ESARBENCH: A BENCHMARK FOR AGENTIC UAV EMBODIED SEARCH AND RESCUE

Anonymous authors
Paper under double-blind review

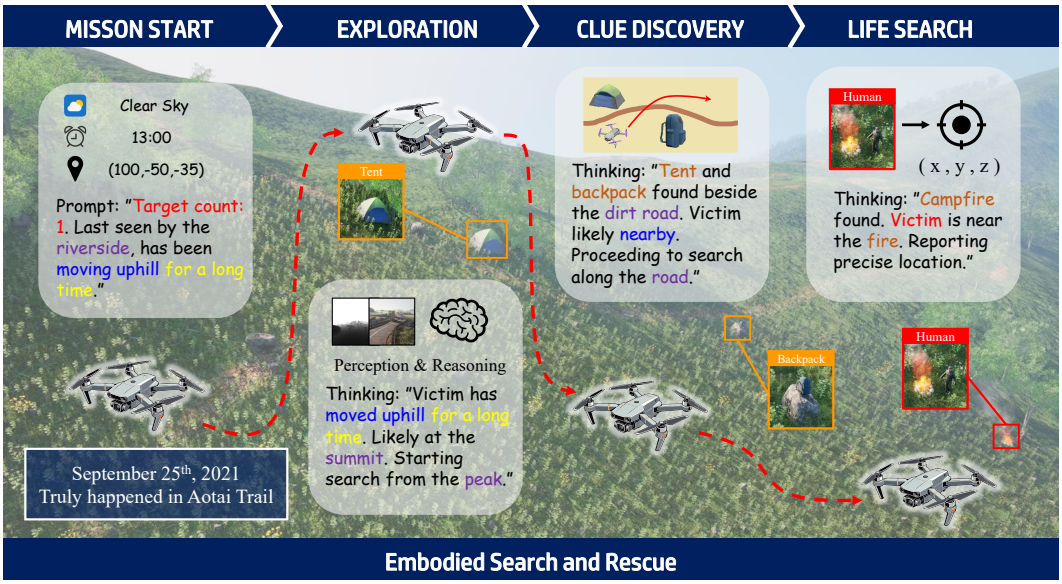


Figure 1: **Illustration of the Embodied Search and Rescue (ESAR) task workflow.** Modeled after real-world cases, the ESAR mission unfolds across four sequential phases: Mission Start, Exploration, Clue Discovery, and Life Search. The UAV agent is initialized with basic environmental conditions and a textual prompt describing the target's last known trajectory. Throughout the flight, the agent utilizes continuous perception and reasoning to identify vital visual clues, such as a tent and a backpack, dynamically adjusting its search strategy based on these findings. Ultimately, the agent locates the human target and reports precise 3D spatial coordinates.

ABSTRACT

The rapid advancement of Multimodal Large Language Models (MLLMs) has empowered Unmanned Aerial Vehicle (UAV) with exceptional capabilities in spatial reasoning, semantic understanding, and complex decision-making, making them inherently suited for UAV Search and Rescue (SAR). However, existing UAV SAR research is dominated by traditional vision and path-planning methods and lacks a comprehensive and unified benchmark for embodied agents. To bridge this gap, we first propose the novel task of **Embodied Search and Rescue (ESAR)**, which requires aerial agents to autonomously explore complex environments, identify rescue clues, and reason about victim locations to execute informed decision-making. Additionally, we present **ESARBench**, the first comprehensive benchmark designed to evaluate MLLM-driven UAV agents in highly realistic SAR scenarios. Leveraging Unreal Engine 5 and AirSim, we construct four high-fidelity, large-scale open environments mapped directly from real-world Geographic Information System (GIS) data to ensure photorealistic landscapes. To rigorously simulate actual rescue operations, our benchmark incorporates dynamic variables including weather conditions, time of day, and stochastic clue placement. Furthermore, we create a dataset of 600 tasks modeled after real-world rescue cases and propose a robust set of evaluation metrics. We evaluate diverse baselines, ranging

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from traditional heuristics to advanced ground and aerial MLLM-based ObjectNav agents. Experimental results highlight the challenges in ESAR, revealing critical bottlenecks in spatial memory, aerial adaptation, and the trade-off between search efficiency and flight safety. We hope ESARBench serves as a valuable resource to advance research on Embodied Search and Rescue domain.

1 INTRODUCTION

The integration of Embodied Artificial Intelligence into Unmanned Aerial Vehicles (UAVs) has emerged as a transformative paradigm(Yao et al., 2025; Anderson et al., 2018), extending the boundaries of robotic autonomy from 2D ground planes(Chang et al., 2017; Chen et al., 2025; Yu et al., 2023) to complex 3D spaces(Tian et al., 2025; Zhao & Lin, 2025; Lin et al., 2025). This trend has driven research into various aerial tasks, such as Aerial Vision Language Navigation (VLN)(Liu et al., 2023; Wang et al., 2025b; Lee et al., 2024) and object goal navigation(Xiao et al., 2025; Ji et al., 2025b; Kim et al., 2025b; Zhang et al., 2026), which aim to equip drones with the capability to observe, understand, and interact with their environments. Among the many applications of intelligent UAVs(Feng et al., 2025; Kim et al., 2025a), Search and Rescue (SAR) stands out as a critical domain. Leveraging their flexibility and extensive field of view, UAVs play an indispensable role in disaster response and wilderness rescue(Yaqoot et al., 2025).

However, two significant gaps hinder the deployment and evaluation of autonomous agents in real-world SAR scenarios.

First, traditional UAV SAR relies on a decoupled stack of classical perception(Sirma et al., 2025; Mishra et al., 2025) and geometric path planning(Dumencic et al., 2025; Döschl et al., 2025). Constrained by a lack of semantic reasoning, these methods depend heavily on narrow, pre-defined operational patterns. This dependency directly results in highly fragmented and task-specific benchmarks, each evaluating custom assumptions rather than generalizable intelligence(Hickling et al., 2025; Bialas et al., 2024; Shuvo et al., 2024). Therefore, they do not provide a unified framework to assess the practical efficacy of UAV agents in real-world SAR tasks.

Moreover, existing embodied UAV researches, such as Aerial VLN(Li et al., 2025b; Cai et al., 2025; Gao et al., 2025), heavily rely on fine-grained, step-by-step linguistic instructions(Wang et al., 2025b; Zhang et al., 2025). These researches lack a high-level, task-centric objective, reducing the agent to a passive instruction follower rather than an active decision-maker. Such setups differ significantly from real-world applications, where instructions are often abstract and goal-oriented. Consequently, current benchmarks fail to comprehensively evaluate a UAV agent's ability to perform long-horizon planning and autonomous exploration under uncertainty(Wu et al., 2025; Xu et al., 2025).

To bridge these gaps, we propose the novel task of **Embodied Search and Rescue (ESAR)**. As illustrated in Figure 1, ESAR requires the agent to demonstrate holistic capabilities, including discovering multi-modal cues, reasoning about environmental semantics, and making autonomous decisions in complex 3D terrains. This task not only meets a vital real-world need but also serves as a benchmark for evaluating the comprehensive capabilities of embodied agents, particularly their potential for task transferability and generalization in open-world settings.

Additionally, to facilitate further research in this domain, we introduce the **ESARBench**, a high-fidelity simulation platform coupled with a comprehensive evaluation framework for UAV agents. Our primary design objective is to minimize the visual sim-to-real gap, ensuring that agent behaviors learned in simulation are robust and transferable to physical reality. To achieve this, we employ Unreal Engine 5 (UE5) for its high-fidelity rendering capabilities and AirSim(Shah et al., 2017) for accurate flight dynamics and physics simulation.

In this framework, we meticulously constructed four large-scale environments based on real-world Geographic Information System (GIS) data and topographical characteristics. These scenarios replicate four distinct and challenging terrains in China, chosen for their representativeness in real-world SAR incidents: Aotai (High Mountain), Lop Nur (Desert), K2 (Snowy Peaks), and Dapeng (Coast). This diversity ensures that the benchmark tests agents across a broad spectrum of topological and visual conditions. Table 1 shows the comparison of ESARBench with existing works.

Table 1: **Comparison of ESARBench with Existing Aerial Robotics Benchmarks.** Our benchmark distinguishes itself through high-fidelity simulation, diverse weather conditions, and large-scale real-world GIS data integration.

Benchmark	Task	Scenario	Simulation	Map Scale	Weather	Real Data
AerialVLN (Liu et al., 2023)	VLN	Urban	UE4	M	4	✗
TravelUAV (Wang et al., 2025b)	VLN	Urban	UE4	M	4	✗
UAV-Flow (Wang et al., 2025a)	VLN/VLA	Urban	UE4	M	4	✗
UAV-ON (Xiao et al., 2025)	ObjectNav	Urban	UE4	M	4	✗
LLM-CRF (Ji et al., 2025a)	SAR	Urban	UE4	S	4	✗
U2UData (Feng et al., 2024)	Perception	Wilderness	UE5	M	7	✓
ESAR-Bench	**ESAR**	**Wilderness**	**UE5**	**L**	**13**	✓

Going beyond static terrain mapping, we aim to reconstruct authentic rescue narratives. Mission-critical clues such as tents, clothes, and illuminants are distributed across the environments based on the specific "Event-Snapshot-Task" generation framework and temporal logic derived from actual rescue cases. Moreover, the benchmark introduces dynamic environmental variables, including shifting weather patterns and time-of-day variations, forcing agents to adapt to changing visibility and lighting conditions.

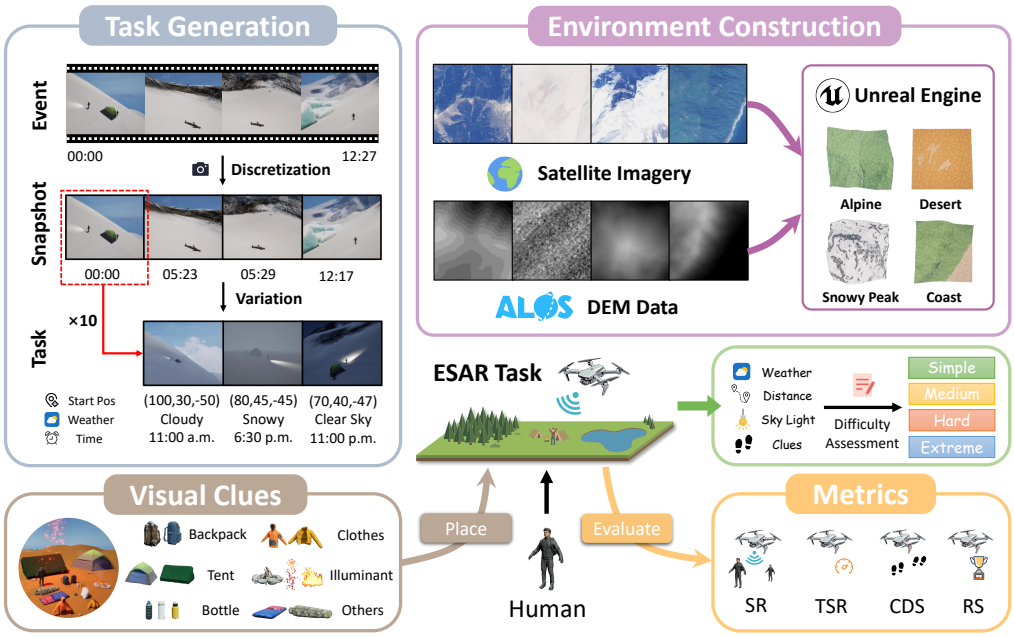


Figure 2: **Overview of the UAV-ESAR Simulator and Benchmark construction pipeline.** The framework consists of two parallel processes: (1) Environment Construction, which utilizes satellite imagery and DEM data to reconstruct high-fidelity terrains in Unreal Engine 5. (2) Task Generation, which discretizes continuous real-world SAR events into static time snapshots with varying parameters (weather, time of day, starting location). These elements are combined to deploy clues and victims, forming complete ESAR tasks stratified across four difficulty levels. Finally, the performance of embodied UAV agents in these tasks is assessed using SR, TSR, CDS, and RS metrics.

Within these dynamic, high-fidelity environments, the UAV agent must process multi-modal sensor inputs to autonomously locate traces of missing persons. To systematically assess their performance in ESAR tasks, we propose a comprehensive suite of evaluation metrics covering Perception Accuracy, Reasoning Capability, and Mission Efficiency. These metrics not only quantify specific SAR performance but also serve to measure the agent's potential for generalizing to other complex embodied tasks.

To establish a baseline for the ESAR task, we evaluate a diverse set of methods, ranging from traditional exploration(Yamauchi, 1997; Chaplot et al., 2020; Yokoyama et al., 2024) to advanced ground and aerial MLLM-based VLN and ObjectNav agents(Yin et al., 2025; Hu et al., 2025; Zhang et al., 2026) with 3D spatial memory. Our experimental results underscore the significant challenges in ESAR. The findings reveal that direct transfer of ground policies is insufficient; aerial agents require aerial-adapted perception, complex reasoning, and spatial memory, alongside balancing critical trade-offs between search efficiency and flight safety.

In summary, our contributions are:

- **Task Definition:** We are the first to formally propose the concept of Embodied Search and Rescue (ESAR), a novel task designed for practical SAR scenario.
- **Simulation Platform:** We develop the first high-fidelity simulation platform for ESAR agents, featuring photorealistic terrains, dynamic event-driven scenarios, and rich multi-modal sensor interfaces.
- **Benchmark & Evaluation:** We establish a comprehensive benchmark and evaluation protocol, providing a standardized metric to assess the core cognitive capabilities required for next-generation autonomous UAVs.

2 Related Work

2.1 UAVs in Search and Rescue

Autonomous Search and Rescue (SAR) remains one of the most critical and impactful applications for UAVs. However, existing UAV SAR methodologies heavily relied on traditional perception and geometric path planning(Strand et al., 2026; Döschl et al., 2025). These approaches are often constrained by their dependency on extensive prior knowledge, such as pre-computed environmental maps(Ge et al., 2024; Yang et al., 2025) or predefined probability models(Cleland-Huang et al., 2025; Dumencic et al., 2025), which are difficult to acquire in the unpredictable dynamics of real-world emergencies. Other research streams have focused on isolated computer vision tasks within SAR contexts, such as standalone object detection(Mishra et al., 2025; Sirma et al., 2025) and tracking(Siktar et al., 2025). These vision-only pipelines inherently lack the cognitive functions required for autonomous reasoning, active planning, and high-level decision-making. While Deep Reinforcement Learning (DRL) has been introduced to address some planning limitations(Ramrakhya et al., 2023; Hickling et al., 2025), they still struggle with complex semantic reasoning and abstract thinking. Most recently, several studies have integrated Multimodal Large Language Models (MLLMs) into SAR applications(Ji et al., 2025a; Döschl & Kiam, 2025), but they still rely on fragmented setups that lack a unified evaluation standard. In conclusion, the community lacks a high-fidelity simulation platform capable of comprehensively validating the interactive capabilities of these aerial agents. To systematically address these critical voids, we introduce the novel task of Embodied Search and Rescue (ESAR) and provide a premier simulator and benchmark tailored for comprehensive agent evaluation.

2.2 Embodied Aerial Agents

To advance the deployment of embodied UAVs into real-world applications(Li et al., 2025a; Zhao et al., 2025), a variety of task paradigms have been proposed to evaluate the capabilities of aerial agents. For instance, Aerial Vision Language Navigation (VLN)(Jiang et al., 2025; Cai et al., 2026) extends traditional ground-based VLN into 3D environments, assessing an agent's ability to navigate by grounding natural language instructions and visual observations. Specifically, works such as UAV-Flow(Wang et al., 2025a) and SPF(Hu et al., 2025) focus on short-horizon navigation guided by concise instructions, while benchmarks like AerialVLN(Liu et al., 2023), TravelUAV(Wang et al., 2025b), CityNav(Lee et al., 2024), and IndoorUAV(Liu et al., 2025) emphasize the execution of long-horizon tasks. Beyond instruction following, Aerial Object Navigation has been extensively studied, with UAV-ON(Xiao et al., 2025), APEX(Zhang et al., 2026), and RAVEN(Kim et al., 2025b) establishing foundational baselines. Furthermore, datasets like AeroDUO(Wu et al., 2025) and U2UData(Feng et al., 2024) explore the dynamics of multi-UAV collaboration in embodied tasks(Hou et al., 2024). Concurrently, several offline datasets(Ferrag et al., 2025; Dai et al., 2025) have been

constructed to explicitly evaluate the perception, reasoning, and decision-making capabilities of agents from an aerial perspective. However, these pioneering works largely focus on isolated sub-tasks lacking direct real-world applicability. Consequently, they fail to adequately assess high-level cognitive capabilities such as complex reasoning and spatial memory, which are essential for practical SAR missions.

3 UAV-ESAR Simulator and ESARBench

3.1 Task Definition

In the aerial embodied ESAR task, the UAV agent is required to navigate complex 3D environments and actively discover mission-critical clues to ultimately locate the victim. At each time step t, the agent receives a current visual observation O_t, maintains an internal state S_t, follows a textual prompt P, and leverages historical context H_t. Unlike traditional navigation tasks that simply output flight actions, the ESAR agent must also explicitly recognize and output the semantic and spatial information of newly discovered clues, denoted as M_{t+1}. Thus, the decision-making process is formulated as a joint output:

$$A_{t+1}, M_{t+1} = \pi(O_t, S_t, P, H_t). \tag{1}$$

The core objective of the agent is to collect critical clues and locate trapped victims within the shortest possible time. First, a successful victim localization is formally defined when the Euclidean distance between the agent's predicted victim coordinates C_{pred} and the actual ground truth coordinates C_{gt} is less than or equal to a predefined error threshold E:

$$||C_{\text{pred}} - C_{\text{gt}}||_2 \leq E. \tag{2}$$

Second, the agent's environmental reasoning capability is scored based on its clue discovery. Let $\mathcal{M}_{\text{gt}}$ represent the set of ground-truth clues distributed in the environment, and $\mathcal{M}_{\text{pred}}$ be the set of clues correctly outputted by the agent during the flight. The task requires the agent to maximize the Clue Recall Rate:

$$\text{CRR} = C_{\text{exact}} = \frac{|\mathcal{M}_{\text{pred}} \cap \mathcal{M}_{\text{gt}}|}{|\mathcal{M}_{\text{gt}}|}, \tag{3}$$

which directly serves as a component of the Clue Discovery Score (CDS).

3.2 UAV-ESAR Simulator

3.2.1 Environment Construction.

To authentically replicate real-world SAR scenarios, we selected four geographical hotspots in China renowned for high frequencies of rescue incidents, meticulously mapping to four representative geomorphologies: the Aotai Trail (Alpine), Lop Nur (Desert), K2 (Snowy Peak), and the Dapeng Peninsula (Coastal cliffs). Based on these, as shown in Figure 3, we constructed four large-scale, open-world simulation environments spanning 2km x 2km, 2km x 2km, 3km x 3km, and 5km x 5km. The UAV-ESAR Simulator precisely maps ALOS PALSAR 12m Digital Elevation Model (DEM) data into UE5, generating natural landscapes that strictly adhere to real-world topographical features. By integrating UE5's high-fidelity rendering pipeline with the flight dynamics of the AirSim-Colosseum plugin, the UAV-ESAR Simulator achieves good performance in SAR simulation.

3.2.2 Scenario Reproduction.

To reconstruct real-world rescue operations, the simulator deploys victims and 12 types of mission-critical clue models—including tents, backpacks, discarded clothing, campfires, and signal flares—at various strategic locations. Crucially, the spatial distribution and contextual placement of these elements are rooted in historical, real-world SAR incidents reported in their respective environments.

3.2.3 Sensors and Environmental Dynamics.

The simulator provides a comprehensive suite of UAV sensors, encompassing an IMU, GPS, LiDAR, alongside multi-view RGB imagery and depth maps. Furthermore, UAV-ESAR supports highly

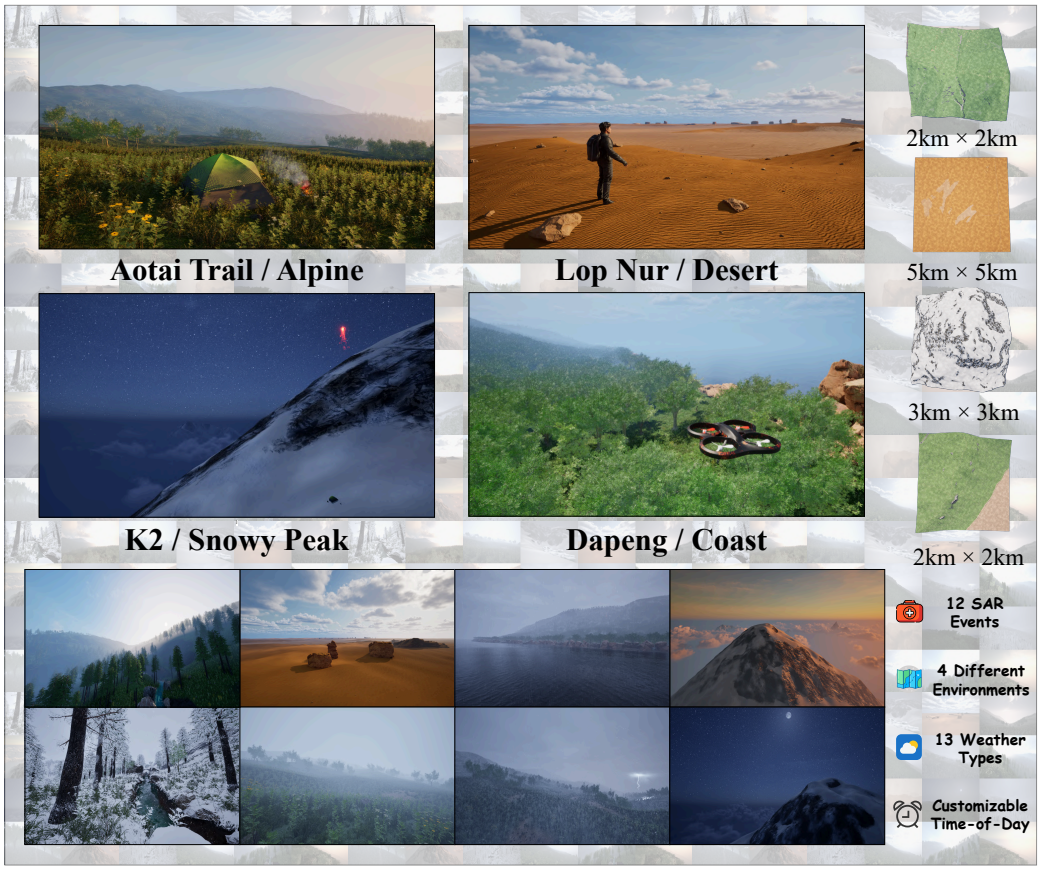


Figure 3: **Environment Construction and Scenario Variations.** The simulation environments are constructed by integrating real-world GIS data to ensure high terrain fidelity. The figure illustrates four distinct geographic environments with varying physical scales, ranging from 2km × 2km to 5km×5km. The platform also features dynamic environmental configurations, supporting 13 different weather types and customizable time-of-day settings.

customizable configurations for weather and time of day. Depending on the specific task, the weather can be dynamically altered among 13 distinct types tailored to the specific environmental climate. Importantly, under specific meteorological conditions, the simulated natural landscapes undergo corresponding physical state changes, such as dynamic snow accumulation, water puddles, and dust coverage—thereby comprehensively testing the perceptual robustness of the embodied agents.

3.3 ESAR Benchmark and Evaluation

3.3.1 Task Data Generation.

As demonstrated in Figure 2, The ESARBench dataset is constructed through a structured, three-tier hierarchical generation framework: Event-Snapshot-Task. An Event represents a complete, longitudinal real-world search and rescue incident that unfolds over an extended period. To ensure experimental fairness, stringent control of variables, and algorithmic reproducibility, we discretize each continuous Event into multiple static time snapshots. Within any given snapshot, the spatial distribution of the victims and clues remains stationary, representing a specific developmental stage of the overall rescue timeline. Finally, from each snapshot, we instantiate multiple distinct Tasks by randomly sampling combinations of environmental and initialization parameters, specifically the time of day, weather conditions, and the UAV's starting location. In total, the ESARBench comprises 12 Events, 60 snapshots, and 600 unique tasks. Figure 7 shows four representative event examples.

3.3.2 Dataset Stratification.

Figure 4 shows the statistics of our task dataset. We stratify the tasks within the ESARBench into four distinct difficulty tiers: Simple, Medium, Hard, and Extreme. This difficulty rating is comprehensively quantified based on a confluence of factors, including weather severity, sky illumination, the average Euclidean distance between the initial starting point and the targets, and the presence of critical clues. Detailed criteria for this difficulty formulation are provided in Appendix A.

3.3.3 Evaluation Metrics.

To systematically assess the performance of UAV agents, we employ four comprehensive metrics:

- **Success Rate (SR):** The ratio of the number of victims successfully located by the UAV to the total number of victims in the environment. We employ the Hungarian algorithm to compute the optimal bipartite matching between the agent's predicted coordinates and the ground truth locations.

$$\text{SR} = \frac{N_{\text{found}}}{N_{\text{total}}}. \tag{4}$$

- **Time-weighted Success Rate (TSR):** A metric that simultaneously evaluates the localization success rate and mission efficiency, where T is the time taken and T_{max} is the maximum allowable task time related to the map.

$$\text{TSR} = \max\left(0, SR \times \left(1 - \frac{T}{T_{\text{max}}}\right)\right). \tag{5}$$

- **Clues Discover Score (CDS):** A comprehensive metric evaluating the UAV's ability to discover mission-critical clues. CDS equally weights pure spatial localization (C_{loc}) and strict exact matching (C_{exact}). C_{exact} requires both spatial proximity (within threshold E) and semantic correctness verified by an LLM evaluator.

$$\text{CDS} = 0.5 \cdot \frac{C_{\text{loc}}}{C_{\text{total}}} + 0.5 \cdot \frac{C_{\text{exact}}}{C_{\text{total}}}, \tag{6}$$

- **Rescue Score (RS):** A holistic metric designed to evaluate the overall capability of the agent in the ESAR task. It balances the primary objective (finding victims) with safe task completion, semantic exploration (finding clues), and temporal efficiency (E_t). The variable $I_{\text{safe}} \in \{0, 1\}$ equals 1 if the agent safely completes the mission, and 0 otherwise. We empirically set the weights to $W_{\text{safe}} = 0.1$, $W_{\text{base}} = 0.3$, $W_{\text{time}} = 0.3$, and $W_{\text{clue}} = 0.3$. The RS is formulated as:

$$\text{RS} = [W_{\text{safe}} \times I_{\text{safe}}] + [\text{SR} \times (W_{\text{base}} + W_{\text{time}} \times E_t)] + [W_{\text{clue}} \times \text{CDS}]. \tag{7}$$

Table 2: The performance of different baseline methods in victim searching tasks, including basic methods, ObjectNav methods without MLLMs, and MLLM-based agents adapted from ground or aerial navigation.

Method	Simple		Medium		Hard		Extreme		Overall	
	SR↑	TSR↑	SR↑	TSR↑	SR↑	TSR↑	SR↑	TSR↑	SR↑	TSR↑
Random	4.68	4.44	4.24	3.87	1.04	0.98	0.00	0.00	2.65	2.47
FBE	13.33	5.64	5.83	0.86	10.52	1.54	2.31	0.00	8.19	2.05
Pure-MLLM	13.21	6.57	0.00	0.00	1.04	0.83	0.00	0.00	3.45	1.80
SemExp	3.23	0.00	9.70	0.55	8.33	3.01	5.56	1.27	6.83	1.21
VLFM	12.10	3.98	10.51	1.69	6.35	2.53	6.96	5.12	9.12	3.17
NavGPT	3.23	1.87	4.65	1.78	6.35	1.45	10.56	0.00	5.92	1.36
UniGoal	8.06	1.75	1.82	0.00	8.96	1.02	7.50	0.10	6.47	0.74
SPF	17.33	1.21	4.85	0.96	7.50	0.93	5.49	0.59	8.84	0.94
APEX	27.83	1.43	10.91	0.99	6.20	0.46	10.83	0.54	13.89	0.87

4 EXPERIMENT

4.1 BASELINES

To establish baselines for the novel ESAR task, we adapt representative methods from several established embodied intelligence settings. These baselines cover several types: basic exploration and direct MLLM control, ObjectNav methods without MLLMs, and MLLM-based agents from ground VLN, ground ObjectNav, aerial VLN, and aerial ObjectNav. For fair comparison, all baselines are connected to the same AirSim interface and use the same four-camera YOLO-World(Cheng et al., 2024) RGB-D module for clue and victim reporting. For the MLLM components, we use Qwen3.5-Plus (Bai et al., 2025) as our base model.

- **Random (Basic):** A lower-bound policy that uniformly samples discrete UAV actions without mapping or task reasoning.
- **FBE(Yamauchi, 1997) (Basic):** A classical frontier-based exploration method using a BEV occupancy map and FMM local planning.
- **Pure-MLLM (Basic):** A direct MLLM-control baseline that maps the front-view observation and task prompt to discrete UAV actions.
- **SemExp(Chaplot et al., 2020) (Ground ObjectNav, non-MLLM):** A semantic-exploration baseline using a BEV semantic map, frontier selection, and FMM planning. We adapt it to ESAR by replacing the original learned global policy with a zero-shot frontier heuristic.
- **VLFM(Yokoyama et al., 2024) (Ground ObjectNav, non-MLLM):** A vision-language frontier baseline that ranks frontiers with a value map. It uses image-text matching to estimate which frontier is most relevant to the task prompt.
- **NavGPT(Zhou et al., 2024) (Ground VLN, MLLM):** A ground VLN-style MLLM agent that selects safe actions from multi-view captions, history, and state. It reasons over textual scene descriptions rather than metric maps, serving as a transfer baseline for indoor instruction-following agents in aerial search.
- **UniGoal(Yin et al., 2025) (Ground ObjectNav, MLLM):** A ground ObjectNav method that uses goal and scene graph matching to guide exploration. It identifies target objects and spatial hints from the SAR prompt, then prioritizes searching in areas that match the current scene graph.
- **SPF(Hu et al., 2025) (Aerial VLN, MLLM):** An aerial VLN-style agent that uses VLM point prediction and depth back-projection for point-and-fly control. Unlike ground VLN baselines, it directly predicts an image-space flight direction and converts it into UAV motion commands.
- **APEX(Zhang et al., 2026) (Aerial ObjectNav, MLLM):** An aerial ObjectNav agent using VLM-guided 3D voxel maps and reward-based discrete action selection. It explicitly models attraction, exploration, and obstacles in 3D space, representing a UAV-specific ObjectNav strategy with spatial memory.

4.2 RESULTS

4.2.1 OVERALL PERFORMANCE AND AERIAL ADAPTATION.

Tables 2 and 3 report the performance of all baselines on victim search, clue discovery, and the comprehensive rescue score. APEX achieves the strongest overall results, ranking first on SR, CDS, and RS with overall scores of 13.89, 4.14, and 13.45, respectively. SPF obtains the second-best RS of 13.12 and remains competitive across both victim search and clue discovery. The advantage of SPF and APEX over the ground MLLM baselines, including NavGPT and UniGoal, indicates that aerial adaptation is crucial for ESAR. UAV agents must handle large-scale outdoor viewpoints, 3D motion, and search-oriented exploration rather than only transferring ground navigation policies.

Table 3: Baseline experiment results in clue discovery tasks and comprehensive scores (incorporating SR, TSR, and CDS).

Method	Simple		Medium		Hard		Extreme		Overall	
	CDS↑	RS↑	CDS↑	RS↑	CDS↑	RS↑	CDS↑	RS↑	CDS↑	RS↑
Random	1.25	9.55	3.49	13.08	1.04	7.48	0.0	8.75	1.51	9.81
FBE	3.65	13.42	5.71	10.43	3.61	9.39	0.27	6.16	3.40	9.97
Pure-MLLM	5.21	11.78	2.28	8.20	1.82	6.11	0.00	7.08	2.39	8.26
SemExp	2.14	7.72	5.50	11.85	1.80	8.32	0.00	7.88	2.47	9.05
VLFM	4.01	12.44	4.24	12.09	1.36	7.13	1.96	10.30	2.92	10.50
NavGPT	4.14	10.47	2.93	11.21	1.69	9.72	4.86	12.54	3.30	10.89
UniGoal	4.71	10.76	1.83	8.92	3.68	9.10	1.13	8.03	2.94	9.27
SPF	5.89	16.94	2.22	12.35	2.94	11.54	3.08	11.49	3.53	13.12
APEX	7.35	18.65	4.06	13.46	1.36	9.59	3.96	12.10	4.14	13.45

4.2.2 Semantic Reasoning in Clue Discovery.

MLLM-based methods show a clear advantage in CDS, with the four adapted MLLM baselines averaging 3.48 overall CDS, compared with 2.70 for the non-MLLM ObjectNav baselines. This trend suggests that semantic understanding and reasoning are important for ESAR, where clues are not merely visual objects but task-relevant evidence that must be interpreted under the rescue prompt. The results suggest that MLLM reasoning is only effective when built into embodied search structures: Pure MLLM models struggle, but SPF and APEX perform better because they add UAV-specific actions and spatial mapping to the model's guidance.

4.2.3 Efficiency Trade-off and Benchmark Difficulty.

Despite their strong RS, SPF and APEX have low TSR scores of 0.94 and 0.87, respectively. This reveals two limitations. First, these agents lack an active mechanism for judging when a multi-target ESAR mission has been sufficiently completed. Second, they struggle to improve efficiency while preserving broad search ability. More generally, all baselines remain far from solving the ESAR Task. Even the best SR and CDS are only 13.89 and 4.14, and the performance gap between the strongest methods and basic baselines is limited. This indicates that ESARBench demands a high level of integration between perception, semantic reasoning, and decision-making.

5 Conclusion

In this paper, we introduce the novel paradigm of Embodied Search and Rescue (ESAR), bridging the gap between embodied intelligence and critical real-world aerial missions. To support this, we develop the first high-fidelity simulation platform designed for UAV ESAR agents, and a comprehensive benchmark to quantitatively evaluate their performance. Our extensive baseline experiments underscore the profound difficulty of the ESAR task. The results demonstrate that autonomous rescue requires an agent's abilities of active perception, semantic reasoning, long-horizon planning, spatial memory, and robust decision-making within complex 3D environments. In the future, we will continuously upgrade the ESARBench, including integrate additional sensor modalities and introduce a broader spectrum of complex task configurations. Furthermore, we will focus on designing more advanced embodied architectures to drive the evolution of fully autonomous aerial rescue systems.

AI USE STATEMENT

We used generative AI tools only for recommended-disclosure tasks: language polishing of the manuscript, formatting, and literature retrieval. We have not used generative AI tools for any required-disclosure task: no synthetic data was generated by AI, and no AI tool was used to develop the conceptual framework, to formulate mathematical claims or proofs, to design the methodology or experiments, to implement methods, or to interpret results. One author checked every AI-assisted edit against the original text. We take responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

This work studies autonomous UAV search and rescue entirely in simulation; no human subjects, no real flights, and no personal data were involved. The real-world rescue incidents used to design our task scenarios are drawn from public reports and no undisclosed personal information about the individuals involved is released. All third-party data and assets are used under their respective licenses with attribution, and our released benchmark contains no personal data.

REFERENCES

Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *CoRR*, abs/1807.06757, 2018. URL https://arxiv.org/abs/1807.06757.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://arxiv.org/abs/2511.21631.

Julian Bialas, Mario Döller, Simone Walch, Michiel Jan Van Veelen, and Abraham Mejia-Aguilar. Optimizing multi-agent coverage path planning UAV search and rescue missions with prioritizing deep reinforcement learning. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2024, Bangkok, Thailand, December 10-14, 2024*, pp. 85–90. IEEE, 2024. doi: 10.1109/ROBIO64047.2024.10907496.

Hengxing Cai, Jinhan Dong, Jingjun Tan, Jingcheng Deng, Sihang Li, Zhifeng Gao, Haidong Wang, Zicheng Su, Agachai Sumalee, and Renxin Zhong. Flightgpt: Towards generalizable and interpretable UAV vision-and-language navigation with vision-language models. *CoRR*, abs/2505.12835, 2025. doi: 10.48550/ARXIV.2505.12835. URL https://arxiv.org/abs/2505.12835.

Hengxing Cai, Yijie Rao, Ligang Huang, Zanyang Zhong, Jinhan Dong, Jingjun Tan, Wenhao Lu, and Renxin Zhong. Airnav: A large-scale real-world UAV vision-and-language navigation dataset with natural and diverse instructions. *CoRR*, abs/2601.03707, 2026. doi: 10.48550/ARXIV.2601.03707. URL https://arxiv.org/abs/2601.03707.

Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 667–676. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00081.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian D. Reid, and Liang Wang. Constraint-aware zero-shot vision-language navigation in continuous environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(11):10441–10456, 2025. doi: 10.1109/TPAMI.2025.3594204.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 16901–16911. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01599.

Jane Cleland-Huang, Pedro Alarcon Granadeno, Arturo Miguel Russell Bernal, Demetrius Hernandez, Michael Murphy, Maureen Petterson, and Walter J. Scheirer. Cognitive guardrails for open-world decision making in autonomous drone swarms. *CoRR*, abs/2505.23576, 2025. doi: 10.48550/ARXIV.2505.23576. URL https://arxiv.org/abs/2505.23576.

Shiqi Dai, Zizhi Ma, Zhicong Luo, Xuesong Yang, Yibin Huang, Wanyue Zhang, Chi Chen, Zonghao Guo, Wang Xu, Yufei Sun, and Maosong Sun. Mm-uavbench: How well do multimodal large language models see, think, and plan in low-altitude UAV scenarios? *CoRR*, abs/2512.23219, 2025. doi: 10.48550/ARXIV.2512.23219. URL https://arxiv.org/abs/2512.23219.

Bjorn Döschl and Jane Jean Kiam. Say'n'fly: An llm-modulo online planning framework to automate UAV command and control. In *34th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2025, Eindhoven, Netherlands, August 25-29, 2025*, pp. 1693–1698. IEEE, 2025. doi: 10.1109/RO-MAN63969.2025.11217764.

Bjorn Döschl, Kai Sommer, and Jane Jean Kiam. AUSPEX: an integrated open-source decision-making framework for uavs in rescue missions. *Frontiers Robotics AI*, 12, 2025. doi: 10.3389/FROBT.2025.1583479.

Stella Dumencic, Luka Lanca, Karlo Jakac, and Stefan Ivic. Experimental validation of UAV search and detection system in real wilderness environment. *CoRR*, abs/2502.17372, 2025. doi: 10.48550/ARXIV.2502.17372. URL https://arxiv.org/abs/2502.17372.

Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. U2udata: A large-scale cooperative perception dataset for swarm uavs autonomous flight. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 7600–7608. ACM, 2024. doi: 10.1145/3664647.3681151.

Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. U2udata-2: A scalable swarm uavs autonomous flight dataset for long-horizon tasks. *CoRR*, abs/2509.00055, 2025. doi: 10.48550/ARXIV.2509.00055. URL https://arxiv.org/abs/2509.00055.

Mohamed Amine Ferrag, Abderrahmane Lakas, and Mérouane Debbah. Uavbench: An open benchmark dataset for autonomous and agentic AI UAV systems via llm-generated flight scenarios. *CoRR*, abs/2511.11252, 2025. doi: 10.48550/ARXIV.2511.11252. URL https://arxiv.org/abs/2511.11252.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Openfly: A versatile toolchain and large-scale benchmark for aerial vision-language navigation. *CoRR*, abs/2502.18041, 2025. doi: 10.48550/ARXIV.2502.18041. URL https://arxiv.org/abs/2502.18041.

Zijian Ge, Jingjing Jiang, and Matthew Coombes. Multi-uav search and rescue in wilderness using smart agent-based probability models. *CoRR*, abs/2411.10148, 2024. doi: 10.48550/ARXIV.2411.10148. URL https://arxiv.org/abs/2411.10148.

Thomas Hickling, Maxwell Hogan, Abdulla Tammam, and Nabil Aouf. Deep reinforcement learning based autonomous decision-making for cooperative uavs: A search and rescue real world application. *CoRR*, abs/2502.20326, 2025. doi: 10.48550/ARXIV.2502.20326. URL https://arxiv.org/abs/2502.20326.

Yukai Hou, Jin Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. UAV swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Trans. Intell. Veh.*, 9(1): 568–578, 2024. doi: 10.1109/TIV.2023.3316196.

Chih Yao Hu, Yang-Sen Lin, Yuna Lee, Chih-Hai Su, Jie-Ying Lee, Shr-Ruei Tsai, Chin-Yang Lin, Kuan-Wen Chen, Tsung-Wei Ke, and Yu-Lun Liu. See, point, fly: A learning-free VLM framework for universal unmanned aerial navigation. *CoRR*, abs/2509.22653, 2025. doi: 10.48550/ARXIV.2509.22653. URL https://arxiv.org/abs/2509.22653.

Kailun Ji, Xiaoyu Hu, Xinyu Zhang, and Jun Chen. An llm-based framework for human-swarm teaming cognition in disaster search and rescue. *CoRR*, abs/2511.04042, 2025a. doi: 10.48550/ARXIV.2511.04042. URL https://arxiv.org/abs/2511.04042.

Yatai Ji, Zhengqiu Zhu, Yong Zhao, Beidan Liu, Chen Gao, Yihao Zhao, Sihang Qiu, Yue Hu, Quanjun Yin, and Yong Li. Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology. *CoRR*, abs/2505.08765, 2025b. doi: 10.48550/ARXIV.2505.08765. URL https://arxiv.org/abs/2505.08765.

Wen Jiang, Li Wang, Kangyao Huang, Wei Fan, Jinyuan Liu, Shaoyu Liu, Hongwei Duan, Bin Xu, and Xiangyang Ji. Longfly: Long-horizon UAV vision-and-language navigation with spatiotemporal context integration. *CoRR*, abs/2512.22010, 2025. doi: 10.48550/ARXIV.2512.22010. URL https://arxiv.org/abs/2512.22010.

Minwoo Kim, Geun Sik Bae, Jinwoo Lee, Woojae Shin, Changseung Kim, Myong-Yol Choi, Heejung Shin, and Hyondong Oh. RAPID: robust and agile planner using inverse reinforcement learning for vision-based drone navigation. *CoRR*, abs/2502.02054, 2025a. doi: 10.48550/ARXIV.2502.02054. URL https://arxiv.org/abs/2502.02054.

Seungchan Kim, Omar Alama, Dmytro Kurdydyk, John Keller, Nikhil Varma Keetha, Wenshan Wang, Yonatan Bisk, and Sebastian A. Scherer. RAVEN: resilient aerial navigation via open-set semantic memory and behavior adaptation. *CoRR*, abs/2509.23563, 2025b. doi: 10.48550/ARXIV.2509.23563. URL https://arxiv.org/abs/2509.23563.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. Citynav: Language-goal aerial navigation dataset with geographic information. *CoRR*, abs/2406.14240, 2024. doi: 10.48550/ARXIV.2406.14240. URL https://arxiv.org/abs/2406.14240.

Nianxin Li, Mao Ye, Lihua Zhou, Song Tang, Yan Gan, Zizhuo Liang, and Xiatian Zhu. Self-prompting analogical reasoning for UAV object detection. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 18412–18420. AAAI Press, 2025a. doi: 10.1609/AAAI.V39I17.34026.

Tianshun Li, Tianyi Huai, Zhen Li, Yichun Gao, Haoang Li, and Xinhu Zheng. Skyvln: Vision-and-language navigation and NMPC control for uavs in urban environments. *CoRR*, abs/2507.06564, 2025b. doi: 10.48550/ARXIV.2507.06564. URL https://arxiv.org/abs/2507.06564.

Fei Lin, Yonglin Tian, Tengchao Zhang, Jun Huang, Sangtian Guan, and Fei-Yue Wang. Airvista-ii: An agentic system for embodied uavs toward dynamic scene semantic understanding. *CoRR*, abs/2504.09583, 2025. doi: 10.48550/ARXIV.2504.09583. URL https://arxiv.org/abs/2504.09583.

Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, and Qi Wu. Aerialvln: Vision-and-language navigation for uavs. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15338–15348. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01411.

Xu Liu, Yu Liu, Hanshuo Qiu, Qirong Yang, and Zhouhui Lian. Indooruav: Benchmarking vision-language UAV navigation in continuous indoor environments. *CoRR*, abs/2512.19024, 2025. doi: 10.48550/ARXIV.2512.19024. URL https://arxiv.org/abs/2512.19024.

Aditya Mishra, Modigari Narendra, Aryan Sinha, and Aakash Kumar. Dynamic backbone optimization of yolov10 for real-time object detection in uav-based search and rescue missions. *IEEE Access*, 13:195975–195986, 2025. doi: 10.1109/ACCESS.2025.3633379.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and RL finetuning for OBJECTNAV. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 17896–17906. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01716.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Marco Hutter and Roland Siegwart (eds.), *Field and Service Robotics, Results of the 11th International Conference, FSR 2017, Zurich, Switzerland, 12-15 September 2017*, volume 5 of *Springer Proceedings in Advanced Robotics*, pp. 621–635. Springer, 2017. doi: 10.1007/978-3-319-67361-5_40.

M. I. R. Shuvo, Nafiul Alam, Awal Ahmed Fime, Hannah Lee, Xiangxu Lin, and Jong-Hoon Kim. A novel large language model (LLM) based approach for robotic collaboration in search and rescue operations. In *50th Annual Conference of the IEEE Industrial Electronics Society, IECON 2024, Chicago, IL, USA, November 3-6, 2024*, pp. 1–6. IEEE, 2024. doi: 10.1109/IECON55916.2024.10905094.

Luka Siktar, Branimir Caran, Bojan Sekoranja, and Marko Svaco. Autonomous UAV navigation for search and rescue missions using computer vision and convolutional neural networks. *CoRR*, abs/2507.18160, 2025. doi: 10.48550/ARXIV.2507.18160. URL https://arxiv.org/abs/2507.18160.

Aykut Sirma, Angelos Plastropoulos, Argyrios Zolotas, and Gilbert Tang. Drespnet: A UAV dataset and yolov8-drn model for aerial instance segmentation of building access points for post-earthquake search-and-rescue missions. *CoRR*, abs/2508.16016, 2025. doi: 10.48550/ARXIV.2508.16016. URL https://arxiv.org/abs/2508.16016.

Sigrid Strand, Thomas Wiedemann, Bram Burczek, and Dmitriy Shutin. Enhancing UAV search under occlusion using next best view planning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 19:1085–1096, 2026. doi: 10.1109/JSTARS.2025.3638881.

Yonglin Tian, Fei Lin, Yiduo Li, Tengchao Zhang, Qiyao Zhang, Xuan Fu, Jun Huang, Xingyuan Dai, Yutong Wang, Chunwei Tian, Bai Li, Yisheng Lv, Levente Kovács, and Feiyue Wang. Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility. *Inf. Fusion*, 122: 103158, 2025. doi: 10.1016/J.INFFUS.2025.103158.

Xiangyu Wang, Donglin Yang, Yue Liao, Wenhao Zheng, Wenjun Wu, Bin Dai, Hongsheng Li, and Si Liu. Uav-flow colosseo: A real-world benchmark for flying-on-a-word UAV imitation learning. *CoRR*, abs/2505.15725, 2025a. doi: 10.48550/ARXIV.2505.15725. URL https://arxiv.org/abs/2505.15725.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic UAV vision-language navigation: Platform, benchmark, and methodology. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b.

Ruipu Wu, Yige Zhang, Jinyu Chen, Linjiang Huang, Shifeng Zhang, Xu Zhou, Liang Wang, and Si Liu. Aeroduo: Aerial duo for uav-based vision and language navigation. *CoRR*, abs/2508.15232, 2025. doi: 10.48550/ARXIV.2508.15232. URL https://arxiv.org/abs/2508.15232.

Jianqiang Xiao, Yuexuan Sun, Yixin Shao, Boxi Gan, Rongqiang Liu, Yanjing Wu, Weili Guan, and Xiang Deng. UAV-ON: A benchmark for open-world object goal navigation with aerial agents. *CoRR*, abs/2508.00288, 2025. doi: 10.48550/ARXIV.2508.00288. URL https://arxiv.org/abs/2508.00288.

Haotian Xu, Yue Hu, Chen Gao, Zhengqiu Zhu, Yong Zhao, Yong Li, and Quanjun Yin. Geonav: Empowering mllms with explicit geospatial reasoning abilities for language-goal aerial navigation. *CoRR*, abs/2504.09587, 2025. doi: 10.48550/ARXIV.2504.09587. URL https://arxiv.org/abs/2504.09587.

Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97 - Towards New Computational Principles for Robotics and Automation, July 10-11, 1997, Monterey, California, USA*, pp. 146–151. IEEE Computer Society, 1997. doi: 10.1109/CIRA.1997.613851.

Guang Yang, Yadong Mo, Chengyu Lv, Ying Zhang, Jian Li, and Shimin Wei. A dual-layer task planning algorithm based on uavs-human cooperation for search and rescue. *Appl. Soft Comput.*, 181:113488, 2025. doi: 10.1016/J.ASOC.2025.113488.

Fanglong Yao, Youzhi Liu, Wenyi Zhang, Zhengqiu Zhu, Chenglong Li, Nayu Liu, Peng Hu, Yuanchang Yue, Kaiwen Wei, Xin He, Xudong Zhao, Zihan Wei, Haotian Xu, Zhiyuan Wang, Gujie Shao, Liu Yang, Dan Zhao, and Yong Yang. Aeroverse-review: Comprehensive survey on aerial embodied vision-and-language navigation. *The Innovation Informatics*, 1(1):100015, 2025. ISSN 3105-8515. doi: 10.59717/j.xinn-inform.2025.100015.

Yasheerah Yaqoot, Muhammad Ahsan Mustafa, Oleg Sautenkov, and Dzmitry Tsetserukou. UAV-VLRR: vision-language informed NMPC for rapid response in UAV search and rescue. *CoRR*, abs/2503.02465, 2025. doi: 10.48550/ARXIV.2503.02465. URL https://arxiv.org/abs/2503.02465.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19057–19066. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01775.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 42–48. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610712.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: leveraging large language models for visual target navigation. In *IROS*, pp. 3554–3560, 2023. doi: 10.1109/IROS55552.2023.10342512.

Daoxuan Zhang, Ping Chen, Xiaobo Xia, Xiu Su, Ruichen Zhen, Jianqiang Xiao, and Shuo Yang. APEX: A decoupled memory-based explorer for asynchronous aerial object goal navigation. *CoRR*, abs/2602.00551, 2026. doi: 10.48550/ARXIV.2602.00551. URL https://arxiv.org/abs/2602.00551.

Weichen Zhang, Chen Gao, Shiquan Yu, Ruiying Peng, Baining Zhao, Qian Zhang, Jinqiang Cui, Xinlei Chen, and Yong Li. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 31292–31309. Association for Computational Linguistics, 2025.

Ganlong Zhao, Guanbin Li, Jia Pan, and Yizhou Yu. Aerial vision-and-language navigation with grid-based view selection and map construction. *CoRR*, abs/2503.11091, 2025. doi: 10.48550/ARXIV.2503.11091. URL https://arxiv.org/abs/2503.11091.

Ji Zhao and Xiao Lin. General-purpose aerial intelligent agents empowered by large language models. *CoRR*, abs/2503.08302, 2025. doi: 10.48550/ARXIV.2503.08302. URL https://arxiv.org/abs/2503.08302.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 7641–7649. AAAI Press, 2024. doi: 10.1609/AAAI.V38I7.28597.