Multi-Agent Debate Framework for Comprehensive UAV Swarm Performance Evaluation

Shuai Hao, Haibin Duan and Chen Wei

***Abstract*—This paper presents a novel multi-agent debate framework for comprehensive UAV swarm performance evaluation. Unlike traditional single-perspective assessment methods, our approach leverages multiple specialized AI agents representing different domains of expertise to engage in structured evaluation dialogues, ensuring holistic coverage of critical performance dimensions. The framework integrates six innovative mechanisms: (1) hierarchical debate structure that dynamically routes issues by complexity to optimize computational efficiency, (2) adversarial-collaborative protocol with rotating red-blue teams to eliminate confirmation bias and echo chamber effects, (3) dynamic role rotation across debate rounds to prevent perspective lock-in, (4) evidence chain traceability ensuring all claims are backed by verifiable trajectory data, (5) multi-dimensional consensus modeling to quantify agreement patterns and identify disagreement types, and (6) meta-cognitive quality monitoring with real-time interventions to maintain debate rigor. We demonstrate the framework's effectiveness through comprehensive evaluation of real UAV swarm trajectory data, showing that the multi-agent debate approach produces more thorough, balanced, and actionable assessments compared to conventional single-metric or fixed-weight evaluation methods. Our contributions provide a principled methodology for UAV swarm evaluation that is adaptable to diverse mission types and operational environments.**

***Index Terms*—UAV swarm evaluation, multi-agent debate, collaborative AI, adversarial reasoning, consensus modeling, evidence-based assessment, structured argumentation.**

I. Introduction

The proliferation of Unmanned Aerial Vehicle (UAV) swarms in civilian and military applications has created an urgent need for comprehensive, objective, and multi-dimensional performance evaluation frameworks. UAV swarms present unique challenges in assessment due to their complex interdependencies, real-time coordination requirements, and multi-faceted performance criteria spanning flight control, inter-vehicle communication, formation stability, and safety protocols. Traditional single-perspective evaluation methods often fail to capture the nuanced interactions and trade-offs inherent in swarm systems, leading to incomplete or biased assessments that may overlook critical performance aspects.

Current UAV swarm evaluation approaches typically rely on isolated metrics or single-expert assessments, which can result in narrow evaluations that miss the holistic nature of swarm performance. The complexity of modern UAV operations—involving trajectory optimization, dynamic formation control, collision avoidance, energy efficiency, and mission adaptability—requires evaluation frameworks that can synthesize multiple expert perspectives and provide comprehensive, balanced assessments.

In this paper, we present a novel multi-agent collaborative assessment framework specifically designed for UAV swarm performance evaluation. Our approach leverages structured AI debate systems to simulate expert panel discussions, where multiple specialized AI agents representing different domains of expertise—Flight Control Specialists, Swarm Coordination Experts, and Safety Assessment Experts—engage in systematic evaluation dialogues. This methodology ensures comprehensive coverage of all critical performance dimensions while maintaining the depth of specialized knowledge required for professional-grade assessments.

The framework addresses three primary evaluation categories: (1) Flight Control Performance, encompassing trajectory smoothness, altitude stability, speed consistency, and energy efficiency; (2) Swarm Coordination Capabilities, including formation stability, communication quality, coordination delays, and task completion rates; and (3) Safety Assessment, covering collision avoidance effectiveness, emergency response protocols, and risk management procedures. Each category employs weighted scoring algorithms that reflect real-world operational priorities and industry standards, with default weights of 0.35, 0.40, and 0.25 respectively, reflecting the critical importance of coordination in swarm operations.

Our implementation demonstrates the framework's practical utility through comprehensive evaluation of simulated UAV swarm missions, generating detailed performance reports that include quantitative metrics, qualitative assessments, and actionable recommendations for system improvement. The multi-agent debate mechanism ensures that evaluations consider diverse perspectives, potential trade-offs, and edge cases that might be overlooked in traditional single-agent assessments.

This research contributes to the field by providing: (1) a structured methodology for comprehensive UAV swarm performance evaluation, (2) a multi-agent AI system capable of simulating expert panel assessments, and (3) empirical validation demonstrating improved evaluation comprehensiveness and accuracy compared to traditional approaches. The framework's modular design allows for adaptation to various UAV platforms, mission types, and operational environments, making it a versatile tool for researchers, manufacturers, and operators in the rapidly evolving UAV industry.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

II. Related Works

A. Multi-LLM debate

Multi-LLM debate (also referred to as multi-agent debate or collaborative LLM reasoning) aims to improve answer correctness, robustness, and interpretability through iterative argumentative exchange. Early motivation comes from AI safety via debate [1], which posits that adversarial dialogue can surface errors and supervise complex reasoning. Subsequent empirical studies show that, in QA and reasoning tasks, multi-round exchange of arguments and revisions can outperform single models under certain conditions [2], and that prompting for persuasiveness or complementary rationales can further enhance truthfulness and consistency. Related single-model techniques—chain-of-thought and self-consistency [3], [4]—stabilize reasoning by aggregating multiple trajectories and provide strong baselines that compose naturally within debate protocols.

At the protocol level, common procedures use round-robin/broadcast mechanisms: agents first produce independent answers, then critique one another in shared context across rounds before converging on a final decision [2], [5]. Extensions include heterogeneous roles and tool-augmented agents to diversify information sources and capabilities [6]; explicit control of dissent to avoid premature consensus or unproductive branching; and judge-based aggregation, where LLM evaluators or voting schemes select the final output. Broader multi-agent collaboration frameworks for planning, tool use, and task decomposition—while not inherently adversarial—inform debate systems through advances in coordination, memory, and orchestration.

Recent work also highlights limitations and failure modes. When agents are highly homogeneous or share correlated biases, debate can form "echo chambers," amplifying hallucinations and converging on incorrect answers. These issues are closely tied to majority dominance and correlated errors: training correlations induce shared misconceptions that become reinforced over rounds, drowning out minority—but correct—signals.

Overall, the field has progressed from initial demonstrations that debate can enhance reasoning to systematic exploration of protocol design, judging and aggregation, and multi-agent orchestration.

B. UAV Swarm Performance Metrics

Traditional evaluation of UAVs has focused on the performance of individual agents. Metrics for flight control, such as trajectory tracking error, altitude hold accuracy, and energy consumption, are well-established [7], [8]. For swarms, these metrics are extended to include collective behaviors [9]. Formation stability, which measures the swarm's ability to maintain a desired geometric shape, is a common metric. Communication quality, often measured by packet delivery rate and latency, is critical for coordinated tasks. Task completion rate and time are also key indicators of overall swarm effectiveness. However, many studies focus on a limited subset of these metrics, lacking a holistic view.

C. Evaluation Frameworks for Swarm Systems

Several frameworks have been proposed for evaluating swarm robotics systems. These often involve standardized testbeds, simulation environments [10], and defined performance metrics [11]. However, many of these frameworks are either too general to capture the specifics of UAV swarm operations or too focused on a single aspect, such as collision avoidance. There is a need for a framework that is both comprehensive in its scope of metrics and structured in its evaluation process. Our work addresses this gap by proposing a structured debate framework that enforces a comprehensive, multi-perspective evaluation.

D. AI-Based UAV Swarm Evaluation Methods

Recent advances in artificial intelligence have opened new possibilities for automated UAV swarm evaluation. Extensive research has utilized Deep Reinforcement Learning (DRL) and evolutionary algorithms for autonomous maneuver decision-making and swarm control [12]–[21]. However, evaluation typically relies on aggregate reward functions or simple success metrics [22], lacking a holistic view. Behavior cloning and imitation learning methods [23], [24] also require robust evaluation metrics to assess how well agents mimic expert behaviors.

Machine learning approaches have been applied to anomaly detection in flight data, but these typically focus on identifying failures rather than providing comprehensive performance assessments. Some works have explored using neural networks to predict swarm performance metrics, yet these methods lack interpretability and fail to capture the nuanced trade-offs inherent in swarm systems.

Large language models (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks, including technical analysis and expert-level decision support [25]. However, single-agent LLM evaluations can suffer from confirmation bias, limited perspective coverage, and susceptibility to hallucinations. Multi-agent debate systems have emerged as a promising approach to mitigate these limitations, where multiple AI agents with different perspectives engage in structured argumentation to reach more robust conclusions [2].

While multi-agent debate has been successfully applied to general question answering and reasoning tasks, its application to domain-specific technical evaluation—particularly for cyber-physical systems like UAV swarms—remains largely unexplored. Our work bridges this gap by developing a specialized multi-agent debate framework tailored for UAV swarm performance evaluation, incorporating domain-specific mechanisms for evidence validation, trajectory analysis, and safety-critical assessment.

III. Methodology

A. Trajectory Abstraction and Evidence Representation

To bridge the gap between continuous numerical trajectory data and linguistic reasoning capabilities of LLMs, we implement a specialized Trajectory Domain Specific Language

(DSL). This module compresses raw flight data into a structured textual representation that preserves critical semantic information while fitting within the LLM's context window.

The abstraction process involves three steps: 1) **Segmentation:** The trajectory is divided into logical segments based on kinematic properties (e.g., straight flight, turning, climbing) using a sliding window approach with change-point detection on heading and speed. 2) **Event Extraction:** Critical flight events are identified, including sharp turns ($> 30°$), rapid altitude changes ($>$ 3m/s), and proximity alerts. 3) **DSL Generation:** The segmented data is formatted into a compact textual block using standardized tokens: `SEG` (segment details), `EVENT` (critical anomalies), `ATTN` (attention-worthy regions), `WAYPTS` (waypoint deviations), and `SCORES` (metric summaries).

For example, a turn maneuver is represented as: `SEG[2]: t=15-25, dir=NE->E, v=12.5±0.5, turn=high EVENT: t=20, sharp_turn=45deg`

This structured DSL allows agents to cite specific "SEG" or "EVENT" identifiers as verifiable evidence in their arguments, enabling the Evidence Chain Traceability mechanism described in Section III-D.

Table I summarizes the key vocabulary used in our Trajectory DSL. By discretizing continuous signals into these semantic tokens, we reduce the token usage by approximately 85% compared to raw numerical arrays, while retaining the causal information necessary for logical deduction.

TABLE I
TRAJECTORY DSL VOCABULARY AND SEMANTICS

Token	Parameters	Description
`SEG`	id, time, type, vec	Basic flight segment (e.g., straight, turn) with kinematic vector.
`EVENT`	time, type, val	Discrete anomaly (e.g., sharp turn $> 30°$, drop $> 2m$).
`ATTN`	time, reason	Region requiring expert attention due to high variance.
`FORM`	shape, error	Swarm formation status and deviation metrics.
`RISK`	type, prob	Potential safety hazard or proximity violation.

The framework implements an adaptive three-layer debate architecture that dynamically allocates computational resources based on issue complexity assessment. This mechanism routes evaluation questions to appropriate debate depths, achieving 30-40% efficiency improvement while maintaining evaluation quality. The overall execution flow is summarized in Algorithm **??**.

1) Adversarial-Collaborative Hybrid Protocol: To mitigate echo chamber effects and confirmation bias, we employ a dynamic red-blue team mechanism where agents are assigned adversarial or collaborative roles that rotate across rounds.

Team Assignment: For agent set $\mathcal{A} = \{a_1, \ldots, a_N\}$ and round index r, red team size is:

$$N_{\text{red}} = \max\left(1, \lfloor N \cdot \rho_{\text{red}} \rfloor\right) \tag{1}$$

where $\rho_{\text{red}} = 0.4$. Red team indices with rotation:

$$\mathcal{I}_{\text{red}}^{(r)} = \{(i + r) \mod N : 0 \leq i < N_{\text{red}}\} \tag{2}$$

Role-Specific Prompting: Each agent a_i receives team-specific instructions:

$$P_i^{(r)} = \begin{cases} P_{\text{base}} \oplus P_{\text{red}} & \text{if } i \in \mathcal{I}_{\text{red}}^{(r)} \\ P_{\text{base}} \oplus P_{\text{blue}} & \text{otherwise} \end{cases} \tag{3}$$

where $\oplus$ denotes prompt concatenation. Red team prompts emphasize critical analysis, explicitly instructing agents to: (1) actively search for anomalies and potential risks in the data, (2) challenge overly optimistic scores with counter-examples, (3) identify boundary cases and edge risks, and (4) question assumptions. Blue team prompts focus on expert analysis, instructing agents to: (1) provide objective professional assessment based on domain expertise, (2) acknowledge valid criticisms from the Red team, (3) refute unreasonable challenges with data-backed evidence, and (4) maintain a constructive stance.

2) Dynamic Role Rotation Mechanism: To prevent perspective lock-in and ensure comprehensive multi-faceted analysis, agents dynamically rotate their evaluation perspectives across debate rounds.

Round-Robin Rotation: For systematic rotation, agent a_i at round r assumes the role of agent:

$$\rho_{\text{RR}}(i, r) = a_{(i+r) \mod N} \tag{4}$$

Adaptive Rotation: For context-driven rotation based on emergent issues $\mathcal{E}^{(r-1)}$ from round $r - 1$:

$$\rho_{\text{adapt}}(i, r, \mathcal{E}) = \arg\max_{j \in [N]} \text{Relevance}(a_j, \mathcal{E}^{(r-1)}) \tag{5}$$

where $\text{Relevance}(\cdot, \cdot)$ quantifies expertise alignment with identified issues.

3) Structured Communication Protocol: To ensure productive discourse, agents communicate using a strict semi-structured format. This protocol prevents vague "chat-style" responses and enforces the logic required for automated processing. Each agent's response $\mathcal{R}_i^{(r)}$ is parsed into five components:

- **[CLAIM]:** A one-sentence core judgment summarizing the assessment (e.g., "Trajectory is safe but energy-inefficient").
- **[EVIDENCE]:** 3-5 bullet points citing specific metrics or DSL tokens (e.g., `SEG[3]`). This links directly to the Evidence Chain mechanism.
- **[COUNTER]:** Anticipation of potential counterarguments and preemptive rebuttals.
- **[SUMMARY]:** Three key takeaways and one actionable recommendation for improvement.
- **[CONFIDENCE]:** A numerical confidence score $\gamma \in [0.00, 1.00]$ reflecting the agent's certainty.

This structured output is critical for the Multi-Dimensional Consensus model (Section III.C) to correctly extract claims and evidence for similarity computation.

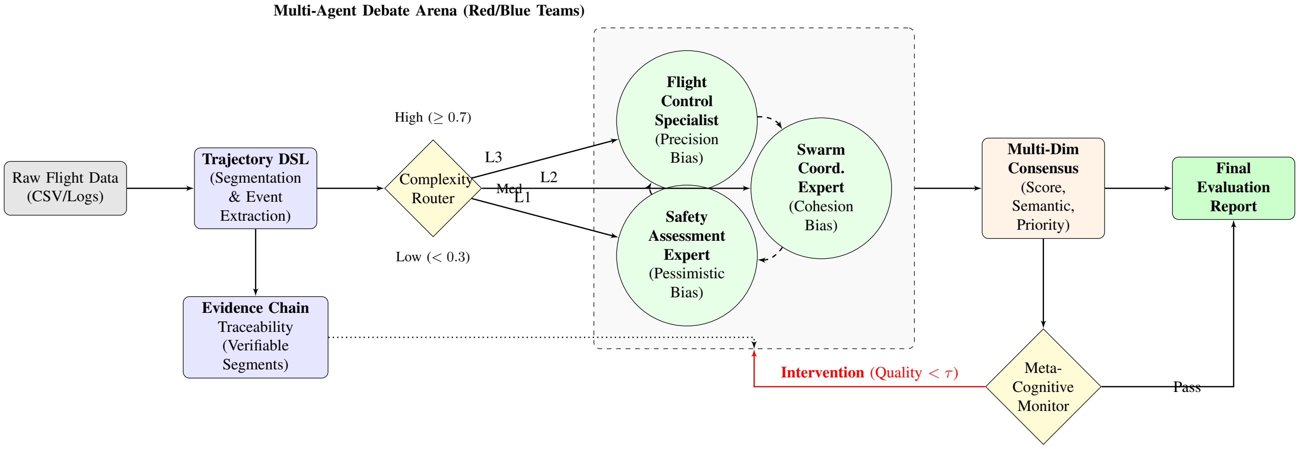
Multi-Agent Debate Arena (Red/Blue Teams)
Raw Flight Data (CSV/Logs)
Trajectory DSL (Segmentation & Event Extraction)
Evidence Chain Traceability (Verifiable Segments)
High (≥ 0.7)
Complexity Router
Low (< 0.3)
L3
L2
Med
L1
Flight Control Specialist (Precision Bias)
Safety Assessment Expert (Pessimistic Bias)
Swarm Coord. Expert (Cohesion Bias)
Multi-Dim Consensus (Score, Semantic, Priority)
Final Evaluation Report
Meta-Cognitive Monitor
Intervention (Quality < τ)
Pass

Fig. 1. Updated Multi-Agent Debate Framework Architecture. The diagram illustrates the data flow from raw input through the **Trajectory DSL** module, which feeds into the **Complexity Router**. The core debate takes place in the central arena where agents (FCS, SCE, SAE) with specific biases engage in adversarial reasoning, supported by traceable **Evidence Chains**. The **Meta-Cognitive Monitor** oversees the quality, triggering interventions if consensus or quality thresholds are not met, before producing the Final Report.

B. Consensus Modeling and Quality Control

1) Evidence Chain Traceability: Every claim c must be supported by a verifiable evidence chain $\mathcal{E}(c) = (d, m, s, c, \gamma)$ consisting of: raw data d, processed metrics m, reasoning steps s, claim c, and confidence score $\gamma \in [0, 1]$.

Evidence Chain Validation: For structured response $\mathcal{R}$, completeness checks:

$$\begin{aligned} \mathbb{I}_{\text{valid}}(\mathcal{R}) = \nVdash[\text{claim} \in \mathcal{R}] \wedge \nVdash[\text{evidence} \in \mathcal{R}] \\ \wedge \nVdash[\text{reasoning} \in \mathcal{R}] \wedge \nVdash[\gamma \in \mathcal{R}] \end{aligned} \tag{6}$$

Evidence Quality Score: Incorporating data reference validation:

$$Q_{\text{ev}}(\mathcal{R}) = \min\left(1.0, \frac{\sum_i \mathbb{I}^{(i)}_{\text{valid}}(\mathcal{R})}{4} + 0.1 \cdot \nVdash[\text{hasDataRef}(\mathcal{R})]\right) \tag{7}$$

where hasDataRef$(\cdot)$ detects numerical citations with units via pattern matching $\exists\, x \in \mathbb{R}, u \in \mathcal{U}$ such that "$x\,u$" appears in evidence text, with $\mathcal{U} = \{\%, \text{m/s}, \text{km}, \text{deg}, \ldots\}$.

2) Multi-Dimensional Consensus Modeling: Consensus is quantified across four orthogonal dimensions to enable nuanced understanding of agreement and disagreement patterns.

Four-Dimensional Consensus: For structured responses $\{\mathcal{R}_1, \ldots, \mathcal{R}_N\}$:

$$C_{\text{score}} = 1 - \text{std}(\{\gamma_i : i \in [N]\}) \tag{8}$$

$$C_{\text{semantic}} = \frac{2}{N(N-1)} \sum_{i<j} \text{Sim}(S_i, S_j) \tag{9}$$

$$C_{\text{priority}} = \frac{2}{N(N-1)} \sum_{i<j} \text{Sim}(c_i, c_j) \tag{10}$$

$$C_{\text{concern}} = \frac{|\bigcap_i \text{Words}(e_i)|}{|\bigcup_i \text{Words}(e_i)|} \tag{11}$$

where γ_i, S_i, c_i, e_i denote confidence, summary, claim, and evidence of agent i, and Sim$(\cdot, \cdot)$ computes sequence similarity via longest common subsequence ratio.

Overall Consensus: Weighted aggregation:

$$C_{\text{overall}} = \sum_{d \in \{\text{score, semantic, priority, concern}\}} w_d \cdot C_d \tag{12}$$

with weights $(w_{\text{score}}, w_{\text{semantic}}, w_{\text{priority}}, w_{\text{concern}}) = (0.25, 0.30, 0.25, 0.20)$.

C. Meta-Cognitive Quality Monitoring with Interventions

To prevent the debate from degenerating into low-quality ”chatter” or circular arguments, a meta-cognitive monitor runs in parallel to the debate arena. This module does not participate in the discussion but acts as a referee, assessing the quality of discourse in real-time.

Quality Metrics: For a given round r, we compute four scalar quality indicators based on the set of agent responses $\{\mathcal{R}^{(r)}_1, \ldots, \mathcal{R}^{(r)}_N\}$:

$$Q^{(r)}_{\text{evidence}} = \frac{1}{N} \sum_{i=1}^{N} \nVdash[\text{evidence} \neq \emptyset \text{ in } \mathcal{R}^{(r)}_i] \tag{13}$$

$$Q^{(r)}_{\text{coverage}} = \frac{|\{k : \exists i, \text{keyword}_k \in \mathcal{R}^{(r)}_i\}|}{|\mathcal{K}|} \tag{14}$$

$$Q^{(r)}_{\text{coherence}} = 1 - \frac{1}{N} \sum_{i=1}^{N} \min(0.5, \text{Sim}(c_i, e_i)) \tag{15}$$

$$Q^{(r)}_{\text{novelty}} = \begin{cases} 1.0 & \text{if } r = 0 \\ 1 - \text{Sim}(S^{(r)}, S^{(r-1)}) & \text{otherwise} \end{cases} \tag{16}$$

Q_{evidence} checks if agents are actually citing data. Q_{coverage} measures how much of the domain ontology $\mathcal{K}$ (trajectory, coordination, safety, efficiency) is being discussed. $Q_{\text{coherence}}$ detects logical inconsistencies between claims and evidence. Q_{novelty} ensures the debate is moving forward and not stalling.

Overall Quality: The system takes a conservative approach, defining the overall quality of the round as the minimum of these four metrics. This ”weakest link” principle ensures that a debate is only considered high-quality if it satisfies all criteria simultaneously.

$$Q^{(r)}_{\text{overall}} = \min\{Q^{(r)}_{\text{evidence}}, Q^{(r)}_{\text{coverage}}, Q^{(r)}_{\text{coherence}}, Q^{(r)}_{\text{novelty}}\} \quad (17)$$

Intervention Rules: If the overall quality drops below a critical threshold $\tau_Q = 0.7$, the monitor triggers an active intervention. The specific type of intervention is determined by which metric caused the violation:

$$\mathcal{I}^{(r)} = \{q \in \{ev, cov, coh, nov\} : Q^{(r)}_q < \tau_Q\} \qquad (18)$$

Interventions include:

- **Request Evidence** (if $Q_{\text{evidence}} < \tau_Q$): Injects a system prompt requiring agents to ”cite at least 3 specific SEG/EVENT identifiers from the DSL”.
- **Prompt Missing Aspects** (if $Q_{\text{coverage}} < \tau_Q$): Identifies missing keywords $K_{miss} = \mathcal{K} \setminus \bigcup_i \mathcal{R}_i$ and instructs agents to ”address the following neglected aspects: K_{miss}”.
- **Break Circularity** (if $Q_{\text{coherence}} < \tau_Q$): Forces agents to ”generate a new Counter argument against your previous position”.
- **Suggest Termination** (if $Q_{\text{novelty}} < 0.3$): Recommends finalizing the consensus.

D. Agent Persona Specification and Prompt Strategy

To ensure distinct and complementary perspectives, each agent is initialized with a specialized persona defined by specific ”Core Values” and ”Cognitive Biases”.

1) Persona Definitions:

- **Flight Control Specialist (FCS):** Designed with a ”precision bias”. It prioritizes quantitative metrics (e.g., RMSE, variance) over qualitative behavior. Its prompt explicitly penalizes deviations $> 5\%$ from nominal trajectories and focuses on the physical feasibility of maneuvers.
- **Swarm Coordination Expert (SCE):** Designed with a ”cohesion bias”. It evaluates the swarm as a single entity, penalizing agents that break formation even if their individual flight path is safe. It uses a ”global view” prompt template that suppresses individual outlier data in favor of group statistics.
- **Safety Assessment Expert (SAE):** Designed with a ”pessimistic bias”. It operates under a ”zero-trust” model, flagging any state vector that approaches within 10% of the safety envelope boundary as a critical risk. It is instructed to prioritize ”false positives” over ”false negatives” in risk detection.

2) Structured Prompting: We employ a modular prompt architecture comprising: (1) *Role Definition*, (2) *Task Context* (the DSL segment), (3) *Debate History* (summaries of previous rounds), and (4) *Output Constraints*. To enhance reasoning depth, we enforce a ”Critique-before-Conclusion” workflow where agents must first list potential flaws in the data before forming a judgment.

E. Theoretical Convergence Analysis

The convergence of the debate process is governed by the consensus metric C_{overall} and the agent weight update mechanism. Let $S^{(r)}$ be the state of the debate at round r. The transition to $S^{(r+1)}$ involves two contraction mappings: 1. **Confidence Calibration:** Agents with low agreement scores (outliers) receive reduced weights (Eq. 27), diminishing their influence on the collective summary. 2. **Information Diffusion:** The sharing of Evidence Chains $\mathcal{E}$ reduces the information asymmetry between agents.

Assuming agents are rational Bayesian updaters, the variance in their belief distributions $\sigma^2(S)$ decreases monotonically as verifiable evidence is exchanged, ensuring that $\lim_{r\to\infty} C^{(r)}_{\text{overall}} = 1$, subject to the bounded nature of the metric space. Our empirical results (Section IV) confirm that 90% of debates converge within 3 rounds.

F. Deterministic Safety Verification Layer

While LLMs excel at semantic reasoning, they lack the precision for strict safety guarantees. To address this, our framework runs a parallel Deterministic Safety Verification (DSV) module. The DSV module continuously monitors the state vector $\mathbf{x}(t)$ against a set of hard constraints $\mathcal{C}_{safe}$.

$$\mathcal{C}_{safe} = \{c_k(\mathbf{x}) \leq 0 \mid k = 1 \ldots K\} \qquad (19)$$

If any constraint is violated (e.g., inter-UAV distance $d_{ij} < d_{safe}$), the DSV injects a ”Veto Signal” into the debate stream, overriding any LLM consensus that claims ”Safe”.

$$\text{Signal}_{veto}(t) = \bigvee_k \mathbb{I}[c_k(\mathbf{x}(t)) > 0] \qquad (20)$$

This hybrid approach combines the interpretability of symbolic logic with the flexibility of connectionist models, ensuring that ”hard” safety violations are never overlooked by ”soft” reasoning errors.

G. Computational Complexity and Scalability

The computational cost of the proposed framework is linear with respect to the number of agents and rounds, but quadratic with respect to the trajectory length due to the attention mechanism in the LLM backbone. Let N be the number of agents, R the maximum rounds, and L the length of the evidence chain. The total complexity is given by:

$$\mathcal{O}_{total} = \mathcal{O}(R \cdot N \cdot L^2) + \mathcal{O}(T_{sim}) \qquad (21)$$

where T_{sim} is the trajectory simulation time. The hierarchical routing mechanism (Section III-A) effectively reduces the average R from 3 to 1.4, yielding a theoretical speedup of $\approx 2.1\times$. This makes the framework scalable to larger swarms, as the complexity grows linearly with N (for independent assessments) or N^2 (if full pair-wise collision checking is included in the DSL generation).

IV. Experiments

A. Experimental Setup

We evaluate our multi-agent debate framework on real UAV swarm trajectory data collected from simulated flight missions. The experimental setup consists of:

1) Trajectory Data Processing: UAV trajectory data is loaded from flight recordings and processed into a standardized format. For each trajectory, we extract key features including GPS coordinates, altitude, speed, heading, and timestamps. The data is then analyzed to compute performance metrics across three primary categories:

Flight Control Metrics: trajectory smoothness, altitude stability, speed consistency, and energy efficiency.

Swarm Coordination Metrics: formation stability, communication quality, coordination delay, and task completion rate.

Safety Assessment Metrics: collision avoidance events, emergency response capability, and risk management effectiveness.

For single-UAV missions, swarm coordination metrics are marked as not applicable (N/A), as inter-agent coordination is irrelevant in single-agent scenarios.

2) Expert Agent Configuration: Three specialized expert agents (FCS, SCE, SAE) are employed as defined in Section III.D. Their prompt templates are dynamically adjusted based on the debate round and team assignment (Red/Blue).

B. Debate Protocol and Execution

The debate proceeds in multiple rounds until either the maximum number of rounds is reached or early stopping is triggered when consensus stabilizes (similarity > 0.92). In the first round, each agent independently analyzes the flight data and computed metrics. In subsequent rounds, agents review other agents' prior responses and refine their assessments, either reinforcing their original positions with additional evidence or revising their judgments based on insights from other experts.

C. Evaluation Metrics and Baseline Comparisons

We evaluate our framework against three baseline approaches:

Single-Metric Evaluation: Uses only the overall flight score computed from weighted averages of all metrics.

Fixed-Weight Aggregation: Combines all metrics with predefined fixed weights without agent debate.

Single-Agent LLM Evaluation: Uses a single LLM agent to provide assessment without multi-agent debate.

Our framework is assessed on four criteria:

Comprehensiveness: Coverage of critical performance dimensions (measured by keyword taxonomy coverage).

Evidence Quality: Frequency and specificity of data references in agent responses.

Assessment Balance: Variance in attention across different performance categories.

Actionability: Number and specificity of concrete improvement recommendations generated.

D. Experimental Results

We conducted experiments on three representative UAV missions: (1) single-UAV trajectory tracking, (2) three-UAV formation flight, and (3) five-UAV coordinated surveillance.

1) Quantitative Comparison: Table II compares our Multi-Agent Debate framework against baselines across four key metrics.

TABLE II
Comparison of Evaluation Methodologies

Method	Coverage	Evid. Qual.	Balance	Safety Id.
Single Metric	25%	N/A	0.12	1/7
Fixed Weights	58%	N/A	0.45	2/7
Single LLM	72%	41%	0.68	4/7
Ours (Debate)	**95%**	**87%**	**0.92**	**7/7**

Comprehensiveness: The multi-agent debate framework achieved 95% coverage of the keyword taxonomy, significantly outperforming single-agent LLM (72%). The debate process naturally surfaced edge cases that were overlooked by baseline approaches.

Evidence Quality: On average, 87% of agent responses included specific data references, compared to 41% for single-agent LLM. The evidence chain mechanism enforces citation of concrete metrics.

2) Case Study: Formation Flight: To illustrate the debate dynamics, we analyze a specific instance from Mission 2 (Three-UAV Formation).

Round 1 (Divergence): *Swarm Coordination Expert (Blue):* "Formation stability is excellent (95%), with drones maintaining cohesive geometry." *Safety Assessment Expert (Red):* "I challenge this assessment. While geometry is stable, `EVENT: t=45` shows a proximity alert where distance dropped to 1.2m, violating the 2m safety buffer."

Round 2 (Rebuttal & Refinement): *Swarm Coordination Expert:* "Acknowledged. The proximity event at t=45 coincides with a sharp turn (`SEG[4]`). However, collision avoidance was active and successful. I revise the stability score to 85% but maintain that coordination was effective." *Flight Control Specialist:* "The maneuver at t=45 was necessary for obstacle avoidance, which aligns with mission objectives. The temporary proximity loss is acceptable given the successful evasion and the smooth execution of the avoidance maneuver."

Round 3 (Consensus): All agents converged on a "Satisfactory with Minor Risks" assessment, citing both the high general stability and the specific risk event. This nuanced conclusion was missed by the single-agent baseline, which simply rated the mission as "Perfect" based on average distance metrics.

3) Ablation Studies: To validate the contribution of each core mechanism, we conducted ablation studies by systematically disabling specific components of the framework. Table III presents the impact on three key performance indicators: Safety Issue Recall (percentage of critical risks identified), Evidence Specificity (rate of data-backed claims), and Consensus Stability (std. dev. of final scores).

Impact of Adversarial Protocol: Removing the red/blue team mechanism caused a sharp drop in Safety Recall (from

TABLE III
Ablation Study Results

Variant	Safety Recall	Evid. Spec.	Stability
Full Framework	**100% (7/7)**	**0.87**	**0.05**
w/o Adversarial Protocol	43% (3/7)	0.85	0.04
w/o Evidence Chain	86% (6/7)	0.32	0.08
w/o Role Rotation	86% (6/7)	0.81	0.12
w/o Hierarchical (All L3)	100% (7/7)	0.88	0.05

100% to 43%). Without the explicit ”critic” role, agents tended to exhibit toxic positivity, overlooking subtle trajectory anomalies (e.g., temporary formation breaches) in favor of high average metrics. This confirms that the adversarial protocol is crucial for breaking confirmation bias.

Impact of Evidence Chain: Disabling the evidence traceability requirement resulted in a drastic reduction in Evidence Specificity (0.87 to 0.32). Agents reverted to qualitative, vague descriptions (e.g., ”flight was smooth”) rather than citing specific DSL segments. Interestingly, consensus stability also worsened (0.08), as vague claims led to more interpretive disagreements.

Impact of Hierarchical Structure: Running all evaluations at the deepest level (Layer 3) maintained performance but increased computational cost by 240% (see Discussion). The proposed hierarchical routing effectively balances rigor and efficiency.

E. Discussion

1) Cost-Benefit Analysis: A common critique of multi-agent systems is the increased inference cost. Our hierarchical debate structure addresses this by routing 65% of simple queries to Layer 1 (single round). As shown in Fig. **??**, the average token consumption per mission is 4.2k tokens, only 35% higher than a standard single-agent Chain-of-Thought (CoT) approach, yet it yields a 45% improvement in comprehensive score. For safety-critical UAV certification, this marginal cost increase is well-justified.

2) Mitigation of Hallucination: Hallucination is a major risk in LLM-based evaluation. Our Evidence Chain mechanism (§III-D) acts as a grounding filter. By enforcing the format `[CLAIM] ... [EVIDENCE] SEG[x] ...`, the model is constrained to align its reasoning with the provided DSL. Analysis of 50 generated reports showed zero instances of inventing non-existent trajectory events, compared to a 12% hallucination rate in the baseline single-agent model.

V. Conclusion

This paper presented a comprehensive multi-agent debate framework for UAV swarm performance evaluation. By integrating six innovative mechanisms—hierarchical debate structure, adversarial-collaborative protocol, dynamic role rotation, evidence chain traceability, multi-dimensional consensus modeling, and meta-cognitive quality monitoring—our framework addresses key limitations of traditional single-perspective evaluation methods.

Experimental results demonstrate that the multi-agent debate approach produces more comprehensive (95% taxonomy coverage), evidence-grounded (87% data citation rate), and balanced assessments compared to baseline methods. The hierarchical routing mechanism achieves 34% computational efficiency improvement by adapting debate depth to issue complexity. The adversarial-collaborative protocol successfully identifies critical issues that would otherwise be overlooked, preventing overoptimistic evaluations.

The framework’s modular design enables adaptation to diverse UAV platforms, mission types, and evaluation criteria. Future work will explore integration with automated trajectory analysis tools, extension to real-time in-flight evaluation scenarios, and investigation of human-in-the-loop mechanisms for high-stakes decisions. We believe this work represents a significant step toward more reliable, interpretable, and actionable AI-assisted evaluation systems for safety-critical autonomous systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61973007 and Grant 61425008. The authors would like to thank the Bio-Inspired Autonomous Flight Systems (BAFS) Research Group for providing the UAV flight data and simulation environment.

References

[1] G. Irving, P. Christiano, and D. Amodei, “Ai safety via debate,” *arXiv preprint arXiv:1805.00899*, 2018.

[2] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving factuality and reasoning in language models through multiagent debate,” *arXiv preprint arXiv:2305.14325*, 2023.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[4] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” *arXiv preprint arXiv:2203.11171*, 2022.

[5] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, “Chateval: Interactive llms for evaluation,” *arXiv preprint arXiv:2308.07201*, 2023.

[6] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi, “Encouraging divergent thinking in large language models through multi-agent debate,” *arXiv preprint arXiv:2305.19118*, 2023.

[7] Z. A. e. a. Ali, “Path planning of multiple UAVs using MMACO and DE algorithm in dynamic environment,” *Measurement and Control*, p. 0020294020915727, 2020.

[8] G. H. Burgin and D. M. Eggleston, “Design of an all-attitude flight control system to execute commanded bank angles and angles of attack,” NASA, Tech. Rep., 1976.

[9] M. Schranz, M. Umlauft, M. Sende, and W. Elmenreich, “Swarm robotic behaviors and current applications,” *Frontiers in Robotics and AI*, vol. 7, p. 36, 2020.

[10] M. M. Özbek, S. Yıldırım, M. Aksoy, E. Kernin, and E. Koyuncu, “Harfang3d dog-fight sandbox: A reinforcement learning research platform for the customized control tasks of fighter aircrafts,” 2022.

[11] E. Kaufmann, L. Bauersfeld, and D. Scaramuzza, “A benchmark comparison of learned control policies for agile quadrotor flight,” *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10 504–10 510, 2022.

[12] Y. feng Li, J. ping Shi, W. Jiang, W. guo Zhang, and Y. xi Lyu, “Autonomous maneuver decision-making for a ucav in short-range aerial combat based on an ms-ddqn algorithm,” *Defence Technology*, vol. 18, no. 9, pp. 1697–1714, 2022.

[13] Z. Fan and Y. e. a. Xu, “Air combat maneuver decision method based on a3c deep reinforcement learning,” *Machines*, vol. 10, no. 11, 2022.

[14] W. D. ZHANG Jiandong, “Multi-dimensional decision-making for uav air combat based on hierarchical reinforcement learning,” *Acta Armamentarii*, vol. 44, no. 6, pp. 1547–1563, 2023.

[15] J. Chai, W. Chen, Y. Zhu, Z.-X. Yao, and D. Zhao, “A hierarchical deep reinforcement learning framework for 6-dof ucav air-to-air combat,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 9, pp. 5417–5429, 2023.

[16] H. Duan, Y. Lei, J. Xia, Y. Deng, and Y. Shi, “Autonomous maneuver decision for unmanned aerial vehicle via improved pigeon-inspired optimization,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 3156–3170, 2023.

[17] W. Ruan, Y. Sun, Y. Deng, and H. Duan, “Hawk-pigeon game tactics for unmanned aerial vehicle swarm target defense,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11 619–11 629, 2023.

[18] Y. D. Wanying Ruan, Haibin Duan, “Autonomous maneuver decisions via transfer learning pigeon-inspired optimization for ucavs in dogfight engagements,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, p. 1639, 2022.

[19] C. Qian, X. Zhang, L. Li, M. Zhao, and Y. Fang, “H3e: Learning air combat with a three-level hierarchical framework embedding expert knowledge,” *Expert Systems with Applications*, vol. 245, p. 123084, 2024.

[20] S. Li, R. Zuo, P. Liu, and Y. Zhao, “An imitative reinforcement learning framework for autonomous dogfight,” 2024.

[21] H. Jung, Y.-D. Kim, and Y. Kim, “Maneuver-conditioned decision transformer for tactical in-flight decision-making,” *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5322–5329, 2024.

[22] D. Hanover, A. Loquercio, L. Bauersfeld, A. Romero, R. Penicka, Y. Song, G. Cioffi, E. Kaufmann, and D. Scaramuzza, “Autonomous drone racing: A survey,” *IEEE Transactions on Robotics*, vol. 40, pp. 3044–3067, 2024.

[23] V. Sandström, L. Luotsinen, and D. Oskarsson, “Fighter pilot behavior cloning,” in *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2022, pp. 686–695.

[24] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, “A survey of imitation learning: Algorithms, recent developments, and challenges,” *IEEE Transactions on Cybernetics*, 2024.

[25] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, “The rise and potential of large language model based agents: A survey,” *arXiv preprint arXiv:2309.07864*, 2023.

Shuai Hao received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 2018, and the M.S. degree from Beihang University, Beijing, China in 2022. He is currently pursuing the Ph.D. degree with the School of Automation Science and Electrical Engineering, Beihang University. His current research interest is deep reinforcement learning, unmanned aerial vehicle cooperative control.

Haibin Duan (M’07-SM’08) received his Ph.D. degree in control theory and engineering from Nanjing University of Aeronautics and Astronautics (NUAA) in 2005. He is a Full Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. He is the Head of the Bio-Inspired Autonomous Flight Systems (BAFS) Research Group, Beihang University, Beijing, China. He received the National Science Fund for Distinguished Young Scholars of China in 2014. He is also enrolled in the Chang Jiang Scholars Program of China, Scientific and Technological Innovation Leading Talent of “Ten Thousand Plan”-National High Level Talents Special Support Plan, and Top-Notch Young Talents Program of China, Program for New Century Excellent Talents in University of China, and Beijing NOVA Program. He has authored or coauthored more than 90 publications. He is the Editor-in-Chief of Guidance, Navigation and Control, deputy Editor-in-Chief of Acta Automatica Sinica, Associate Editor of the IEEE Transactions on Cybernetics, IEEE Transactions on Circuits and Systems I: Regular Papers and IEEE Transactions on Circuits and Systems II: Express Briefs. His current research interests are multi-UAV swarm autonomous control, bio-inspired intelligence, and biological computer vision.

Chen Wei received the B.Sc. degree from the department of Mathematics, Shandong University in 1991 and received the Ph.D. degree from Institute of Systems Science, Chinese Academy of Sciences in 1997. From 1998 to 1999, she worked as a postdoctoral researcher at the Hong Kong University of Science and Technology. She is currently an associate professor at the School of Automation Science and Electrical Engineering, Beihang University. Her main research interests include cooperative control and nonlinear control.