UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents

Finley Holt

2025-12-25

Table of contents

1 UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents

Authors: Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjing Wu, Weili Guan, Xiang Deng **Institution**: Harbin Institute of Technology, Shenzhen **Source**: https://arxiv.org/html/2508.00288 **ArXiv ID**: arXiv:2508.00288v4 [cs.RO] **Published**: August 22, 2025 **Type**: Benchmark Paper **Code**: https://github.com/Kyaren/UAV_ON

1.1 Overview

UAV-ON introduces the first large-scale benchmark for instance-level Object Goal Navigation (ObjectNav) by aerial agents in open-world environments. Unlike existing Vision-and-Language Navigation (VLN) paradigms that rely on sequential, step-by-step linguistic instructions, UAV-ON enables UAVs to operate based on high-level semantic goals, promoting greater autonomy and scalability.

1.1.1 Key Innovation

The benchmark shifts from VLN's detailed instruction-following to ObjectNav's goal-driven exploration, where agents receive only a semantic instruction `c = {name, size, description}` and must autonomously navigate to locate the target object without GPS, global maps, or external guidance.

1.1.2 Benchmark Scope

- **14 high-fidelity Unreal Engine environments** spanning urban, natural, and mixed-use settings
- **1,270 annotated target objects** with instance-level instructions
- **11,000+ navigation tasks** across diverse semantic regions
- **Physically grounded simulation** requiring discrete, parameterized actions with collision dynamics
- **Total coverage**: Approximately 9 million square units with average object density of 1.4 per 100 square units

---

1.2 Decision Impact for Flyby-F11

1.2.1 ADOPT - High Confidence

ObjectNav Paradigm over VLN: - **Evidence**: UAV-ON demonstrates semantic goal specification enables greater autonomy than step-by-step VLN instructions - **Decision**: Adopt high-level semantic goal specification for flyby-f11 mission planning - **Rationale**: Natural language mission intents ("locate vehicle convoy") map directly to ObjectNav formulation rather than detailed movement commands

Safety as Critical Constraint: - **Evidence**: All baseline methods exhibit 37-65.5% collision rates, making them unsuitable for real deployment - **Decision**: Implement explicit safety constraints as hard requirements, not learned behaviors - **Rationale**: Collision avoidance cannot be treated as a soft objective in RL reward functions; ontology-based preconditions must enforce safety a priori

Semantic Goal Specification Format: - **Evidence**: UAV-ON's `c = {name, size, description}` instruction format successfully captures instance-level object specification - **Decision**: Use structured semantic instructions for flyby-f11 mission goals - **Rationale**: Format supports both category-level ("locate any vehicle") and instance-level ("red sedan near treeline") specifications

1.2.2 CONSIDER - Needs Validation

LLM for High-Level Planning Only: - **Evidence**: AOA demonstrates strong semantic reasoning (26.30% OSR) but poor control safety (45-65.5% collision rates) - **Decision**: Consider LLM for mission interpretation and high-level planning, NOT for direct control - **Rationale**: Leverage LLM semantic understanding while constraining low-level control via ontology + RL - **Validation Needed**: Test whether LLM can reliably translate natural language mission intents into formal ontology queries

Variable vs. Fixed Control Granularity: - **Evidence**: AOA-V shows better exploration (26.30% OSR, 45% collision) vs. AOA-F (17.50% OSR, 65.5% collision) - **Decision**: Investigate

context-dependent switching between variable and fixed-step control - **Rationale**: Variable control benefits exploration; fixed control improves reliability in cluttered environments - **Validation Needed**: Determine ontological conditions for switching between control modes

Multi-View Sensor Configuration: - **Evidence**: UAV-ON uses 4-view RGB-D (front, left, right, down) for comprehensive egocentric awareness - **Decision**: Evaluate whether flyby-f11's sensor suite (ZED 2i stereo, ToF) requires similar multi-view setup - **Rationale**: Single front-facing camera may be insufficient for safe 3D navigation; need lateral/downward awareness - **Validation Needed**: Simulate occlusion scenarios to determine minimum sensor coverage

1.2.3 AVOID - Evidence Against

Pure LLM Control: - **Evidence**: AOA-F achieves 7.30% SR but 65.5% collision rate; AOA-V achieves 4.20% SR with 45.0% collision rate - **Decision**: Do NOT use LLM directly for low-level motion control on flyby-f11 - **Rationale**: Unacceptable collision rates demonstrate LLMs cannot reliably reason about spatial dynamics and obstacle avoidance - **Alternative**: Use LLM for semantic reasoning, ontology for safety constraints, RL for control

Unconstrained Learning Approaches: - **Evidence**: CLIP-H (51.4% collision), AOA-V (45.0%), AOA-F (65.5%) all exceed acceptable safety thresholds - **Decision**: Reject pure end-to-end learning without explicit safety constraints - **Rationale**: Real-world deployment requires <5% collision rate; learned policies without hard constraints are insufficient - **Alternative**: Ontology-constrained RL with formally verified safety properties

VLN Paradigm for Autonomous Missions: - **Evidence**: VLN requires step-by-step instructions, limiting autonomy and scalability (Table 1 comparison) - **Decision**: Do NOT adopt VLN paradigm for flyby-f11 mission architecture - **Rationale**: Communications-denied operations require goal-driven autonomy, not detailed instruction-following - **Alternative**: ObjectNav paradigm with semantic goal specification

1.2.4 INVESTIGATE - Open Questions

How to Close OSR-SR Gap: - **Observation**: AOA-V achieves 26.30% OSR but only 4.20% SR (6.2x gap) - **Implication**: Agent can locate targets but cannot reliably decide when to terminate - **Research Question**: Can ontology-based goal satisfaction predicates improve termination decisions? - **Proposed Experiment**: Define formal stopping criteria in ontology (e.g., `goalSatisfied(state) ← visualConfidence(target) > distance < 20`), measure impact on SR

Termination Logic Design: - **Observation**: Separate termination from exploration appears critical (AOA struggles with multitasking) - **Implication**: Need dedicated module for goal satisfaction assessment - **Research Question**: Should termination use (a) learned classifier, (b) ontological rule, or (c) hybrid approach? - **Proposed Experiment**: Implement all three on UAV-ON test set, measure precision/recall of Stop command

Object-Scene Co-Occurrence Effectiveness: - **Observation**: UAV-ON uses LLM-generated object placement based on semantic regions - **Implication**: Ontology could encode similar co-occurrence priors for search guidance - **Research Question**: Can ontological knowledge of `likelyLocation(objectType, sceneRegion)` improve exploration efficiency? - **Proposed Experiment**: Compare random exploration vs. ontology-guided region prioritization on UAV-ON

Scalability to Real Hardware: - **Observation**: UAV-ON uses high-fidelity Unreal Engine

simulation; real-world deployment faces sim-to-real gap - **Implication**: Need to assess computational feasibility on Jetson Orin NX (50 TOPS) - **Research Question**: Can ontology-constrained RL run at real-time inference rates on edge hardware? - **Proposed Experiment**: Benchmark inference latency for perception → ontology reasoning → RL policy on Jetson Orin NX

1.2.5 Critical Evidence Table

Baseline Collision Rates (Table 3 data):

Method	Collision Rate	Success Rate	Oracle Success Rate	Safe Distance (m)
Random	37.9%	3.70%	8.00%	31.68
CLIP-H	51.4%	6.20%	11.90%	125.21
AOA-V (LLM variable)	45.0%	4.20%	26.30%	232.08
AOA-F (LLM fixed)	**65.5%**	7.30%	17.50%	144.25

Key Insights: 1. **No baseline achieves <40% collision rate** - all current approaches are unsafe for real deployment 2. **Best task performance (AOA-F: 7.30% SR) has worst safety (65.5% collision)** - fundamental safety-performance tradeoff 3. **LLM approaches show promise (26.30% OSR) but critical safety gaps** - cannot be used without constraints 4. **Exploration capability (OSR) inversely correlated with collision rate** - need safety constraints that preserve exploration

Decision Threshold: Real-world deployment requires <5% collision rate. Current methods exceed this by 7.5x to 13x.

1.2.6 Benchmark Applicability Assessment

Using UAV-ON to Test Ontology-Constrained RL:

Strengths: - **Standardized Evaluation**: SR, OSR, DTS, SPL metrics enable direct comparison with baselines - **Safety Tracking**: Collision rate measurement critical for validating ontology safety constraints - **Diverse Scenarios**: 14 environments with varying obstacle densities test generalization - **Training Data**: 10,000 episodes with A* shortest paths support RL training and imitation learning - **Semantic Complexity**: Instance-level object descriptions stress ontological reasoning capabilities

Limitations: - **Simulation-Only**: No real-world validation; sim-to-real gap unknown - **Static Environments**: No dynamic obstacles or moving objects (unlike real missions) - **Limited Weather/Lighting**: High-fidelity rendering but no adverse environmental conditions - **Computational Assumptions**: Assumes unlimited onboard compute; edge deployment constraints not modeled

Recommended Usage: 1. **Phase 1 - Proof of Concept**: Implement ontology-constrained RL on UAV-ON, target >15% SR with <10% collision rate (2x better SR, 4.5x better safety than AOA-F) 2. **Phase 2 - Ablation Studies**: Measure impact of (a) ontology safety constraints, (b) semantic co-occurrence reasoning, (c) termination logic design 3. **Phase 3 - Generalization Testing**: Evaluate on UAV-ON's 4 novel environments to assess ontology transfer capability 4. **Phase 4 - Hardware Validation**: Port best-performing approach to flyby-f11 simulation environment, then real hardware

Success Criteria for Benchmark Validation: - **Safety**: <10% collision rate (vs. 37-65% baseline) - **Task Performance**: >15% SR (vs. 7.30% best baseline) - **Efficiency**: >10% SPL (vs. 4.15% best baseline) - **Exploration**: Maintain >20% OSR (comparable to AOA-V's 26.30%) - **Interpretability**: Provide ontological explanations for all decisions (novel capability)

If ontology-constrained RL meets these criteria on UAV-ON, it provides strong evidence for deployment on flyby-f11 hardware.

1.3 Task Definition

1.3.1 Problem Formulation

At episode initialization, the UAV receives: - **6-DoF pose**: `P = [x, y, z, 0, , 0]` where `(x, y, z)` [3] is position and is yaw angle - **Semantic instruction**: `c = {name, size, description}` specifying: - Target object category - Estimated physical footprint (small/medium/large) - Instance-level visual descriptors (e.g., "a child wearing a pale green shirt and dark pants") - **Search constraint**: Target guaranteed within 50-unit horizontal radius

1.3.2 Sensory Configuration

The UAV operates with **four synchronized RGB-D cameras**: - **Front view**: Forward-facing egocentric perception - **Left view**: Lateral left awareness - **Right view**: Lateral right awareness - **Down view**: Ground-facing for altitude and terrain monitoring

Critical constraint: No GPS, global maps, or privileged information (object poses, semantic maps, scene geometry). Navigation relies entirely on egocentric visual observations and internal memory.

1.3.3 Action Space

Parameterized continuous control with motion primitives:

Translation Actions: - Move Forward (distance parameter) - Move Left (distance parameter) - Move Right (distance parameter) - Ascend (distance parameter) - Descend (distance parameter)

Rotation Actions: - Rotate Left (angular displacement parameter) - Rotate Right (angular displacement parameter)

Termination: - Stop (when agent estimates proximity within 20 units)

Unlike prior aerial benchmarks using fixed-step discrete actions or teleportation, UAV-ON requires **physically executed movements** with collision detection. Any obstacle contact constitutes failure.

1.3.4 Episode Termination

Episodes end when: 1. Agent issues Stop command 2. Collision with obstacle occurs 3. Maximum step limit (150 steps) reached

Success criterion: Agent stops within 20 units of target object.

1.4 Methodology

1.4.1 Scene Construction

Semantic Region Labeling: Environments segmented into categories: - Urban: village, town, city - Infrastructure: park, road - Natural: forest, mountain, snowy mountain, water area

Object Placement Strategy: Prompt-based object mapping using LLM-generated co-occurrence priors: - Benches and trash bins → parks - Bicycles → roads - Boats → water areas

This encourages **object-scene co-occurrence reasoning** rather than position memorization.

Spatial Diversity: - Scene scales range from 350×250 to 1400×1250 units - Varying elevation, obstacle density, and open space ratios - 1,270 unique targets across 14 environments

1.4.2 Dataset Analysis

Object Distribution (Figure 2): - **Small objects**: 18.05% - **Medium objects**: 22.30% - **Large objects**: 59.65%

Semantic Diversity: - Category word cloud shows broad object types (humans, vehicles, structures, natural features) - Description word cloud highlights rich linguistic cues (colors, poses, materials, spatial arrangements)

1.4.3 Dataset Split

Training Set (10 environments): - 10,000 navigation episodes - Proportional allocation by scene size and object distribution - Agents access absolute position and Euclidean distance to target - 3D voxelized grids (1-unit resolution) with A* shortest paths for imitation learning

Test Set (1,000 episodes): - 10 training environments + 4 novel environments - Mix of familiar/novel scenes and seen/unseen object categories - Replaced targets in training environments for generalization assessment

1.4.4 Evaluation Metrics

1. Success Rate (SR↑):

```
SR = (1/N) Σ  1{d    }
```

Where `d` is final distance to target, `= 20` units.

2. Oracle Success Rate (OSR↑):

```
OSR = (1/N) Σ  1{min_t(d , )   }
```

Measures if agent ever came within threshold during episode (upper bound to SR).

3. Distance to Success (DTS↓):

```
DTS = (1/N) Σ  d
```

Continuous assessment of final distance to goal.

4. Success-weighted Path Length (SPL↑):

```
SPL = (1/N) Σ  1{d    } · (l  / max(p , l ))
```

Where `l` is shortest geodesic path (computed via 3D A* on 1-unit occupancy grid) and `p` is actual trajectory length. Penalizes inefficient paths.

1.5 Baseline Methods

1.5.1 1. Random Baseline

Design: Randomly selects movement direction at each timestep with fixed translation/rotation step sizes. - No observation or goal information used - Stop disabled for first 10 steps to prevent premature termination - Serves as lower-bound performance anchor

1.5.2 2. CLIP-based Heuristic Exploration (CLIP-H)

Architecture: - Pre-trained CLIP model encodes four RGB views - Cosine similarity matching against textual target description - If similarity exceeds threshold → Stop - Otherwise, move toward highest-similarity direction

Post-processing Rules: - Scene boundary suppression (prevents out-of-bounds exploration) - Minimum flight altitude enforcement (collision avoidance)

Strengths: Strong visual similarity matching, efficient path following **Weaknesses**: Limited semantic reasoning, struggles with ambiguous objects

1.5.3 3. Aerial ObjectNav Agent (AOA)

Core Concept: Zero-shot modular policy built on pre-trained Multimodal Large Language Model (MLLM).

System Architecture (Figure 3):

Input Processing: 1. **RGB Views**: Encoded into textual captions via Qwen-VL model 2. **Depth Views**: Min-pooled into compact 3×3 spatial matrices 3. **Pose History**: Last 10 poses as 4D vectors `[x, y, z, ]` 4. **Target Information**: Structured semantic instruction

Structured Prompt to GPT-4o mini:

```
# Prompt Header: Role & Execution Rule
You are an autonomous UAV navigating in an open-world environment.
Follow the information and constraints below to decide your next action.

# Task Constraints
- Search radius: 50 units
```

```
- Maximum steps: 150
- Success threshold: within 20 units of the target

# Target Information
- Name: [category]
- Size: [estimated spatial footprint]
- Description: [visual attributes]

# RGB Captions
- Front: [caption from front RGB view]
- Left: [caption from left RGB view]
- Right: [caption from right RGB view]
- Down: [caption from downward RGB view]

# Depth Image Matrices
- FrontDepth: [3×3 depth grid]
- LeftDepth: [3×3 depth grid]
- RightDepth: [3×3 depth grid]
- DownDepth: [3×3 depth grid]

# Previous UAV Poses
- Pose history: [10-step trajectory of UAV [x, y, z, yaw]]
```

Output: GPT-4o mini generates semantic action command with continuous control magnitude: - Example: "Move forward 3 units" - Example: "Rotate left 60 degrees"

Two Variants:

AOA-Fixed (AOA-F): - Uses constant translation and rotation parameters - Compatible with discrete control benchmarks - Mirrors previous UAV navigation baselines (AerialVLN, CityNav)

AOA-Variable (AOA-V): - LLM determines both action type and magnitude - Greater behavioral adaptability - Exploits GPT's spatial reasoning for flexible control

Post-processing: - Out-of-bound movement actions replaced with in-place rotations - Addresses LLM's limited spatial boundary awareness - No minimum altitude enforcement (relies on depth cameras)

Operating Mode: Entirely zero-shot (no training or fine-tuning)

1.6 Key Findings

1.6.1 Quantitative Results (Table 2)

Overall Performance:

Method	DTS↓	SR↑	OSR↑	SPL↑
Random	42.57	3.70%	8.00%	2.66%
CLIP-H	46.31	6.20%	11.90%	4.15%
AOA-V	49.58	4.20%	26.30%	0.87%

Method	DTS↓	SR↑	OSR↑	SPL↑
AOA-F	48.37	**7.30%**	17.50%	**4.06%**

Performance by Object Size:

Small Objects (18.05% of dataset): - AOA-V: 25.44% OSR (best exploration) - AOA-F: 4.45% SR (best task completion) - CLIP-H: 1.51% SPL (most efficient)

Medium Objects (22.30% of dataset): - CLIP-H: 10.95% SR (best performance) - AOA-V: 27.62% OSR (best exploration) - CLIP-H: 7.17% SPL

Large Objects (59.65% of dataset): - AOA-F: 14.29% SR (best performance) - AOA-V: 27.95% OSR - AOA-F: 10.66% SPL

1.6.2 Termination and Safety Analysis (Table 3)

Episode Termination Types:

Method	Stop	Max Step	Collision	Avg. Steps	Safe Dist. (m)
Random	62.1%	0.0%	**37.9%**	12.09	31.68
CLIP-H	36.8%	11.8%	**51.4%**	30.25	125.21
AOA-V	19.9%	35.1%	**45.0%**	85.65	232.08
AOA-F	30.6%	3.9%	**65.5%**	36.88	144.25

Critical Safety Findings: - **All methods exhibit collision rates exceeding 37%**, with AOA-F reaching **65.5%** - Safe navigation distance (total collision-free trajectory) correlates with exploration capability - AOA-V achieves 232.08m safe distance despite 45% collision rate (extensive exploration before failure) - Fixed-step control (AOA-F) shows higher collision rate (65.5%) than variable-step (45.0%)

1.6.3 Performance Insights

1. Exploration vs. Termination Tradeoff: - **AOA-V**: Highest OSR (26.30%) but low SR (4.20%) - Strong semantic grounding and exploration - Poor termination decision-making - LLM struggles with multitasking (semantic understanding + motion planning + stop control)

2. Control Strategy Impact: - **AOA-F**: Better SR (7.30%) and SPL (4.06%) - Fixed-step motion simplifies control - More reliable trajectory execution - Reduced exploratory reach vs. AOA-V

3. Visual Similarity vs. Semantic Reasoning: - **CLIP-H**: Highest SPL (4.15%) - Efficient path following via visual similarity - Lower OSR (11.90%) suggests shallow semantic understanding - Struggles with ambiguous or visually complex objects

4. Object Size Dependency: - Large objects (59.65% of dataset) show best overall performance - Small objects most challenging (lowest SR across all methods) - Visual saliency directly impacts both exploration and termination accuracy

1.7 Technical Contributions

1.7.1 1. First Large-Scale Aerial ObjectNav Benchmark

Paradigm Shift: - From VLN's dense step-by-step supervision → high-level semantic goal autonomy - 11,000+ tasks with compact semantic instructions - 14 high-fidelity outdoor scenes with realistic object placements

1.7.2 2. Physically Grounded 3D Navigation

Distinguishing Features: - Parameterized continuous action space (vs. fixed discrete actions) - Physical execution with collision dynamics (vs. teleport-based control) - Multi-view RGB-D egocentric perception (no global information) - Object-scene co-occurrence reasoning requirements

1.7.3 3. Comprehensive Baseline Suite

Methodological Diversity: - Random (lower bound) - CLIP-H (visual similarity heuristic) - AOA-F (fixed-step LLM control) - AOA-V (variable-step LLM control)

Zero-Shot LLM Framework: - Novel integration of multimodal inputs (RGB captions, depth matrices, pose history) - Structured prompting for spatial reasoning - No task-specific training required

---

1.8 Challenges Identified

1.8.1 1. Semantic Navigation Difficulty

Evidence: - Best SR: 7.30% (AOA-F) - Best OSR: 26.30% (AOA-V) - Large gap between "ever reaching" (OSR) and "successfully stopping" (SR)

Implications: Combining semantic reasoning, exploration, and precise termination remains fundamentally difficult.

1.8.2 2. Safety Criticality

Evidence: - All methods: >37% collision rate - AOA-F (best SR): 65.5% collision rate - Fixed-step control increases collision risk in cluttered environments

Implications: Current approaches unsafe for real-world deployment. Need explicit safety constraints.

1.8.3 3. LLM Spatial Reasoning Limitations

Observed Behaviors: - AOA-V struggles with termination decisions despite strong exploration - Out-of-bound movement attempts require post-processing intervention - Multitasking (perception + planning + control) dilutes attention

Implications: Monolithic LLM policies insufficient. Need modular architectures with specialized safety modules.

1.8.4 4. Scale and Complexity Challenges

Environmental Factors: - Large search radius (50 units) - Maximum 150 steps for exploration - Diverse semantic regions with varying obstacle densities - Instance-level (not just category-level) object identification

Agent Limitations: - Limited field-of-view from egocentric cameras - No global localization or mapping - Real-time processing constraints

---

1.9 Relevance to Drone Autonomy

1.9.1 1. Semantic Navigation as Open Problem

Current State: - Low success rates (7.30% best) demonstrate fundamental unsolved challenges - Gap between indoor ObjectNav (solved to ~60-80% SR in structured environments) and outdoor aerial ObjectNav

Implications: - Need for novel approaches beyond pure learning - Potential for hybrid architectures combining classical planning with learned components

1.9.2 2. LLM-Based Approaches: Promise and Peril

Strengths: - AOA demonstrates zero-shot generalization capability - Multimodal reasoning shows semantic grounding (26.30% OSR) - Natural language interface enables flexible goal specification

Critical Weaknesses: - 45-65.5% collision rates unacceptable for real systems - Poor termination decisions (OSR-SR gap) - Limited spatial awareness (boundary violations)

Implications: Pure LLM policies insufficient. Need safety-constrained architectures.

1.9.3 3. Control Granularity Insights

Fixed-Step (AOA-F): - Better success metrics (SR, SPL) - Worse collision rate (65.5%) - Simpler, more predictable behavior

Variable-Step (AOA-V): - Better exploration (OSR) - Lower collision rate (45.0%) - Adaptive to environmental context

Implications: Constrained control improves reliability but reduces adaptability. Optimal policy may need context-dependent switching.

1.9.4 4. Safety-Performance Tension

Observed Tradeoff: - Methods achieving higher SR also show higher collision rates - Exploration (OSR) inversely correlated with path efficiency (SPL) - No baseline successfully balances safety, efficiency, and task completion

Implications: Need explicit safety constraints that don't sacrifice exploration capability.

1.10 Connection to Ontology-Constrained RL

1.10.1 Proposed Solution Framework

Problem: UAV-ON reveals that pure learning approaches (LLM, CLIP-based) exhibit: 1. High collision rates (safety failure) 2. Poor semantic-to-action grounding 3. Inefficient exploration strategies

Hypothesis: Ontology-constrained Reinforcement Learning could address these gaps:

1.10.2 1. Ontology for Safety Constraints

Mechanism: - Formal ontology defines safe state space and action preconditions - Example: `canExecute(moveForward, state) ← minDepth(state.frontView) > safetyThreshold` - Hard constraints prevent collision-prone actions a priori

Expected Impact: - Reduce collision rates from 37-65% → <10% - Maintain exploration capability (OSR) while improving safety - Explicit reasoning about obstacle proximity

1.10.3 2. Ontology for Semantic Grounding

Mechanism: - Object-scene co-occurrence encoded in ontology knowledge base - Example: `likelyLocation(bench, park)  likelyLocation(boat, waterArea)` - Guides exploration toward semantically plausible regions

Expected Impact: - Improve exploration efficiency (reduce DTS) - Better handle instance-level descriptions via ontological reasoning - Close OSR-SR gap through improved termination logic

1.10.4 3. RL for Optimal Navigation

Mechanism: - RL learns exploration strategy within ontology-constrained action space - Reward shaping based on: - Distance reduction to target - Semantic region relevance (from ontology) - Collision avoidance (enforced by ontology constraints)

Expected Impact: - Achieve adaptive control (like AOA-V) with safety guarantees (better than AOA-F) - Learn efficient navigation policies while respecting hard constraints - Generalize across diverse environments via ontological abstractions

1.10.5 4. Modular Architecture Benefits

Compared to AOA: - AOA: Monolithic LLM handles perception + planning + control + termination - Ontology-RL: Specialized modules with formal interfaces - Perception → ontology-based scene understanding - Planning → RL policy constrained by ontology - Control → parameterized actions validated by safety ontology - Termination → ontological goal satisfaction checking

Expected Impact: - Reduce cognitive load per component - Enable formal verification of safety properties - Improve interpretability and debugging

---

1.11 Benchmark Utility

1.11.1 For Ontology-Constrained RL Development

1. Training Environment: - 10,000 episodes with ground-truth positions and A* shortest paths - Supports both RL (interaction-based) and imitation learning (demonstration-based) - Voxelized 3D grids enable discrete ontological state representations

2. Evaluation Protocol: - Standardized metrics (SR, OSR, DTS, SPL) enable direct comparison with baselines - Mix of familiar/novel scenes tests generalization - Instance-level object descriptions stress semantic reasoning

3. Safety Analysis: - Collision rate tracking essential for real-world deployment assessment - Trajectory visualization in high-fidelity Unreal environments - Provides failure cases for ontology refinement

---

1.12 Future Research Directions

1.12.1 Identified by Paper

1. **Improved Termination Logic**: Separate exploration from stopping decisions
2. **Hierarchical Planning**: Multi-scale reasoning (region → local → object)
3. **Active Perception**: Adaptive viewpoint selection for semantic disambiguation
4. **Sim-to-Real Transfer**: Bridging simulation-reality gap for outdoor deployment

1.12.2 Additional Opportunities (Ontology-Constrained RL Context)

1. **Formal Safety Verification**: Prove ontology-constrained policies satisfy collision-free properties
2. **Semantic Map Building**: Integrate SLAM with ontological object categorization
3. **Multi-Agent Coordination**: Extend to cooperative search with shared ontology
4. **Continual Learning**: Update ontology from failed episodes (e.g., new object-scene associations)

1.13 Experimental Comparisons with Related Work

1.13.1 ObjectNav Benchmarks (Table 1)

Benchmark	Viewpoint	Task	Goal Type	Goal Specification
AI2-THOR	Ground	ObjNav	Category	Category Label
Gibson	Ground	ObjNav	Category	Category Label
RoboTHOR	Ground	ObjNav	Category	Category Label
HM3D	Ground	ObjNav	Category	Category Label
GeoText	Aerial	VLN	Location	Movement Instruction
AerialVLN	Aerial	VLN	Location	Movement Instruction
CityNav	Aerial	VLN	Location	Movement Instruction
TravelUAV	Aerial	VLN	Location	Movement Instruction

Benchmark	Viewpoint	Task	Goal Type	Goal Specification
OpenFly	Aerial	VLN	Location	Movement Instruction
UAV-ON	**Aerial**	**ObjNav**	**Instance**	**Semantic Instruction**

Key Distinctions: - Only aerial ObjectNav benchmark (vs. ground-based or aerial VLN) - Instance-level goals (vs. category-level) - Semantic instructions (vs. step-by-step movement commands)

1.13.2 Aerial Navigation Evolution

Early Work (conventional): - Map-based planning - Visual navigation with GPS

Recent Work (VLN paradigm): - AerialVLN: First UAV VLN benchmark - GeoText-1652, CityNav: Geospatial grounding - TravelUAV, OpenFly: Scalable environments with trajectory datasets

UAV-ON Contribution: - Removes step-by-step dependency - Requires autonomous goal interpretation - Physical action execution (vs. teleport-based)

1.14 Implementation Details

1.14.1 Simulation Platform

Engine: Unreal Engine + Microsoft AirSim - High-fidelity 3D rendering - Realistic physics simulation - Large-scale environment support

Sensor Configuration: - **RGB cameras**: 4 views (front, left, right, down), synchronized - **Depth cameras**: Corresponding depth maps for all RGB views - **Resolution**: Consistent across all views (specific values not detailed in paper) - **Field of View**: Consistent (specific angles not detailed)

Action Execution: - Physical movement with continuous collision detection - Parameterized control magnitudes - Pose updates follow selected action semantics

1.14.2 Code and Data Availability

- **GitHub Repository**: https://github.com/Kyaren/UAV_ON
- **Benchmark Environments**: 14 Unreal Engine scenes
- **Annotations**: 1,270 target objects with semantic instructions
- **Training Data**: 10,000 episodes with A* shortest paths
- **Test Data**: 1,000 episodes with generalization splits

1.15 Citation

```
@article{xiao2025uavon,
  title={UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents},
  author={Xiao, Jianqiang and Sun, Yuexuan and Shao, Yixin and Gan, Boxi and Liu, Rongqiang and
  journal={arXiv preprint arXiv:2508.00288},
  year={2025}
}
```

1.16 Summary

UAV-ON establishes aerial ObjectNav as a distinct and challenging problem domain, revealing fundamental limitations in current approaches:

1. **Low success rates (7.30% best)** demonstrate that open-world semantic navigation remains unsolved
2. **High collision rates (37-65%)** show critical safety gaps unsuitable for real deployment
3. **OSR-SR gap** highlights difficulty of combining exploration with precise termination
4. **LLM approaches** show promise (zero-shot generalization) but suffer from safety failures and spatial reasoning limitations

The benchmark provides strong motivation for **ontology-constrained RL** approaches that can: - Enforce safety through formal constraints (address collision rates) - Leverage semantic knowledge for efficient exploration (improve SR/SPL) - Decompose complex reasoning into modular, verifiable components (close OSR-SR gap) - Learn adaptive policies while respecting hard constraints (combine AOA-V flexibility with AOA-F reliability)

UAV-ON offers a rigorous evaluation platform with standardized metrics, diverse environments, and realistic challenges essential for advancing autonomous aerial systems toward real-world deployment.