IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

NEUSIS: A Compositional Neuro-Symbolic Framework for Autonomous Perception, Reasoning, and Planning in Complex UAV Search Missions

Zhixi Cai*‡, Cristian Rojas Cardenas*, Kevin Leo*, Chenyuan Zhang*, Kal Backman*, Hanbing Li*, Boying Li, Mahsa Ghorbanali, Stavya Datta, Lizhen Qu, Julian Gutierrez, Alexey Ignatiev, Yuan-Fang Li†, Mor Vered†, Peter J. Stuckey†, Maria Garcia de la Banda† and Hamid Rezatofighi†

***Abstract*—This paper addresses the problem of autonomous UAV search missions, where a UAV must locate specific Entities of Interest (EOIs) within a time limit, based on brief descriptions in large, hazard-prone environments with keep-out zones. The UAV must perceive, reason, and make decisions with limited and uncertain information. We propose NEUSIS, a compositional neuro-symbolic system designed for effective UAV search and navigation in realistic scenarios. NEUSIS integrates neuro-symbolic visual perception, reasoning, and grounding (GRiD) to process raw sensory inputs, maintains a probabilistic world model for environment representation, and uses a hierarchical planning component (SNaC) for efficient path planning. Experimental results from simulated urban search missions using AirSim and Unreal Engine show that NEUSIS outperforms state-of-the-art baselines for both perception and planning. These results demonstrate the effectiveness of our compositional neuro-symbolic approach in handling complex scenarios, making it a promising solution for autonomous UAV systems in search missions.**

***Index Terms*—Aerial Systems: Perception and Autonomy, Recognition, Planning under Uncertainty**

I. Introduction

THE development of autonomous agents capable of safely completing Intelligence, Surveillance, and Reconnaissance (ISR) missions in complex environments presents significant challenges [6]. Uncrewed Aerial Vehicles (UAVs) are increasingly utilized in these missions due to their ability to cover large areas and access hazardous locations with minimal risk to human life [22]. However, designing fully autonomous UAV systems for ISR tasks in unpredictable and complex environments with uncertain knowledge, is a formidable challenge.

In this paper, we focus on complex scenarios in which a UAV must, within a designated time limit, autonomously search for a number of specific entities of interest (EOIs) based on brief descriptions, e.g., find "*a red SUV vehicle*" or "*a pedestrian carrying a blue umbrella*", in a large suburban or urban environment that may contain hazards or keep-out zones (KOZs). These hazard zones represent significant risks that the UAV must carefully avoid while efficiently searching through the given areas of interest (AOIs) [27]. To successfully operate in such scenarios, an autonomous UAV must actively and reliably perceive the environment from onboard sensory measurements, reason about the environment, and make decisions based on the mission description and partial or uncertain information about the surroundings.

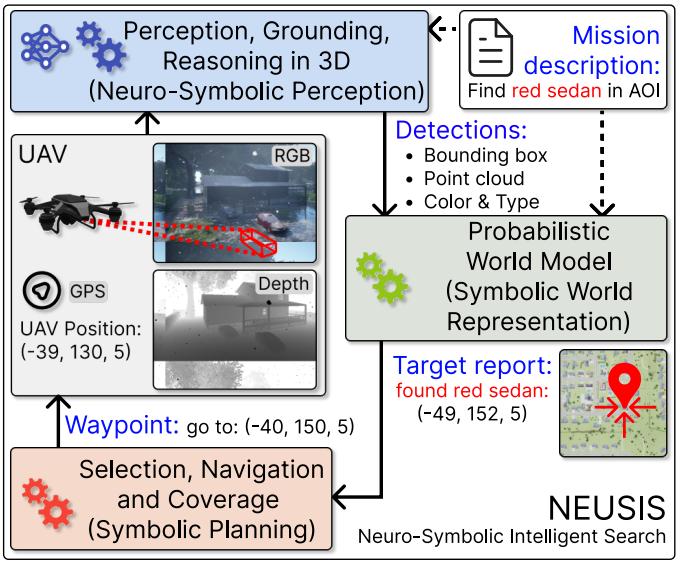


Fig. 1: Overview of NEUSIS. Neuro-symbolic *Perception, Grounding, Reasoning in 3D* (GRiD); Symbolic Probabilistic World Model; and *Selection, Navigation and Coverage* (SNaC) components autonomously complete UAV search missions by processing sensor inputs to find targets, such as the red sedan required by the mission description.

Recent advances in Large Multimodal Models (LMMs) [26], [29], [30] have shown promise in different robotics tasks. However, their reliance on diverse, large-scale datasets for training as monolithic end-to-end models imposes significant computational demands. These models often lack interpretability when they fail and struggle to generalize beyond their training domains, especially in adversarial settings [12], [33]. Furthermore, LMMs lack explicit components to model the world state or update their knowledge, which is crucial for complex tasks like searching in unconstrained environments. As an alternative, many autonomous robotics systems employ compositional approaches [16], [20], [42], integrating explicit

Manuscript received: March, 10, 2025; Revised June, 2, 2025; Accepted July, 6, 2025.

This paper was recommended for publication by Editor Abhinav Valada upon evaluation of the Associate Editor and Reviewers' comments.

*These authors contributed equally to this work as the joint first authors

†These authors contributed equally to this work as the joint last authors

‡Corresponding author (mor.vered@monash.edu)

All authors except Julian Gutierrez are with Monash University, Australia. Julian Gutierrez is with University of Sussex, UK

This work is supported by the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) program under award number FA8750-23-2-1016 and the Australian Research Council Discovery Project ARC DP2020102427

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Fig. 2: Screenshots from the Neighborhood environment illustrating different real-world challenges for UAVs.

perception and planning to perform tasks. These systems use neural-based perception models to process sensory data into abstract representations like segmentation, detection, or captions, which a neural planner then uses for navigation. While these approaches offer better interpretability and generalizability than monolithic models, they still lack explicit visual reasoning and world state representation. Neural planners require substantial training data, are task-constrained, and may be less efficient than model-based or symbolic planners. They also remain vulnerable to adversarial conditions, making them less suitable for search problems in unconstrained environments, the focus of this paper.

A viable baseline for such search problems is to use state-of-the-art (SoTA) vision or multimodal language models [4], [15], [17] to translate mission specifications and sensor data to a robust abstract level. This can then be integrated with model-based planners [19], [21], [30], [39] that offer better generalizability, robustness, and efficiency. However, this approach still lacks explicit visual reasoning and a persistent world model, which limits its ability to maintain an interpretable representation of the environment and make informed decisions.

To address these limitations, we introduce NEUSIS (**Neu**ro-**S**ymbolic **I**ntelligent **S**earch), a novel compositional neuro-symbolic framework comprised of three main components (see Figure 1): *(i)* a Neuro-symbolic component for *Perception, Grounding and Reasoning in 3D*, GRiD, which handles the perception, visual reasoning and localization of entities of interest in a 3D world using UAV visual sensors; *(ii)* a *Probabilistic (Symbolic) World Model*, which refines the potentially noisy outputs from GRiD and updates the belief about entities of interest based on world knowledge and probabilistic belief, maintaining a coherent and interpretable representation of the environment that enables robust reasoning and decision-making; *(iii)* a *Hierarchical Model-Based (Symbolic) Planning* component, SNaC, which uses high-level planning to determine the overall search strategy, mid-level planning for navigating to an allocated area, and low-level planning to efficiently and effectively search within allocated areas while avoiding obstacles. Although individual modules such as GroundingDINO or A* are well-known, integrating them into a closed-loop, UAV search system that interprets free-form natural-language prompts in real time is far from trivial. We designed a modular, neuro-symbolic architecture that tightly couples a multi-modal visual perception and reasoning module (GRiD), a novel probabilistic world model, and a hierarchical symbolic planner (SNaC). The carefully engineered interfaces and real-time coordination across modules are key innovations.

We evaluate NEUSIS on a search mission benchmark developed by Keno *et al.* [14] based on the AirSim [31] simulator for Unreal Engine as part of the DARPA-Assured Neuro-symbolic Reasoning (ANSR) program.[1] This benchmark presents complex scenarios, including various challenging environmental settings (e.g., different weather conditions). Our results demonstrate that NEUSIS significantly outperforms a strong compositional baseline in terms of localization performance, and navigation efficiency, marking a significant advance in end-to-end autonomous UAV systems. To summarize, we propose NEUSIS, a novel compositional neuro-symbolic framework for UAV systems, integrating three essential components with the following contributions:

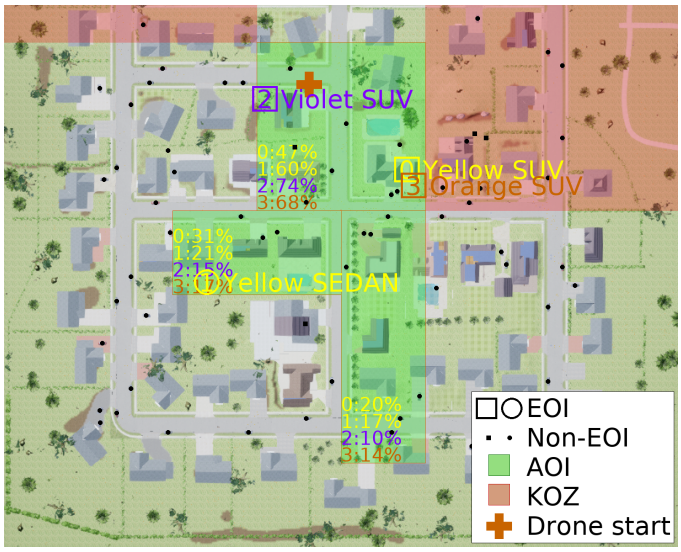


Fig. 3: Example mission scenario with 4 entities of interest (EOIs) across 3 areas of interest (AOIs). Prior likelihood of EOI presence is shown in the bottom left corner of each AOI.

1) Our perception module GRiD integrates camera, depth, and GPS data to ground objects in a 3D grid using a customized state-of-the-art multimodal compositional model, enabling robust visual grounding and interpretable reasoning.
2) Our probabilistic world model refines noisy predicted outputs from GRiD by incorporating prior knowledge through Bayesian filtering and temporal accumulation, ensuring a dynamic and accurate environmental representation for informed decision-making.
3) Our hierarchical planner SNaC leverages symbolic reasoning and constraint optimization for AOI selection, obstacle-aware navigation, and fine-grained search execution.

II. Related Work

A main challenge in autonomous UAV ISR missions is bridging the gap between raw sensory perception and effective decision-making under uncertainty. We next review previous works in visual perception, world modeling, and planning.

[1] https://www.darpa.mil/program/assured-neuro-symbolic-learning-and-reasoning

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

1) Visual Perception and Grounding: Visual grounding methods such as YOLO-World [4], GLIP [15], and GroundingDINO [17] are able to detect objects via textual queries. Despite their effectiveness in controlled scenarios, these neural-only methods struggle to generalize in dynamic, realistic environments and lack a persistent, interpretable representation of the scene—an essential feature for robust UAV navigation in uncertain settings [32].

2) World Modeling and Planning: Coverage Path Planning (CPP), or *coverage*, is the task of computing a path that passes over all locations in a given area, while avoiding obstacles. This is regarded as a specialized form of visual navigation and remains a fundamental challenge in robotics and planning [7], [35]. Map-based approaches [1], [21], [35], typically assume static environments, limiting their adaptability in dynamic contexts. While deep learning methods [11], [24] improve adaptability to unknown environments, they suffer from high computational costs, long training times, and sparse reward issues [35]. Partially Observable Markov Decision Processes (POMDPs) have been widely used as a probabilistic framework for modeling UAV search and rescue missions [8], [36]. These approaches typically represent the environment using a grid-based abstraction, with empirically defined transition and observation functions. Based on this setup, researchers have proposed greedy and potential-based strategies to guide UAV behavior. However, due to the inherent scalability issues of POMDPs, particularly the exponential growth of their state space, these methods often rely on overly simplified environmental models. This limits their applicability and effectiveness in realistic, continuous-world scenarios. Probabilistic world modeling [32] remains under-explored for UAV missions, further restricting adaptability in dynamic settings. When considered as a whole, both end-to-end and compositional approaches do not support explicit visual reasoning and/or cannot maintain an interpretable, probabilistic world model. NEUSIS is designed to directly address these shortcomings, enabling robust UAV search in complex, dynamic environments.

III. Environment and Mission

We evaluate our system using the same protocol as the DARPA ANSR program, specifically the Hybrid AI Mission Environment for Rapid Training and Testing (HAMERITT) system [14]. HAMERITT, built on the AirSim [31] plugin for Unreal Engine, enables dynamic scenario generation and realistic sensory data collection, including RGB camera feeds, depth sensors, and GPS data from a UAV platform.

Simulation-based evaluation is crucial for this mission for two key reasons. First, deploying UAVs in urban environments is highly restricted in many countries, making controlled, repeatable real-world experiments impractical. Second, simulation provides advantages beyond regulatory constraints, allowing systematic testing across diverse environmental conditions while ensuring fair and reproducible comparisons between different approaches.

By adopting this protocol, we align our evaluation with the DARPA ANSR program, ensuring consistency with existing research and a rigorous benchmarking framework for assessing our system's performance.

The UAV's mission is to identify as many entities of interest (EOI) as possible within the specified areas of interest (AOIs) and time constraints (5 minutes) without entering the keep-out-zones (KOZs). EOIs are new (i.e., non pre-populated) cars specified by combining a type and color description (e.g., "red SUV") with a probability for being within each AOIs. EOIs are known to be the only cars with their description in AOIs for which they have non-zero probability. To succeed, the UAV must prioritize its focus on the most promising AOIs and allocate its time wisely. Visualization of a potential mission scenario is shown in Figure 3. To challenge robustness under adversarial conditions, distracting non-EOI cars that partially match an EOI description (e.g., with the correct type and color, but positioned outside any AOI) may also be present.

IV. The NEUSIS System

NEUSIS, Neuro-Symbolic Intelligent Search, is a compositional framework (see Figure 4) comprised of three main components: a neuro-symbolic visual perception, reasoning and grounding component (GRiD); a symbolic world model; and a symbolic hierarchical planning component (SNaC).

A. Perception, Grounding, Reasoning in 3D (GRiD)

The GRiD component processes UAV sensor data, i.e., RGB, depth, and GPS to localize entities of interest (EOIs) in 3D space and infer their attributes. This is achieved by integrating visual perception, grounding, and neuro-symbolic reasoning.

GRiD builds on recent advances in neuro-symbolic compositional visual reasoning methods [2], [9], [18], [32], [34], [40], which tackle complex visual reasoning tasks by decomposing them into sub-tasks. These sub-tasks are individually solved using vision foundation models and large language model (LLM)-generated code, with the results combined to complete the overall task. For GRiD, we adopt HYDRA [13], a state-of-the-art neuro-symbolic reasoning system that combines reinforcement learning with LLM-driven code generation to enable dynamic, compositional visual understanding. Since HYDRA is designed for 2D image-based reasoning tasks (e.g., visual grounding and question answering), it requires adaptation for perception, reasoning, and grounding from the sensor stream of visual data in the UAV's 3D search mission.

To adapt HYDRA for this mission, we expand GRiD's toolkit to include 2D target bounding boxes with attributes recognition, instance segmentation, object tracking, and 3D coordinates projection. The following Python APIs are implemented to meet mission requirements: **segment**, **track**, **project_to_3d**, **classify_object_attributes**, and **classify_object_types**. We integrate state-of-the-art VFMs for these tasks. CLIP [28] was fine-tuned for classifying object attributes and types (**classify_object_attributes**, **classify_object_types**), while pretrained models [17], [38] were used for grounding (**find**, **segment**), and a symbolic method [3] was used for object tracking (**track**). Further implementation details are provided in Section V-A.

To avoid computational bottlenecks and resource overuse, we implemented a caching mechanism for the LLM-generated

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

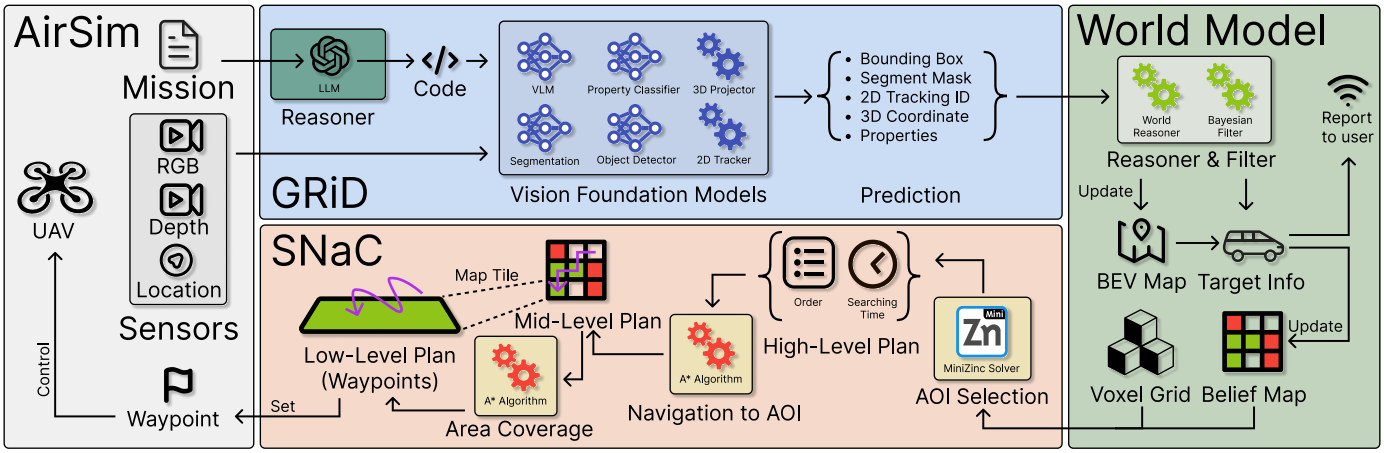


Fig. 4: The pipeline of our proposed neuro-symbolic system, NEUSIS. The UAV operates in a simulated environment (AirSim) and is equipped with sensors including RGB camera, depth camera, and GPS. The *Perception, Grounding, Reasoning in 3D* (GRiD) component processes sensor data using a reasoner (code generator) and Vision Foundation Models (VFMs), including neuro-based segmentation, object detection, property classification, and symbolic 2D tracker and 3D projector, to generate predictions. Predictions are sent to the world model, which maintains a belief map, and generates target reports. The *Selection, Navigation and Coverage* (SNaC) component generates a hierarchical plan, with the Selection, Navigation, and Area Coverage sub-tasks producing high-level, mid-level, and low-level plans.

reasoner code where the executable code is generated in advance for similar scenarios, and then directly executed without re-generation by the LLM. This allows reuse of reasoner plans across similar mission queries in different scenarios.

B. Probabilistic World Model

Due to the noise in sensor inputs, GRiD is rarely 100% confident in its output. The world model accumulates localization information from GRiD to maintain a persistent probabilistic representation of the environment and provide a mechanism for identifying and reporting entities of interest (EOIs).

The world model is initialized with a ground-truth voxel occupancy grid and a birds-eye view (BEV) segmentation map, indicating the locations of pre-populated objects like walls, trees, and roads. It is also provided with the initial prior belief map from the mission description. On each frame, it receives noisy 3D localization data and attribute likelihoods from GRiD to perform the following three tasks:

1) World Reasoning: Refines 3D localization by using domain knowledge to remove (i) infeasible points from masked depth data (e.g., spurious points from leaves or power lines) thus computing more accurate 3D center points, and (ii) detections that violate physical constraints (e.g., cars high above ground or inside walls).

2) Information Accumulation: Improves 3D localization and attribute classification by accumulating detections about the same objects over time. A naïve approach would compute the average position of detections, but this does not consider the uncertainty of GRiD's outputs. Instead, we use (i) the Hungarian algorithm for data association, (ii) linear Kalman Filters for position refinement, and (iii) discrete attribute distribution ranking for more accurate attribute likelihood updates.

3) Reporting: Generates online reports by evaluating whether any tracked objects match the EOI descriptions. To do this, it reasons about the probability of a match, and reports any candidate exceeding a confidence threshold. A final offline report summarizing the best detection of each EOI is produced at the end of the mission.

In addition to the tasks mentioned above, the World Model also maintains environmental information relevant to the planning component, including the voxel occupancy grid and belief map, to support path planning operations.

C. Selection, Navigation and Coverage (SNaC)

The SNaC component is designed to generate a trajectory that efficiently searches the areas of interest (AOIs) by maximizing the likelihood of encountering entities of interest (EOIs) within the allocated time. The component first retrieves the belief map and other environmental information from the World Model component, and then generates a sequence of waypoints to be sent to the UAV's control unit. While this task is closely related to area coverage and object goal navigation, the existence of *multiple* EOIs within the AOIs across a large environment introduces two main complexities: lack of a fixed order for visiting the AOIs, and the need to identify as many EOIs as possible within the given time, which is usually insufficient to cover all AOIs.

To this end, SNaC employs a hierarchical approach, dividing the task into three distinct sub-tasks: **S***election* (high-level planning), **Na***vigation* (mid-level planning), and **C***overage* (low-level planning). The *Selection* sub-task leverages the belief map to compute a high-level route between AOIs and to allocate an exploration time for each AOI. Based on the output from the *Selection* sub-task and other relevant information from the world model, the *Navigation* sub-task then plans a path to reach the selected AOI. Once there, the *Coverage* sub-task plans how to systematically search for EOIs in the area.

1) Selection: To efficiently explore AOIs and maximize EOI detection, we formulate the task as a constraint optimization problem. The solution determines the optimal sequence

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

of AOI visits and the corresponding time allocation for each, based on the following key pre-estimated or known factors: **AOI area** in m^2; **Prior probability of EOIs:** The likelihood of EOIs being in each AOI; **Detection rate:** The ratio of prior probability to AOI area; **Travel time:** Estimated based on the distances between AOIs and the UAV's speed; **Mission time constraint**. Formally, for each AOI a_i, we define the decision variables: allocated exploration time $0 \leq t_i \leq s_i/v_c$ (where s_i is the AOI area, v_c coverage speed) and visit-order permutation $o_1, \ldots, o_n$. The objective maximizes the expected number of detected EOIs: $\sum_{i=0}^{n} \mathbb{1}_{l_i<T}\left(\frac{t_{o_i}}{s_{o_i}/v_c} \sum_{j=1}^{m} p_{o_i,j}\right)$, where l_i tracks elapsed time including travel and exploration, $p_{i,j}$ refers to the prior probability of EOI_i in the AOI a_j, and $\mathbb{1}_{l_i<T}$ ensures AOIs are explored within mission duration T. This optimization is modeled with MiniZinc [23] and solved efficiently by the Chuffed solver [5]. The selected next AOI is then passed to the subsequent module for execution.

2) Navigation: Once an AOI is selected, this sub-task generates a path (as a sequence of waypoints) for travelling to that area. It does this by first constructing a visibility graph [25] using information regarding keep-out zones (KOZs) and voxel occupancy. It then executes an A* [10] algorithm on the visibility graph to determine the optimal path, ensuring avoidance of both obstacles and KOZ.

3) Area Coverage: After reaching an AOI, this sub-task plans the low-level search for EOIs. It begins by converting the AOI into a grid, thus representing the coverage task as the exploration of all accessible grid points. To achieve this it creates an open set of points to be visited, greedily finds the nearest non-visited point from the starting position, and uses the A* algorithm to navigate to that point while avoiding any obstacles or KOZ. Subsequently, the UAV navigates towards that point along the computed path, and removes visited points from the open set. Once the open set becomes empty, or all EOIs are found, the search concludes. Note that the belief map in the world model is updated based on the EOIs found in that AOI, and the updated belief map is then used by the *Selection* sub-task to select the next AOI.

V. Experiments

A. Implementation Details

We implemented our proposed system by integrating the GRiD, World Model, and SNaC components, and compared its performance against a framework built using state-of-the-art (SoTA) solutions for perception and planning. This comparison highlights the contribution of our system to the specific problem. Additionally, we conducted several ablation studies for each component to assess the impact of their (sub)tasks and the specific features they contribute.

Toolkit in GRiD. For GRiD, we use GroundingDINO [17] for grounding, linear probed CLIP [28] for color/type classifiers, and OCSort [3] as the 2D tracker to assign tracking IDs to targets. After generating the 2D bounding box for the detected target by the visual grounding model, we use EfficientSAM [38] to obtain the pixel mask of the target. Using the depth sensor data, we compute the 3D coordinates of all pixels within the mask as a point cloud. The 3D location of the target is determined by averaging these points, and then sent to the world model. In ablation studies, we follow HYDRA [13] to use GLIP [15] for grounding and XVLM [41] for zeroshot color/type classifiers, as the original VFMs.

Baseline. The baseline system is composed of either YOLO-World [4] or GroundingDINO [17] for the perception component and Fields2Cover [21] for the planning component. A subset of these components was selected by the DARPA ANSR program, ensuring their relevance and suitability for our task. YOLO-World and GroundingDINO represent the state-of-the-art in vision-language models (VLMs), being known for their efficiency and high performance in 2D grounding tasks. They do not provide estimated segmentation masks for EOIs, so we compute their 3D coordinates by projecting the center of the 2D bounding box using the available depth data. On the planning side, Fields2Cover is a symbolic, model-based planner widely used for autonomous planning due to its robustness and efficiency in navigating complex environments.

B. Metrics

The UAV's mission is to identify as many entities of interest (EOIs) as possible within the given mission time. To fulfill this objective, we establish the primary requirements for the task: a reliable system should not only successfully detect EOIs but also maintain consistency in decision-making throughout the mission. Moreover, it should navigate efficiently to minimize detection time while ensuring accurate reporting. Based on these considerations, we employ a set of evaluation metrics that effectively capture these aspects.

The *offline F1-score*, along with *precision* and *recall*, serves as the primary metric for assessing the system's overall performance. These offline metrics evaluate final reports that consolidate all information gathered during the mission. A reported EOI is considered correct if its reported position is within 5 meters of the ground truth. In addition, we also include a popular metric in the domain, *Success Rate* (SR), defined as n/N, where N is the total number of EOIs and n the number of successfully detected EOIs. While SR is widely used, it has limitations, because correctly reporting an EOI once does not guarantee the correctness of previous and/or subsequent decisions. SR solely reflects the quality of planning, not the system's reasoning ability. Hence, it serves only as a supplementary reference and is used in the ablation study for the planning component.

To independently evaluate the perception component, we use the *online F1-score* as the primary metric, accompanied by *precision* and *recall*. Unlike offline metrics, which evaluate final consolidated reports, online metrics assess frame-wise detection performance based on immediate reports, without incorporating reasoning ability. The results across all frames are accumulated to compute overall online F1-score, precision, and recall for the scenario.

For evaluating the planning component, we measure *penalized detection time* as an indicator of navigation efficiency. Detection time is meaningful only when all approaches achieve the same SR; hence, we include a penalty system. Specifically, for each detected EOI (1st, 2nd, 3rd, and 4th), if an EOI is not

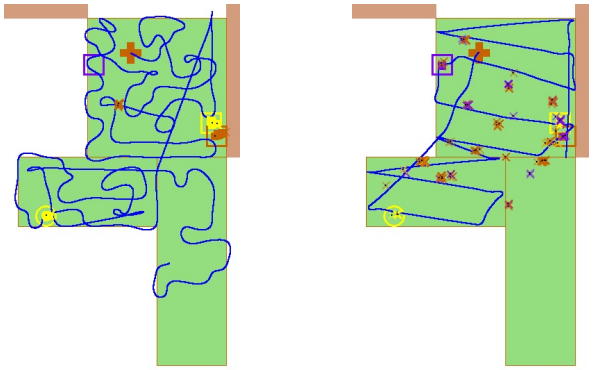
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Planning Component	Perception Component	World Model	F1-score (↑) Offline	F1-score (↑) Online	Success Rate (↑)	Detection Time (↓) 1st	Detection Time (↓) 2nd	Detection Time (↓) 3rd	Detection Time (↓) 4th	Penalized Detection Time (↓) 1st	Penalized Detection Time (↓) 2nd	Penalized Detection Time (↓) 3rd	Penalized Detection Time (↓) 4th
Fields2Cover [21]	YOLO-World [4]	✗	04.40	11.40	29.58	78.30	137.57	179.79	256.37	159.59	237.73	281.97	299.07
SNaC	YOLO-World [4]	✗	06.17	08.24	31.67	74.02	93.64	96.93	120.09	122.98	224.34	279.69	297.00
Fields2Cover [21]	GroundingDINO [17]	✗	17.37	21.12	31.88	75.90	91.63	167.53	196.14	154.40	236.73	289.33	292.89
SNaC	GroundingDINO [17]	✗	18.62	19.87	33.75	63.20	91.92	105.24	135.19	149.03	239.18	284.33	292.88
Fields2Cover [21]	GRiD	✗	24.44	29.07	36.17	66.78	100.95	162.81	186.02	136.75	210.43	254.27	290.50
Fields2Cover [21]	GRiD	✓	30.56	40.15	36.32	60.39	134.13	167.68	165.92	126.31	188.21	245.84	283.61
SNaC	GRiD	✗	41.27	50.29	40.27	63.27	104.54	122.85	149.60	96.19	189.97	243.74	275.84
SNaC	GRiD	✓	**52.07**	**54.12**	**61.82**	61.25	98.19	138.98	134.78	**87.30**	**138.55**	**209.24**	**263.95**

TABLE I: Quantitative comparison between the proposed methods with baselines.

SoTA VFMs	Color/Type Classifiers	2D tracking	3D Projection on center of	World Model	F1-score (↑) Offline	F1-score (↑) Online	Precision (↑) Offline	Precision (↑) Online	Recall (↑) Offline	Recall (↑) Online
✗	Zeroshot	✗	2D bbox	✗	05.39	30.08	05.79	32.61	05.08	33.69
✗	Zeroshot	✗	3D pointcloud	✗	25.28	35.24	28.72	40.09	23.21	39.60
✓	Zeroshot	✗	3D pointcloud	✗	30.73	40.12	40.28	57.92	26.39	39.92
✓	Linear-probing	✗	3D pointcloud	✗	41.27	50.29	45.28	65.36	38.75	42.56
✓	Linear-probing	✗	3D pointcloud	✓	49.29	53.40	57.58	67.92	44.89	46.76
✓	Linear-probing	✓	3D pointcloud	✓	**52.07**	**54.12**	**59.82**	**68.71**	**47.77**	**47.34**

TABLE II: Ablation study of GRiD component.



(a) NEUSIS (GRiD, World Model, SNaC) (b) Baseline (YOLO-World, Fields2Cover)

Fig. 5: Comparison of (a) our system and (b) the baseline method on the scenario depicted in Figure 3. Filled, colored shapes denote correct EOI reports, colored crosses denote false positives, and blue curves represent the UAV's flight path.

detected, we assign it the maximum mission time (300 seconds in our experiments), treating it as if the method had detected all missing EOIs at the very end.

C. Quantitative Comparison

We compared the F1-score, success rate, and detection time across different configurations of planning and perception components, as shown in Table I. When replacing Fields2Cover with SNaC (row 2), we often observe a significant improvement in penalized detection time. Replacing YOLO-World (rows 1 and 2) with GroundingDINO (rows 3 and 4) improves F1-score and success rate, and often reduces penalized detection time, indicating more reliable detections. Similarly, comparing rows 1, 3, and 5, where YOLO-World and GroundingDINO are replaced with GRiD, there is a dramatic improvement in EOI localization performance, as indicated by the higher F1-scores and the reduction in penalized detection time. It is worth noting that noisy reports from YOLO-World and GroundingDINO lead to higher success rates (around 30%) than would be expected, as only one report needs to be correct, and success rate does not adequately penalize incorrect reports. The F1-score metric gives a stronger indication of the actual performance of perception systems, and GRiD outperforms YOLO-World on this metric by around 20%. Further, the combination of GRiD with SNaC (row 7) leads to a substantial increase in mission F1-score, success rate, and detection time (penalized or not). The routes that SNaC produces allow the GRiD and world model to see cars in the environment from more directions, allowing for higher confidence to be built before making a report. Finally, with the addition of the world model in rows 6 and 8, we see a further improvement, in particular in terms of the success rate, offline F1-score, and penalized detection time, demonstrating the effectiveness of our compositional neuro-symbolic approach.

D. Qualitative Comparison

To better understand the results of our experiments, we examined visualisations of the behaviour of the different configurations. Figure 5 provides a representative example of flight paths and entities of interest (EOI) reports from a) our proposed system NEUSIS, and b) the baseline system based on Fields2Cover and YOLO-World. The Fields2Cover approach employs a deliberate back and forth search strategy that systematically covers the areas of interest (AOIs). However, YOLO-World produces many false positives (colored crosses), and also generated noisy output near the ground truth targets. NEUSIS's planning component SNaC performs a more bespoke exploration that allows GRiD and the world model to see potential EOIs from more angles, thus making more confident reports. Overall, the qualitative visualisation shows the advantages of our integrated neuro-symbolic system in both navigation efficiency and target detection performance.

E. Ablation Studies

GRiD. We conducted extensive ablation studies to evaluate the impact of different tasks in the GRiD component. Table II presents the results using online and offline perception metrics (F1-score, precision and recall) for comparison. The first two rows highlight the impact of 3D projection methods, demonstrating that point cloud-based 3D projection significantly outperforms projecting the center point of 2D bounding boxes. The results from rows 2, 3, and 4 show the substantial positive contribution of state-of-the-art (SOTA) VFMs and color/type classifiers. Finally, rows 4, 5, and 6 show the effectiveness of integrating the world model and 2D tracking, both of which lead to notable performance improvements.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

World Model. Ablation studies for the world model are presented in Table III. Starting with a version that only performs basic world reasoning, we see that the addition of information accumulation with naïve filtering (using the average 3D position), only provides a small improvement for online F1-score ($42.57\% \rightarrow 44.62\%$). When Bayesian filtering is added we see a 10% increase in online F1-score, demonstrating the importance of correctly handling uncertainty.

World Model Component Information Accumulation	F1-score (↑) Offline	 Online	Success Rate (↑)
World reasoning only	48.68	42.57	59.78
+ Naïve Accumulation	48.55	44.62	58.33
+ Bayesian Filtering	**52.07**	**54.12**	**61.82**

TABLE III: Ablation study of World Model.

SNaC. Table IV presents the ablation study for the SNaC component with ground truth perception, where targets are reported within a 25-meter range, and the mission must be completed within 300 seconds. We use GT perception to remove the random effect from noisy output of perception module to the performance of the planning module, ensuring a fair evaluation of SNaC's effectiveness.

Planning Component	Success Rate (↑)	Penalized Detection Time (↓) 1st	 2nd	 3rd	 4th
Baseline	20.83	128.80	197.62	254.49	300.00
+ Selection	43.75	116.93	202.70	252.23	297.01
+ Area coverage	**54.51**	109.66	**174.78**	**243.97**	**292.07**

TABLE IV: Ablation study of SNaC component.

Starting with the baseline version, which uses Fields2Cover [21] for area coverage and computes a route based on the closest AOIs, the introduction of the *Selection* sub-task, incorporating MiniZinc optimization based on the belief map, significantly improves the success rate, nearly doubling it (from 20.83% to 43.75%). This optimization enhances efficiency in determining the areas of interest (AOIs) visitation order and exploration time allocation by considering not only distance but also the probability of finding a target. Finally, the addition of the proposed *Coverage* sub-task further improves performance. While our maneuver takes longer to cover specific sections of the map, as illustrated in Figure 5, it raises the success rate to 54.51%, demonstrating its effectiveness in improving low-level search coverage throughout the environment. The vertices used to generate a coverage path are sampled from the belief map, resulting in visits to specific areas based on the current probability of finding a target at different locations on the map. This approach proves to be more effective at locating targets compared to the baseline, which often fails to detect the last target due to gaps in its coverage strategy.

F. Analysis of GPS Noise

In practice, the location information received from a real GPS sensor will be noisy [37]. To evaluate how NEUSIS would perform under realistic conditions, we reran our test scenarios while injecting zero-mean Gaussian noise (noise power is parametrized by a standard deviation value) to the UAV's horizontal position measurements, at a frequency under 10 Hz, simulating real-world GPS inaccuracies. Crucially, no modifications were made to the perception (GRiD), world model, or planner (SNaC) modules, allowing us to isolate and assess the impact of pose noise on the system's end-to-end performance. The results of these experiments are shown in Figure 6. They show that within the 95% confidence interval of typical GPS noise (0–1.82 m) [37], NEUSIS maintains more than 95% of its nominal performance, with only a minor drop (around 4.8%) in F1-score. Compared with NEUSIS, the decline in performance of other baselines is much more significant, as reflected in the higher relative percentage drop. Furthermore, compared with NEUSIS without world model, the effect from noise is increased, but still better than the baselines. Therefore, this robustness stems from the system design: our probabilistic world model absorbs perceptual uncertainty, while the planner operates on a discrete grid (1 m resolution), reducing sensitivity to small positional errors.

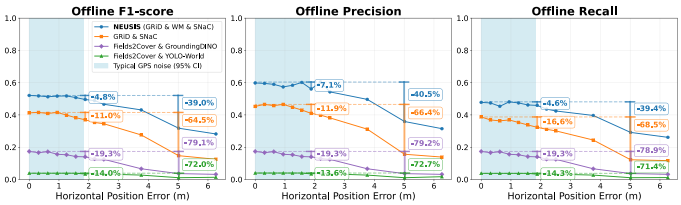


Fig. 6: Analysis on the noise error. The x-axis shows the horizontal positional error in meters, and the y-axis the value of the performance metrics (offline F1-score, offline precision and offline recall). The blue zone highlights the 95% confidence interval of the typical GPS noise range [37].

VI. Conclusion

This paper presented NEUSIS, a compositional neuro-symbolic system for autonomous UAVs in complex search missions. By integrating neuro-symbolic perception (GRiD), a probabilistic world model, and a hierarchical symbolic planning component (SNaC), our approach enables efficient target detection, reasoning, and navigation. Extensive experiments demonstrate that NEUSIS significantly outperforms baselines for both perception and planning.

Broader Impact. NEUSIS has potential for real-world applications such as search-and-rescue missions, improving UAVs' ability to locate targets in hazardous environments. We acknowledge that the autonomous search capability we develop here has the potential for use in harmful applications.

Limitations. While NEUSIS shows strong performance in a high-fidelity simulation, real-world UAV deployment remains an open challenge due to the need for additional hardware integration, communication optimization, and regulatory clearance. Extending the system for physical field trials is an important direction for future work.

References

[1] T. M. Cabreira, L. B. Brisolara, and P. R. Ferreira Jr., "Survey on coverage path planning with unmanned aerial vehicles," *Drones*, vol. 3, no. 1, 2019.

[2] Z. Cai, F. Ke, S. Jahangard, M. Garcia de la Banda, R. Haffari, P. J. Stuckey, and H. Rezatofighi, "NAVER: A neuro-symbolic compositional automaton for visual grounding with explicit logic reasoning," *arXiv preprint arXiv:2502.00372*, 2025.

[3] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9686–9696.

[4] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 901–16 911.

[5] G. Chu, "Improving combinatorial optimization," Ph.D. dissertation, University of Melbourne, Australia, 2011.

[6] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, and M. Schwager, "Foundation Models in Robotics: Applications, Challenges, and the Future," *The International Journal of Robotics Research*, vol. 44, no. 5, pp. 701–739, 2025.

[7] E. Galceran and M. Carreras, "A survey on coverage path planning for robotics," *Robotics and Autonomous systems*, vol. 61, no. 12, pp. 1258–1276, 2013.

[8] A. Gupta, D. Bessonov, and P. Li, "A decision-theoretic approach to detection-based target search with a uav," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5304–5309.

[9] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 953–14 962.

[10] P. Hart, N. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[11] J. Hu, H. Niu, J. Carrasco, B. Lennox, and F. Arvin, "Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 413–14 423, 2020.

[12] J. Huang, S. Xie, J. Sun, Q. Ma, C. Liu, D. Lin, and B. Zhou, "Learning a Decision Module by Imitating Driver's Control Behaviors," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 1–10, iSSN: 2640-3498.

[13] F. Ke, Z. Cai, S. Jahangard, W. Wang, P. D. Haghighi, and H. Rezatofighi, "HYDRA: A hyper agent for dynamic compositional visual reasoning," in *European Conference on Computer Vision*. Springer, 2024, pp. 132–149.

[14] H. Keno, N. J. Pioch, C. Guagliano, and T. H. Chung, "Simulation-based Scenario Generation for Robust Hybrid AI for Autonomy," Sept. 2024, arXiv:2409.06608 [cs].

[15] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 965–10 975.

[16] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2Motion: from natural language instructions to feasible plans," *Autonomous Robots*, vol. 47, no. 8, pp. 1345–1365, Dec. 2023.

[17] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 38–55.

[18] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] S. Macenski, F. Martín, R. White, and J. Ginés Clavero, "The marathon 2: A navigation system," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[20] P. Mahmoudieh, D. Pathak, and T. Darrell, "Zero-Shot Reward Specification via Grounded Natural Language," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, June 2022, pp. 14 743–14 752, iSSN: 2640-3498.

[21] G. Mier, J. Valente, and S. de Bruin, "Fields2Cover: An open-source coverage path planning library for unmanned agricultural vehicles," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2166–2172, 2023.

[22] S. M. S. Mohd Daud, M. Y. P. Mohd Yusof, C. C. Heo, L. S. Khoo, M. K. Chainchel Singh, M. S. Mahmood, and H. Nawawi, "Applications of drone in disaster management: A scoping review," *Science & Justice*, vol. 62, no. 1, pp. 30–42, 2022.

[23] N. Nethercote, P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, and G. Tack, "Minizinc: Towards a standard cp modelling language," in *Principles and Practice of Constraint Programming – CP 2007*, C. Bessière, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 529–543.

[24] F. Niroui, K. Zhang, Z. Kashino, and G. Nejat, "Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 610–617, 2019.

[25] B. Oommen, S. Iyengar, N. Rao, and R. Kashyap, "Robot navigation in unknown terrains using learned visibility graphs. part i: The disjoint convex obstacle case," *IEEE Journal on Robotics and Automation*, vol. 3, no. 6, pp. 672–681, 1987.

[26] N. D. Palo, A. Byravan, L. Hasenclever, M. Wulfmeier, N. Heess, and M. Riedmiller, "Towards A Unified Agent with Foundation Models," in *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, Mar. 2023.

[27] S. Primatesta, G. Guglieri, and A. Rizzo, "A risk-aware path planning strategy for uavs in urban environments," *Journal of Intelligent & Robotic Systems*, vol. 95, 08 2019.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, July 2021, pp. 8748–8763, iSSN: 2640-3498.

[29] D. Shah, B. Osiński, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," in *Proceedings of The 6th Conference on Robot Learning*. PMLR, Mar. 2023, pp. 492–504, iSSN: 2640-3498.

[30] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A Foundation Model for Visual Navigation," in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 711–733, iSSN: 2640-3498.

[31] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[32] A. Stanić, S. Caelles, and M. Tschannen, "Towards truly zero-shot compositional visual reasoning with llms as programmers," *arXiv preprint arXiv:2401.01974*, 2024.

[33] J. Sun, H. Sun, T. Han, and B. Zhou, "Neuro-Symbolic Program Search for Autonomous Driving Decision Module Design," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 21–30, iSSN: 2640-3498.

[34] D. Sur'is, S. Menon, and C. Vondrick, "Vipergpt: Visual inference via python execution for reasoning," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 854–11 864, 2023.

[35] C. S. Tan, R. Mohd-Mokhtar, and M. R. Arshad, "A comprehensive review of coverage path planning in robotics using classical and heuristic algorithms," *IEEE Access*, vol. 9, pp. 119 310–119 342, 2021.

[36] S. Waharte, A. Symington, and N. Trigoni, "Probabilistic search with agile uavs," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 2840–2845.

[37] J. William and H. T. C. N. T. Team, "Global positioning system standard positioning service performance analysis report," *FAA GPS Performance Analysis Report*, vol. 7, 2014.

[38] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, R. Krishnamoorthi, and V. Chandra, "EfficientSAM: Leveraged masked image pretraining for efficient segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 111–16 121.

[39] L. Yang, J. Qi, J. Xiao, and X. Yong, "A literature review of uav 3d path planning," in *Proceeding of the 11th world congress on intelligent control and automation*. IEEE, 2014, pp. 2376–2381.

[40] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. A. Ayyubi, K.-W. Chang, and S.-F. Chang, "Idealgpt: Iteratively decomposing vision and language reasoning via large language models," *arXiv preprint arXiv:2305.14985*, 2023.

[41] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 994–26 009.

[42] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, and et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 2165–2183.