 7000 episodes with extended episode length of 600 steps. Additional tracking metrics were collected including the number of evasions per episode and the average number of Adversary passes survived.

Fig. 1 shows Scenario 1 (single-attack, blue) and Scenario 2 (persistent adversary, red) across 7,000 episodes. Hero goal-reach rate for both scenarios converges above the 70% target threshold, with Scenario 2 requiring approximately three times as many episodes to stabilize. Collision and evasion rates show rapid suppression of collisions as Phase 3 fine-tuning progresses, with Scenario 2 retaining a higher residual evasion rate due to repeated adversary repositioning. Average cumulative reward over the last 100 episodes confirms stable convergence for both policies, with Scenario 2 exhibiting a characteristic dip near episode 4,000 as the Hero adapts to the persistent repositioning mechanic. Closest approach distance trends upward in both scenarios and consistently remains above the 1.5 m collision radius (dashed line), demonstrating that the learned policy maintains safe separation throughout training.

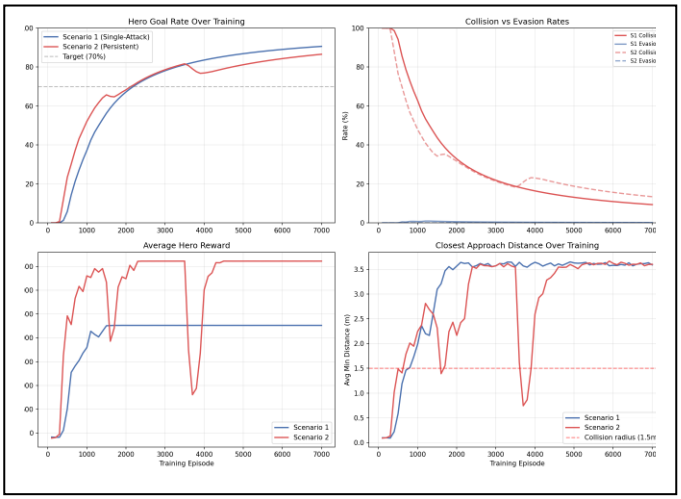

Fig. 1. Training progression curves for Scenario 1 and Scenario 2.

D. Explainability Analysis

A comprehensive explainability analysis was conducted on both Scenario 1 and Scenario 2 to provide interpretability into the Hero UAV's learned collision avoidance decisions. The primary XAI technique used is gradient-based saliency analysis. This method was selected because it provides rapid, model-agnostic feature importance rankings that directly measure how sensitive the policy's decisions are to the changes in each input feature. Saliency was particularly well suited to this project because the policy relies on structured safety-critical variables like relative position, time-to-collision, and boundary clearance.

Following the completion of the RL training, 200 evaluation episodes were run per scenario and analyzed with ten interpretability methods: (1) gradient saliency per feature, (2) grouped saliency by feature category, (3) action distribution versus adversary distance, (4) policy divergence (Kullback-Leibler) relative to Phase 1 baseline, (5) 3D trajectory visualization, (6) training curve analysis, (7) partial dependence for feature sensitivity, (8) policy entropy over the flight episode, (9) inference latency benchmarking over 1,000 trials with 100-trial warm-up, and (10) feature attribution consistency measured as Spearman rank correlation across five episode splits. A decision transparency score is defined as 1 – (mean entropy / max entropy), reflecting the ratio of confident (low-entropy) decisions.

IV. Simulation Results

A. Safety Performance

Table I summarizes goal achievement, collision rates and timeouts across all training phases and scenarios.

TABLE I. GOAL AND COLLISION RATE ACROSS TRAINING PHASES

Phase/Scenario	Goal Rate	Collision Rate	Time outs
Phase 1 (Hero + Friend)	97.1%	—	2.9%
Phase 2 (Adversary vs. Hero)	24.6%	75.4%	0%
Scenario 1 (Single attack)	90.6%	9.4%	0%
Scenario 2 (Persistent attacks)	86.6%	13.4%	0%

In Phase 1, Hero and Friend achieve a goal rate of 97.1% without any threat from the Adversary. The Phase 1 Hero and Friend reaching the goal, without adversary, is depicted in Fig. 2. Phase 2 confirms the Adversary is a credible threat achieving a 75.4% collision rate, creating a 62-percentage-point gap that Phase 3 must recover.

Both Phase 3 scenarios recover to 87–91% goal rate, approximately a 62% improvement over the undefended Hero in Phase 2. Notably, Scenario 2 achieves 86.6% despite the adversary repositioning up to three times per episode, demonstrating that the learned policy generalizes to persistent multi-pass threats. Zero timeout events across both Phase 3 scenarios indicate the Hero consistently commits to either reaching the goal or avoiding collision rather than stalling in uncertain states.

The 4% gap between Scenario 1 and Scenario 2 is attributed to the increased geometric complexity of the persistent repositioning. The adversary's teleportation to 60% of the Hero-to-goal path forces the Hero to solve a continuously refreshed interception problem rather than a single evasion. The observed degradation is modest and consistent with the added task difficulty.

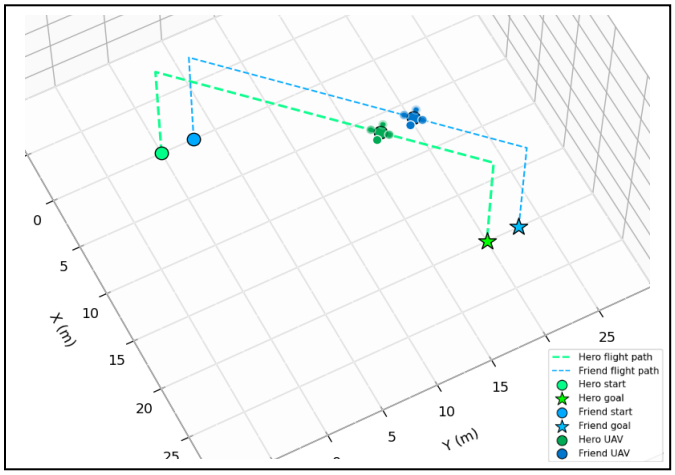

Fig. 2. Phase 1- Trained Hero and Friend reaching goal, with no Adversary.

Fig. 2 shows the Phase 1 training trajectory of the Hero and Friend UAVs in the 3D simulation environment without adversarial interference. They follow a smooth takeoff–cruise–landing path toward their respective goals, while maintaining a safe distance of 3m throughout the mission. During Phase 1 training, both UAVs achieve a goal-reaching success rate of 97.1%. The remaining 2.9% ended in timeout, where the UAVs exceed the maximum step limit without reaching their goals.

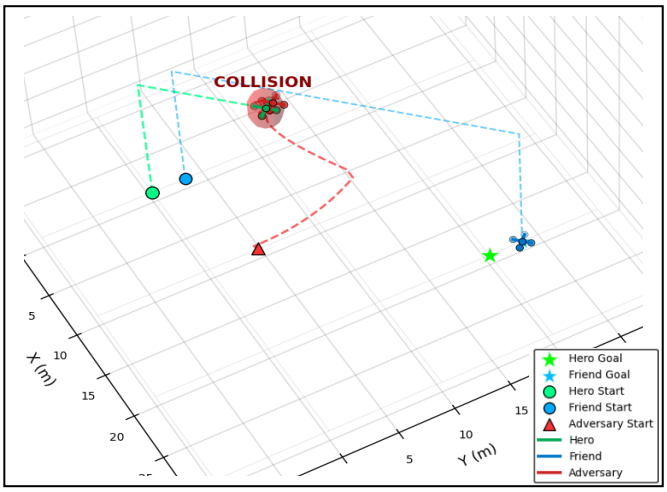

Fig. 3. Phase 2- Trained Adversary colliding with Hero.

Fig. 3 shows Phase 2 training where the adversary enters the picture and is trained to intercept with the Phase 1 Hero during its cruise segment, while the companion Friend UAV continues along its nominal path toward its assigned goal. The resulting trajectory illustrates a successful mid-course collision between the adversary and the Hero, showing that the learned adversarial policy can reliably disrupt the Hero before goal completion. Phase 2 training reduces the Hero's goal-reaching rate to 24.6%, while increasing the collision rate to 75.4%, confirming that the adversary has learned an effective interception strategy.

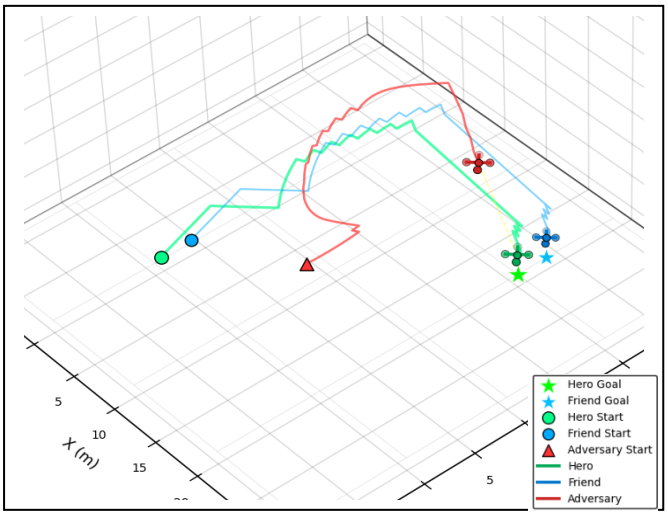

Fig. 4. Phase 3 Retrained Hero and Friend avoiding Adversary to reach goal.

Fig. 4 shows results of phase 3 under Scenario 2, in which the adversary persistently pursues Hero rather than disengaging after a single intercept attempt. The retrained Hero and Friend UAVs deviate laterally to evade the persistent Adversary while preserving safe boundary clearances throughout the maneuver. After the avoidance turn, both agents realign with the mission corridor and successfully reach their respective goals.

The added curvature in the flight paths reflects a modest increase in path length and travel time, indicating that the learned policy trades short-term efficiency for collision avoidance before resuming path to goal. The quantitative results show the retrained Hero achieving a goal-reaching rate of 90.6%

in Scenario 1 (single attack) and 86.6% in Scenario 2 (persistent attacks). Despite the more challenging condition shown in the figure, the Hero still maintains a high mission success rate, demonstrating that the learned avoidance policy remains effective under sustained adversarial pressure.

B. *Computational Efficiency*

Table II reports inference latency, confirming real-time capability. Both policies operate orders of magnitude faster than the minimum control rate required for UAV flight. Mean latency of 0.064 ms in Scenario 1 and 0.056 ms in Scenario 2 yields maximum achievable decision rates of 15,732 Hz and 17,951 Hz respectively, far exceeding the 10 Hz minimum for real-time UAV control. The lower variance in Scenario 2 (σ = 0.014 ms versus 0.046 ms in Scenario 1) suggests the persistent-adversary policy learned a more uniform decision pathway, reducing computational variance alongside decision uncertainty at confident timesteps.

TABLE II. COMPUTATIONAL EFFICIENCY METRICS

Metric	Scenario 1	Scenario 2
Mean inference latency	0.064 ms ± 0.046 ms	0.056 ms ± 0.014 ms
Max decision rate	15,732 Hz	17,951 Hz
Real-time capable (≥10 Hz)	Yes	Yes

C. *Explainability Analysis*

Table III shows the XAI evaluation results, which form the core contribution of this work toward regulatory-grade transparency. Feature attribution consistency near 1.0 in both scenarios (ρ = 0.998 and 0.995) indicates that gradient saliency rankings are stable and reproducible across independent episode subsets. This stability is a necessary condition for trustworthy explanations.

TABLE III. EXPLAINABILITY METRICS

Metric	Scenario 1	Scenario 2
Feature attribution consistency (Spearman ρ)	0.998 ± 0.002	0.995 ± 0.002
Decision transparency score	1.000	0.794
Confident decision ratio	100.0%	71.9%

The decision transparency score of 1.000 in Scenario 1 indicates that every action taken by the Hero in the single-attack scenario was a maximally confident, low-entropy decision. The policy had effectively learned a deterministic mapping for the threat. The reduction to 0.794 in Scenario 2, with 71.9% of decisions classified as confident, reflects appropriate epistemic uncertainty under the more complex persistent repositioning scenario. A transparency score of 1.0 in a high-uncertainty environment such as scenario 2 would indicate overconfident policy behavior. The results of Scenario 2 demonstrate that the system correctly modulates its certainty with task difficulty.

Gradient saliency analysis on Fig. 5 shows that the policy's most influential inputs were the allowed boundary clearance (cl_ left, cl_right, cl_down), goal position and threat cues such as adversary relative position, velocity, time-to-collision (TTC). Across both scenarios, the saliency rankings were highly stable across episode subsets, indicating that the learned policy relied on a consistent set of decision cues rather than unstable or overfit shortcuts. This suggests the controller balances safe maneuvering within the flight corridor with progress toward the goal, while incorporating adversary information as part of that broader avoidance strategy. The convergence of these attribution patterns across both single-attack and persistent-adversary conditions further confirms that the learned avoidance behavior is scenario-invariant, strengthening its credibility as a generalizable and trustworthy decision-making policy.

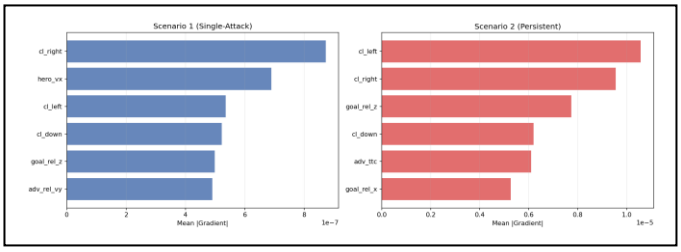

Fig. 5. Top Gradient-Saliency Features for both scenarios.

D. *Discussion*

The three-phase curriculum proved essential to avoiding catastrophic forgetting. Fine-tuning Phase 1 rather than training from scratch allowed Phase 3 to retain the goal-seeking behavior while grafting avoidance capability onto the existing trained model. The suppression of separation reward when the adversary lies on the Hero-to-goal path (Scenario 2) was similarly critical. Without it, preliminary experiments showed Hero converging to safe but mission-defeating retreat strategy.

The XAI analysis reveals that the learned policy is not only interpretable in a statistical sense, but also semantically valid. The dominant features are precisely the situational variables a human pilot or air traffic controller would prioritize in a detect-and-avoid judgment. This semantic validity strengthens the case for regulatory acceptance, as it demonstrates that the AI system reasons the problem in a manner consistent with established aviation safety logic.

V. CONCLUSION AND FUTURE WORK

This paper presented an Explainable Deep Reinforcement Learning framework for real-time multi-UAV collision avoidance in shared dynamic airspace. A three-phase PPO curriculum produced a Hero UAV achieving a 90.6% goal rate under single-attack conditions and 86.6% against a persistent three-pass adversary. Both policies operate at sub-millisecond inference latency with decision rates exceeding 15,000 Hz, confirming real-time viability. Post-hoc XAI analysis across ten interpretability methods demonstrated feature attribution consistency above Spearman ρ = 0.995 and decision transparency scores of 0.794–1.000. Gradient saliency consistently identified boundary clearances, adversary relative position, and time-to-collision as the dominant decision features. The semantic validity of these attributions, combined with their reproducibility across episode subsets, provides evidence that AI-based separation assurance may be feasible for certification under existing UTM frameworks.

Future work will address four extensions: (1) a continuous action space replacing the current six-action discrete scheme to enable finer-grained trajectory control; (2) larger UAV fleet scenarios beyond three agents to test scalability of the curriculum and XAI framework; (3) wind disturbance and sensor noise injection to evaluate policy robustness under realistic environmental conditions; and (4) simulation-to-real transfer validation using FPGA implementation or hardware UAV platforms benchmarked against the EuRoC MAV dataset. An adversary observation space ablation study is also planned to empirically quantify the impact of observation dimensionality on threat competence. This work establishes a foundation for AI-driven separation assurance in autonomous airspace, demonstrating that high-performance collision avoidance and interpretability are not mutually exclusive goals.

REFERENCES

[1] FAA, "Advanced Air Mobility (AAM) Ecosystem Working Groups," Federal Aviation Administration, Washington, DC, Tech. Rep., 2023.

[2] NASA, "Urban Air Mobility (UAM) Market Study," NASA CR-2018-219981, 2018.

[3] J. Tang, S. Lao, and Y. Wan, "Systematic review of collision-avoidance approaches for unmanned aerial vehicles," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4356–4367, Sep. 2022.

[4] FAA, "Safety Assurance Considerations for Uncrewed Aircraft Systems," Advisory Circular AC 90-109A, 2021.

[5] A. Sengupta, F. Bai, and N. Chakraborty, "Deep reinforcement learning-based collision avoidance in UAV environment," *IEEE Internet of Things J.*, vol. 9, no. 20, pp. 20150–20160, Oct. 2022.

[6] D.-G. Thomas, D. Olshanskyi, K. Krueger, T. Wongpiromsarn, and A. Jannesari, "Interpretable UAV collision avoidance using deep reinforcement learning," arXiv:2105.12254, 2021.

[7] M. Zhang, C. Yan, W. Dai, A. Xiang, and K. H. Low, "Tactical conflict resolution in urban airspace for unmanned aerial vehicles operations using attention-based deep reinforcement learning," Green Energy and Intelligent Transportation 2 (2023) 100107.

[8] C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao, and W. Liu, "A survey on interpretable reinforcement learning," *Mach. Learn.*, vol. 113, pp. 5847–5890, 2024.

[9] A. Degas, M. R. Islam, C. Hurter, S. Barua, H. Rahman, M. Poudel, D. Ruscio, M. U. Ahmed, S. Begum, M. A. Rahman, S. Bonelli, G. Cartocci, G. Di Flumeri, G. Borghini, F. Babiloni, and P. Arico, "A Survey on Artificial Intelligence (AI) and eXplainable AI in Air Traffic Management: Current Trends and Development with Future Research Trajectory," *Appl. Sci.* 2022, 12, 1295.

[10] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Caonfalonieri, R. Guidotti, J. Del Ser, N. Diaz-Rodriguez, and F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Information Fusion 99 (2023) 101805.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.

[13] C. B. Browne, J. Lim, A. Vasan, and A. T. Hayes, "PyFlyt: UAV simulation environments for reinforcement learning research," arXiv:2304.01645, 2023.