Towards Autonomous UAV Visual Object Search in City Space: Benchmark and Agentic Methodology

Yatai Ji[1, 3], **Zhengqiu Zhu**[*1, 3], **Yong Zhao**[1, 3], **Beidan Liu**[1, 3], **Chen Gao**[2], **Yihao Zhao**[2], **Sihang Qiu**[1, 3], **Yue Hu**[*1, 3], **Quanjun Yin**[1, 3]

[1]State Key Laboratory of Digital Intelligent Modeling and Simulation, National University of Defense Technology
[2]BNRist, Tsinghua University
[3]College of Systems Engineering, National University of Defense Technology
{jiyatai_1209, zhuzhengqiu12, zhaoyong15, liubangdan, huyue11}@nudt.edu.cn, {chgao96, yh-zhao21}@mails.tsinghua.edu.cn, sihangq@acm.org, yinquanjun@nudt.edu.cn

Abstract

Aerial Visual Object Search (AVOS) tasks in urban environments require Unmanned Aerial Vehicles (UAVs) to autonomously search for and identify target objects based on visual inputs without external guidance. Existing approaches struggle in complex urban environments due to redundant semantic processing, similar object ambiguity, and the exploration-exploitation dilemma. To advance research and support the AVOS task, we introduce CityAVOS, the first benchmark dataset for autonomous search of static urban objects. It features 2,420 tasks of varying difficulty across six object categories, designed to rigorously evaluate UAV search strategies. To solve the AVOS task, we also propose **PRPSearcher** (**P**erception-**R**easoning-**P**lanning **Searcher**), a novel agentic method powered by multi-modal large language models (MLLMs) that enables a UAV agent to reason like humans on visual cues when searching for objects. Specifically, PRPSearcher constructs three specialized 3D maps: an object-centric dynamic semantic map enhancing spatial perception, a cognitive map based on semantic "attraction" values for target reasoning, and an uncertainty map for balanced exploration-exploitation search. Moreover, we propose a denoising mechanism to mitigate interference from similar objects and design an "Inspiration Promote Thought" prompting mechanism for adaptive action planning. Experimental results on CityAVOS demonstrate that PRPSearcher surpasses existing baselines in both success rate and search efficiency (on average: +37.69% SR, +28.96% SPL, -30.69% MSS, and -46.40% NE). Our work paves the way for future advances in embodied visual target search.

Code — https://github.com/jiyatai/CityAVOS

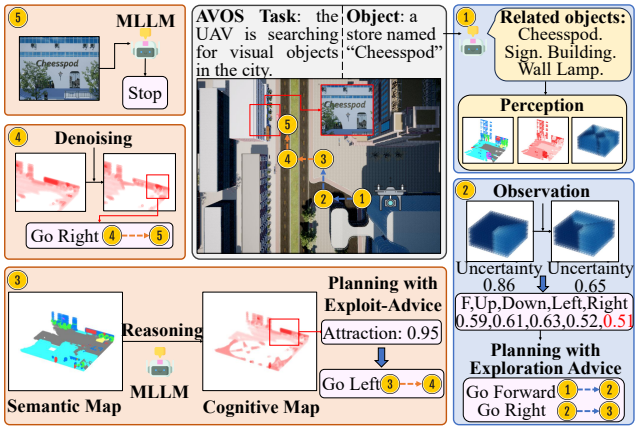

Figure 1. An illustration case of a UAV performing the AVOS task in an unfamiliar urban environment. In the search process, the UAV agent perceives the surrounding urban environment and reasons about the potential locations of the target object. In steps 1 and 2, the agent plans exploratory actions to probe the unknown space with the highest uncertainty. In steps 3 and 4, the agent plans exploitative actions to search in the area with the highest attractions. The agent's adaptive switching between exploration and exploitation is governed by the IPT mechanism. Finally, in step 5, the agent finds the target object and stops.

1 Introduction

Unmanned Aerial Vehicles (UAVs) have found extensive applications in object search missions within city environments. Notable use cases encompass last-mile delivery in logistics systems (She and Ouyang 2021) and rescue search operations in emergency response scenarios (Zhao et al. 2019). Traditional solutions for UAV-based object search typically leverage metaheuristics or deep reinforcement learning methods to improve search efficiency through optimized flight path planning (Xing et al. 2022; Hou et al. 2023). However, the potential of dynamic visual observations of UAVs is often overlooked in these traditional methods. Recent advancements in embodied intelligence have enabled UAV-based agents driven by Multi-modal Large Language Models (MLLMs) to exhibit human-like proficiency in visual understanding, cognitive reasoning, and action decision-making (Liu et al. 2024; Huang et al. 2025). Consequently, the traditional object search task is transitioning towards Aerial Visual Object Search (AVOS) tasks, in which UAVs are expected to autonomously find visual objects (initial input: the images and text descriptions of the

*Corresponding author.

target object are provided) in unfamiliar urban settings based on dynamic visual inputs without any navigational assistance or external instructions.

The AVOS task shares a similar task formulation (but different) from the indoor Image Goal Navigation (ImageNav) task. In the ImageNav domain, modular approaches (Al-Halah, Ramakrishnan, and Grauman 2022) often construct semantic maps to retrieve semantic information related to the image goal from memory, subsequently navigating step-by-step to the goal location (Mezghan et al. 2022). In indoor environments, agents can typically infer the goal location through world knowledge and navigate to it by analyzing the relationship between the image goal and its surroundings (Li et al. 2025). However, in urban environments, extensive spaces contain abundant semantic information and similar objects. Successfully locating the image goal in such scenarios requires the agent not only to infer potential target locations from visual cues but also to possess efficient spatial exploration capabilities, which represents the primary difference between the AVOS task and the ImageNav task. In addition, since image goals in outdoor environments often contain multiple semantic objects, a text description is added to each AVOS task to specify which particular object in the image is the target.

This paper investigates the AVOS task in city spaces, facing three unique challenges compared with previous studies:

1) **Complex and rich semantics of objects pose challenges to spatially-aware environmental representations**: Existing approaches primarily rely on point clouds or semantic grid maps for spatial awareness, but they often fall short in computational efficiency and mapping accuracy due to the redundant semantic information in complex urban environments. Therefore, a critical need exists for novel semantic mapping methods designed for urban contexts that are both computationally efficient and accurate.

2) **Visual resemblance among objects poses challenges to target reasoning and identification**: Urban scenes often feature multiple similar objects, such as shops, billboards, and cars, which are hard to distinguish remotely. Accurate identification typically requires closer observation. Therefore, a key challenge lies in mitigating interference from these visually analogous yet incorrect targets during the target reasoning.

3) **Vast urban space and complex spatial structures pose challenges to action planning**: In large, complex urban settings, building, tree, and other occlusions can create visual blind spots in agent-constructed semantic maps. This leads to a difficult trade-off: searching only for semantic targets ignores unexplored areas, while exploring broadly is often inefficient. Thus, balancing this exploration-exploitation dilemma in action planning is a challenge.

As an initial step, we develop a benchmark dataset, CityAVOS, to evaluate a UAV agent's performance on AVOS tasks. We summarize the differences between this dataset and other benchmark datasets in Appendix. The CityAVOS dataset categorizes six common target types and defines two levels of search difficulty. The targets are described by both images and text, within complex scenes featuring intricate semantic information and spatial structures. Notably, a UAV agent receives no guiding instructions, requiring itself to perform a zero-shot autonomous search.

To address AVOS tasks, we introduce PRPSearcher (**P**erception-**R**easoning-**P**lanning UAV **Searcher**), a novel agentic method powered by MLLMs, designed to follow the three-stage process that humans typically employ when performing search tasks, as illustrated in Fig. 1. **In the perception phase**, PRPSearcher leverages the semantic reasoning capabilities of MLLMs to filter out irrelevant semantic elements during the segmentation process, thereby constructing an object-centric 3D semantic map. The semantic filtering and dynamic updating mechanisms enhance the efficiency and accuracy of semantic map construction. Moreover, PRPSearcher constructs and updates a 3D uncertainty map to record the extent of spatial exploration in the surrounding environment. **In the reasoning phase**, a 3D cognitive map is created based on 'attraction' values deducted by the MLLM that quantifies the semantic attraction of an object to a UAV agent. Moreover, we design a denoising mechanism to eliminate the influence of non-target objects. **In the planning phase**, we introduce an Inspiration Promote Thought (IPT) prompting mechanism based on the cognitive map and the uncertainty map to help the agent strike a balance between exploration and exploitation during the action decision-making process. Results show that PRPSearcher achieves 53.50% of SR and 40.57% of SPL in CityAVOS tasks, significantly surpassing the performance of the SOTA baselines. In summary, the contributions of this work are as follows:

- To the best of our knowledge, we are the first to introduce a benchmark dataset for the AVOS task in city space, namely CityAVOS.
- Inspired by human three-tier cognition, we propose an MLLM-based agentic method to address the AVOS task. It is mainly owing to the construction of three maps — a semantic map, a cognitive map, and an uncertainty map — to enhance agents' spatial perception, target reasoning, and action planning capabilities.
- Experimental results demonstrate that our approach outperforms existing baselines in tackling the AVOS task. However, the performance gap compared to humans highlights opportunities for future research to improve semantic reasoning and spatial exploration in AVOS tasks.

2 Related Work

2.1 Image Goal Navigation

Image-goal navigation tasks require agents to navigate in response to a target described by an image (Chaplot et al. 2020; Li et al. 2025; Qin et al. 2025). Since objects in indoor environments are typically correlated with their locations, most studies (Mousavian et al. 2019; Al-Halah, Ramakrishnan, and Grauman 2022; Du et al. 2023) train end-to-end models using deep reinforcement learning (DRL) to translate visual information into agent actions. Although these approaches have demonstrated promising results in simulators, they still suffer from high computational complexity and poor generalization capabilities. Modular approaches effectively address these issues to some extent.

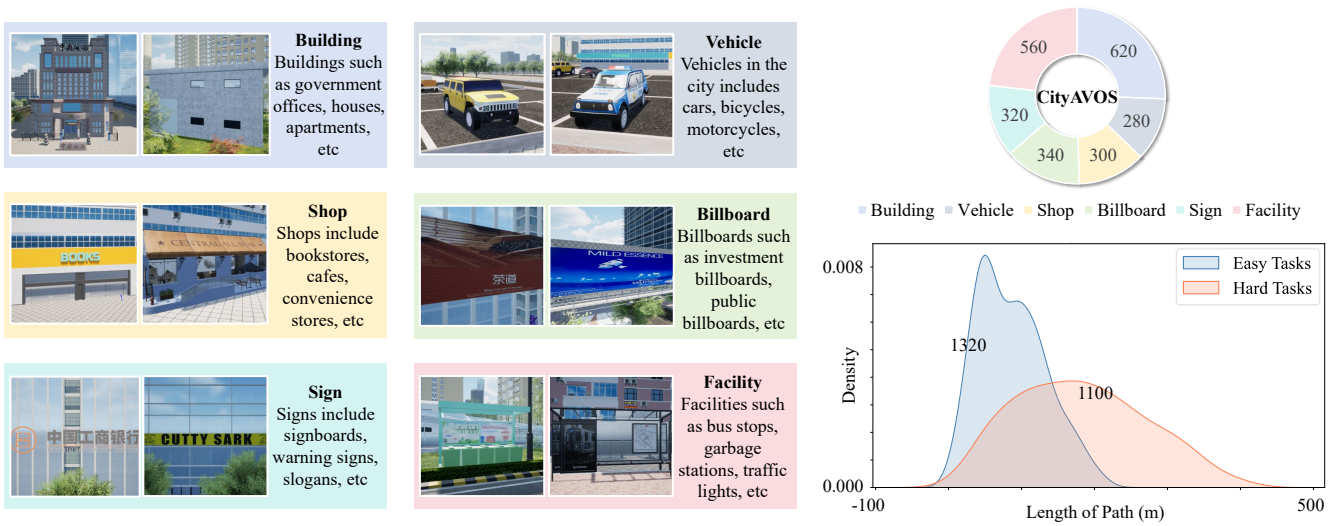

Figure 2. Examples of six object categories and dataset statistics of the CityAVOS.

Kim *et al.* (Kim et al. 2023) introduced a pretrained semantic segmentation model to process visual inputs. UniGoal (Yin et al. 2025) proposed a multi-stage exploration strategy, utilizing graph matching techniques to align navigation targets. While Qin *et al.* (Qin et al. 2025) developed a relational policy network to predict agent actions. Moreover, Large Language Models (LLMs) have been widely applied in these indoor object navigation methods (Dorbala, Mullen, and Manocha 2023; Cai et al. 2024a). For instance, L3MVN (Yu, Kasaei, and Cao 2023) used LLMs for commonsense reasoning to improve ImageNav efficiency.

These studies primarily focus on indoor scenes and can hardly be directly applied to AVOS tasks in urban environments. However, their semantic mapping and cognitive reasoning approaches offer meaningful insights, which inspire us to develop outdoor exploration techniques that mimic human cognition, improving agents' spatial perception, target reasoning, and action planning capabilities.

2.2 Urban Object Search

Traditional urban object search methods (Xing et al. 2022; Hou et al. 2023) typically rely on optimization algorithms like meta-heuristics (Wu et al. 2023) to generate search paths. Some other approaches incorporate graph neural networks (Zhang et al. 2025) with DRL (Wu et al. 2019) to address this problem. However, these approaches often lack the capability to process or incorporate the initially given image, linguistic instructions, and on-board visual observations effectively. Recent advancements in embodied intelligence and LLMs have significantly advanced urban object search methodologies. For instance, Doschl *et al.* (Döschl and Kiam 2024) proposed Say-REAPEx, an LLM-modulo online planning framework that prunes target-irrelevant actions from the planning process. To enhance LLM interpretability within urban contexts, NEUSIS (Cai et al. 2024b) integrates neuro-symbolic methods to aid environmental reasoning. Currently, research on AVOS tasks within city spaces remains in its nascent stage, but the rapid evolution of urban embodied environments and benchmark platforms, such as AerialVLN (Liu et al. 2023), OpenUAV (Wang et al. 2024), EmbodiedCity (Gao et al. 2024a), has already provided fertile ground for its growth.

Nevertheless, there remains a notable absence of a dedicated AVOS benchmark tailored for urban environments, as well as a corresponding effective baseline model. Thus, this work contributes a comprehensive benchmark dataset for the AVOS task, and an effective MLLM-based agent baseline for autonomous visual search in urban environments.

3 CityAVOS Dataset

In this section, we first define the AVOS task. Then, we introduce the construction process of the CityAVOS dataset and perform the dataset statistics.

3.1 Task Definition

In an AVOS task, a UAV agent is required to explore an unseen urban environment and search for a visual object with task information $G = (I, T)$, where I represents the visual information (image) of the object and T represents the text information of the object. At each step t, the agent perceives the RGB image V_t and depth image D_t in its current pose $P_t = [pos_t, ori_t]$. With observations $O_t = \{V_t, D_t, P_t\}$, the agent establishes an estimation of the visual object E_t. Then, a search policy $\pi(a_t|G, E_t)$ is employed to generate an action a_t. The agent determines whether to search or locate the target based on observations. Finally, the search task ends when the agent executes the stop action.

3.2 Dataset Construction

We develop CityAVOS based on EmbodiedCity (Gao et al. 2024a), a platform built on Unreal Engine 5.3 that features high-fidelity simulations of urban streets, buildings, trees, vehicles, and pedestrians (Zhao et al. 2025). By integrating AirSim (Shah et al. 2018), the platform provides a realistic

environment for evaluating the performance of autonomous UAVs in urban settings. Using this environment, we define six distinct search scenarios (e.g., streets, neighborhoods, parks), with areas ranging from 5,600 to 82,800 square meters. To adapt these scenarios for the AVOS task, we embed specific recognizable objects within the scenes.

3.3 Dataset Statistics

The CityAVOS dataset contains 2,420 AVOS tasks and their corresponding trajectories, which consist of objects in the following six categories: *building*, *vehicle*, *shop*, *billboard*, *sign*, and *facility*. The distribution of these categories of tasks is illustrated in the top right corner of Fig. 2. The tasks in the dataset are categorized into two levels of difficulty: easy and hard. For easy tasks, the agent is required to locate a unique object within a small-scale scene. Hard tasks require the agent to identify non-unique targets in a large-scale scene. The bottom right corner of Fig. 2 illustrates the distribution of the corresponding difficulty levels.

4 The Agentic Method

4.1 Overview

An overview of the proposed PRPSearcher agent for the AVOS task is illustrated in Fig. 3. **In the perception phase**, the UAV agent creates an object-centric 3D dynamic semantic map of its surroundings by employing an MLLM to reason about target-related objects and extract corresponding semantics. To provide a quantitative measure of the environment's exploration status at each timestep, PRPSearcher also updates a 3D uncertainty map within the UAV's field of view. **In the reasoning phase**, the UAV agent uses a 3D cognitive map to estimate the target's position. The map created by an MLLM encodes an "attraction" field that quantifies the semantic pull each object exerts on the UAV agent. To ensure accuracy, this work designs a denoising mechanism to mitigate the influence of objects unrelated to the target. Finally, **in the planning phase**, we introduce the Inspiration Promotes Thought prompting mechanism for the UAV agent's action planning. This mechanism inputs target location estimates into the prompt as "exploitation advice", guiding the agent's search and target identification. This is balanced by selectively adding "exploration advice" from the 3D uncertainty map, serving as "Inspiration" to encourage exploring unknown areas alongside exploiting known ones.

4.2 Object-Centric 3D Dynamic Semantic Map Construction Based on Spatial Perception

To represent semantic distribution in urban environments, we construct an object-centric 3D dynamic semantic map S in a 3D-grid form based on visual observations and the UAV pose. For each task n, we employ an MLLM to reason about target-related objects based on task information (including image I_n and text description T_n) and obtain relevant semantic elements:

$$E_n = \text{MLLM}(Prompt_{rel}, I_n, T_n), \tag{1}$$

where $Prompt_{rel}$ is the prompt input for an MLLM to generate the relevant semantic elements. These elements are integrated into the prompt for segmentation, eliminating semantics unrelated to the target object. The semantic segmentation process is defined as:

$$R_s = Segment(E_n, V), \tag{2}$$

where V is the RGB image from observation, R_s denotes the results of semantic segmentation, including masks, boxes, and labels of each semantic element. $Segment()$ represents the semantic segmentation process by Ground-SAM model (Bousselham et al. 2024). Based on the semantic segmentation results R_s and the depth image D, the world coordinates corresponding to each semantic element can be obtained.

For each grid (i, j, k) in the semantic map S, we count all the semantic labels it contains, and select the most frequently occurring semantic label as the semantic representation of that grid:

$$S(i,j,k) = \arg\max_{c \in E_n} \sum_{(u,v) \in \text{pixels in } (i,j,k)} \mathbb{I}(L(u,v) = c), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $L(u, v)$ is the semantics of the pixel (u, v) stored in the results of semantic segmentation R_s, and c is the semantic category. This step is executed at each observation of the agent to achieve dynamic updates of the semantic map.

4.3 Attraction-Driven Target Reasoning

Based on the semantic map, we represent the agent's estimation of the target's position by constructing a 3D cognitive map. Additionally, a denoising mechanism is applied to eliminate interference from non-target objects during the search process.

3D Cognitive Map. The 3D cognitive map C is a 3D grid map that is equal in size to the semantic map S. We employ an MLLM to measure how strongly an object's semantics attract the UAV agent, i.e., how relevant the object is to the target. For each semantic category c in S, the attraction value is computed as:

$$A = \text{MLLM}(Prompt_{att}, c \in S). \tag{4}$$

By calculating the attraction values $A(S_{i,j,k})$ for each grid (i, j, k) in the semantic map, we can assign these values to the corresponding grids in the cognitive map:

$$C_{i,j,k} = A(S_{i,j,k}). \tag{5}$$

Denoising Mechanism. A mirrored cognitive map C' is created to keep track of whether each grid has been recognized by the UAV agent. The state of each grid in C' is represented as follows:

- $C'(i, j, k) = 1$ if the grid (i, j, k) has not been recognized.
- $C'(i, j, k) = 0$ if the grid (i, j, k) has been recognized.

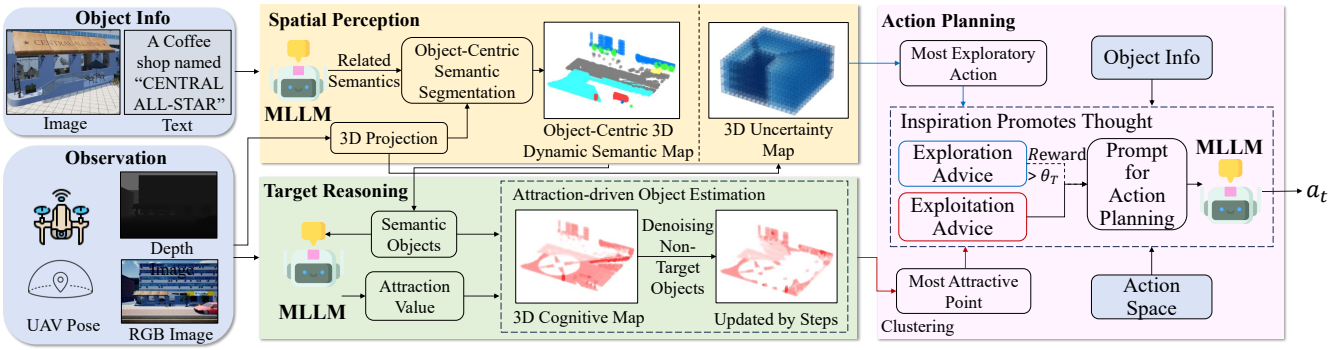

Figure 3. Overview of the agentic method–PRPSearcher.

When the UAV agent performs an observation action, it leverages its current position and viewing angle to determine which grid cells in the cognitive map are visible. For each visible grid (i, j, k), if it is within the distance defined by the step size of the agent, it is updated in the mirrored cognitive map C' as recognized: $C'(i, j, k) = 0$.

To enhance the quality of the cognitive map by filtering out noise from recognized areas, we apply a denoising process using the mirrored cognitive map, formulated as below:

$$C_{i,j,k} = C_{i,j,k} \cdot C'(i, j, k). \tag{6}$$

4.4 E-E Balanced Action Planning

To find the target with higher efficiency and success rate, we require the UAV agent to achieve an exploration-exploitation balance in action planning.

3D Uncertainty Map. Each cell (i, j, k) of the 3D uncertainty map is associated with an uncertainty value $U_{i,j,k} \in [0, 1]$. At the start of the search, all cells have an uncertainty value of 1, indicating complete uncertainty.

A UAV agent performs an observation at position $\mathbf{p} = (X, Y, Z)$ and orientation $\mathbf{o} = (o_x, o_y, o_z)$. Based on the current position and orientation, the set of visible grid cells $\mathcal{V}$ is computed. For each visible cell $(i, j, k) \in \mathcal{V}$, we attenuate its uncertainty $U_{i,j,k}$ based on distance. The uncertainty of different faces of a cell is calculated independently. The attenuation function $f(d)$ is defined as:

$$f(d) = 1 - e^{-\alpha \cdot d}, \tag{7}$$

where $d = \sqrt{(X - x_i)^2 + (Y - y_j)^2 + (Z - z_k)^2}$ is the Euclidean distance from the grid cell (i, j, k) to the agent's position $\mathbf{p}$, α is the attenuation coefficient, controlling the rate at which uncertainty decreases with distance. Each time the agent performs an observation, the above process is repeated, and the 3D uncertainty map is updated as follows:

$$U^{\text{new}}_{i,j,k} = \begin{cases} U^{\text{old}}_{i,j,k} \cdot f(d) & \text{if } (i, j, k) \in \mathcal{V} \\ U^{\text{old}}_{i,j,k} & \text{otherwise} \end{cases}. \tag{8}$$

Exploration Advice. To model the exploration process with the 3D uncertainty map, we define a reward function that quantifies the reduction in uncertainty achieved by each potential action within the agent's action space. The reward $R_{exploration}(a)$ for an action a is defined as follows.

$$R_{exploration}(a) = \sum_{(i,j,k) \in \mathcal{V}_a} \left(U^{\text{old}}_{i,j,k} - U^{\text{new}}_{i,j,k}\right), \tag{9}$$

where $U^{\text{new}}_{i,j,k}$ is the updated uncertainty for grid cell (i, j, k) after executing action a, computed by Formula 8. The action that maximizes the reward can be formulated as:

$$a^*_{exploration} = \arg\max_{a \in \mathcal{A}} Reward(a), \tag{10}$$

where $a^*_{exploration}$ is the exploration advice for the agent.

Exploitation Advice. The 3D cognitive map reflects the "attraction" of these semantic elements to the search object. Areas with the highest attraction values are the most likely locations for the target object. Let $\mathcal{G}$ be the set of high-relevance grids, defined as:

$$\mathcal{G} = \{(i, j, k) \mid C_{i,j,k} = max(C_{i,j,k})\}. \tag{11}$$

By using the DBSCAN clustering method (Schubert et al. 2017), several clusters $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n$ can be identified as high-relevance regions. For the largest cluster $\mathcal{C}_m$, the center point $\mathbf{p}_m = (X_m, Y_m, Z_m)$ is calculated as the target point for the exploitation process. The action $a^*_{exploitation}$ that navigates to the point $\mathbf{p}_m$ is the generated exploitation advice for the UAV agent.

IPT-based E-E Balanced Planning. When humans search for objects, they typically begin by considering the most likely locations of the target and then investigate those areas thoroughly. During the process, spontaneous thoughts such as "There's a place I haven't checked yet" often arise—this type of inspiration helps avoid overlooking potential locations. Such behavior reflects a natural balance between exploration and exploitation in human cognition. Motivated by this insight, we replicate this cognitive process by proposing the IPT prompting mechanism.

This mechanism integrates exploitation advice as long-term guidance into the agent's action planning prompt to continuously guide the agent in finding and identifying known objects. In contrast, exploration advice will be selectively incorporated into the prompt in the form of "Inspiration". To facilitate this, we introduce a threshold θ_T to

Method	Easy Tasks				Hard Tasks				Total Tasks			
	SR↑	MSS↓	SPL↑	NE↓	SR↑	MSS↓	SPL↑	NE↓	SR↑	MSS↓	SPL↑	NE↓
Human	**85.45**	**17.40**	**76.58**	**20.74**	**70.61**	**17.13**	**64.30**	**47.84**	**78.68**	**17.26**	**70.92**	**32.90**
RE	10.30	49.35	6.90	89.00	5.05	74.40	2.47	182.94	7.93	60.97	4.90	131.41
FBE	13.64	39.47	10.04	97.48	8.07	59.57	6.34	195.98	11.07	48.62	8.33	142.33
L3MVN	26.82	34.51	21.54	87.89	7.13	59.90	4.02	187.06	17.87	46.05	13.57	132.90
WMNav	20.62	38.54	18.05	75.06	7.75	72.45	5.10	146.07	14.82	54.02	12.20	106.99
STMR	32.68	34.25	23.86	66.07	20.96	57.24	13.23	121.09	27.35	44.70	19.03	91.33
PRPSearcher w/o exploitation	16.36	38.87	13.25	95.25	3.28	60.32	2.20	143.09	10.41	48.63	8.23	116.57
PRPSearcher w/o exploration	60.47	30.22	47.89	50.19	35.82	45.94	28.73	116.65	49.19	37.37	39.06	80.16
PRPSearcher	66.32	28.85	49.82	43.62	38.31	42.89	29.76	90.67	53.50	35.26	40.57	64.86

Table 1. Performance comparisons with SOTA baselines on CityAVOS benchmark.

Method	Total			
	SR↑	MSS↓	SPL↑	NE↓
free-prompt	50.52	37.89	38.11	84.03
human-design	38.46	41.41	30.27	105.68
object-centric	**53.50**	**35.26**	**40.57**	**64.86**

Table 2. Ablation study of the object-centred 3D dynamic semantic map for PRPSearcher.

θ_T	SR	MSS	SPL	NE	N_θ
1	49.19	37.37	39.06	80.16	0.00
0.5	49.38	38.44	38.83	78.62	0.00
0.2	51.38	35.30	39.89	67.78	4.37
0.1	**53.50**	**35.26**	**40.57**	**64.86**	8.62
0.05	43.99	41.71	32.48	89.47	26.09
0.02	38.20	44.59	30.05	97.58	44.59
0	38.37	44.82	29.90	95.11	44.82

Table 3. Ablation study of the IPT prompting mechanism for PRPSearcher.

assess whether the benefits of exploration actions are significant enough. When the benefits exceed this threshold, exploration advice will be added to the planning prompt to remind the agent to shift its focus toward exploring unknown spaces.

$$Prompt_{plan} = Advice_{exploit} + (Reward(a^*) > \theta_T) \cdot Advice_{explore} \quad , \quad (12)$$

where $\mathbb{I}()$ is the Boolean function, $\mathbb{I}(Reward(a^*) > \theta_T) = 1$ when $Reward(a^*) > \theta_T$ is true, otherwise $\mathbb{I}(Reward(a^*) > \theta_T) = 0$.

5 Experiments

5.1 Experiment Setup

Evaluation Metrics. We adopt four standard metrics to measure the performance, i.e., Success Rate (SR), Success Rate Weighted by Inverse Path Length (SPL) (Wu et al. 2024), Mean Search Steps (MSS) (Zhao et al. 2022), and Navigation Error (NE) (Ramakrishnan et al. 2022; Liu et al. 2023).

Implementation Details. For PRPSearcher, the input image is resized to 640 × 480 for convenient processing. The MLLMs leveraged for visual analysis and reasoning during the spatial perception, target reasoning, and action planning phases is GPT-4o in this work. The dataset used for the experiment is CityAVOS, and the platform is the EmbodiedCity modified for AVOS. Due to API limitations, 605 tasks (25%) are randomly selected from the CityAVOS dataset for extensive experiments.

Baselines. We compare against a diverse range of baseline methods, including basic methods such as Random Exploration (RE) and Frontier-Based Exploration (FBE), indoor image goal navigation methods such as L3MVN (Yu, Kasaei, and Cao 2023) and WMNav (Nie et al. 2025), and the outdoor approach STMR (Gao et al. 2024b). In addition, we invite five postgraduates with drone-operating expertise to participate in the experiment as the Human Agent.

5.2 Comparisons with SOTA Methods

As shown in Tab. 1, our proposed approach significantly outperforms the baseline methods (on average: +37.69% SR, +28.96% SPL, -30.69% MSS, and -46.40% NE) in total tasks, demonstrating the effectiveness of the designed mechanisms and the constructed maps. Some important observations can be obtained:

- **Basic Methods.** The random exploration method and frontier-based exploration methods perform poorly on tasks of various difficulties. As both types of methods are blind space exploration approaches, their performance reflects that the AVOS tasks cannot be solved through basic space exploration patterns.
- **Indoor ImageNav Methods.** The L3MVN method has increased the success rate (SR) by 13.18% on the simple difficulty task set compared to the basic method. The WMNav method achieves good performance on hard tasks through a curiosity mechanism. These results not only highlight the importance of understanding semantics for AVOS tasks but also reflect the limitations of indoor methods in city environments.
- **Outdoor Methods.** STMR facilitates the storage of outdoor semantic information by constructing a Top-down

map in the air. Meanwhile, it enhances the ability of agent action planning based on the CoT reasoning. Therefore, the SR in hard tasks can reach 20.96%. This result reflects the importance of the reasoning ability of agents in highly difficult tasks.

- **Human Agent.** Human agents performed best in all task classifications. With the increase of task difficulty, the performance of human agents also decreased slightly, which indicates that there are certain challenges for human beings to successfully complete AVOS tasks.

Overall, the comparisons with baseline models reveal that *excluding interference from redundant object information during semantic extraction and effectively distinguishing target-like objects in urban environments* are crucial for improving search efficiency. Additionally, achieving a higher success rate in AVOS tasks *depends on striking an optimal balance between exploitation and exploration*.

5.3 Ablation Study

Effect of the object-centred 3D dynamic semantic map. We design two other semantic segmentation prompts: free-prompt and human-designed. The former does not provide prompts to the semantic segmentation model, allowing the model to determine the segmentation targets on its own. The latter involves humans actively setting the prompts, without further adjustments for different tasks. The experimental results indicated in Tab. 2 show that the human-designed semantic segmentation prompts achieve the worst experimental results, and the performance of the free-prompt is slightly lower than our method. The results demonstrate that our proposed object-centric semantic map can reason about the most relevant semantics for different task, enabling the agent to locate the target object more efficiently.

Effect of the exploration and exploitation design. We validate the effectiveness of our proposed 3D uncertainty map and 3D cognitive map by separately ablating exploration and exploitation suggestions for the agent. The experimental results are shown in Tab. 1. The **PRPSearcher w/o exploration** method still maintains a high performance, but the absence of suggestions for exploring unknown spaces results in a decline in both SR and SPL. Conversely, the **PRPSearcher w/o exploitation** method performs poorly, yet still outperforms the FBE method.

Effect of the IPT prompting mechanism. The IPT prompt mechanism is designed to balance exploration and exploitation during the agent's action planning. A key parameter in this mechanism, denoted as θ_T, controls the frequency of exploration advice provided to the agent. We conducted numerical experiments to evaluate the impact of different θ_T values, and the results are summarized in Table 3. When $\theta_T = 0.5$ or $\theta_T = 1$, the number of exploration prompts received by the agent drops to zero ($N_\theta = 0$), leading to a decline in performance due to the lack of exploratory guidance. Conversely, when $\theta_T = 0$, the agent receives exploration advice at every decision step, which overwhelms its decision-making process and significantly reduces the SR. Results show that $\theta_T = 0.1$ is the optimal setting in balancing the exploration and exploitation.

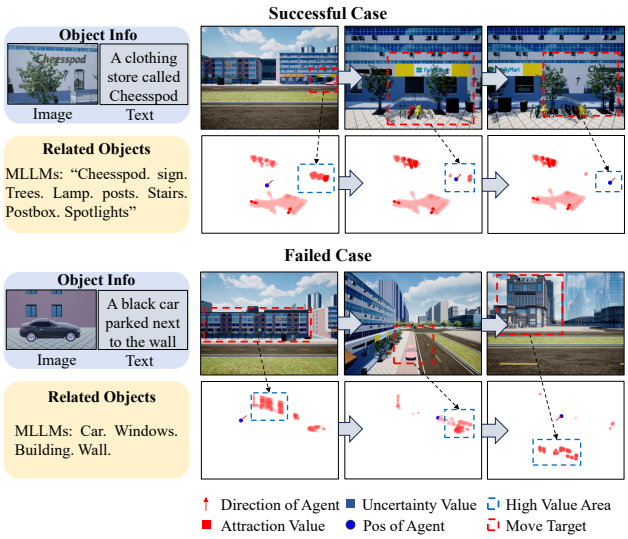

Figure 4. Two case studies of PRPSearcher.

5.4 Case Study

As shown in Fig. 4, we present a successful case and a failed case of PRPSearcher. In the successful case, the agent first explores its surroundings based on exploration advice. Guided by exploitation suggestions from the 3D cognitive map, the agent searches along a row of shops with nearby trees under a building. The denoising mechanism helps the agent remain focused, ignoring similar but irrelevant shops, ultimately locating the target. Crucially, associating trees with the target enhances efficiency by directing the search toward the correct area.

In a representative failure case, the target is "A black car parked next to the wall." Sparse visual information results in limited semantic cues: "Car, Windows, Building, Wall." Initially, PRPSearcher directs the UAV agent toward buildings. However, the search ultimately fails due to step limitations. Among evaluated baseline methods, only human agents and the FBE method succeed in this scenario. This highlights PRPSearcher's limitations in spatial exploration efficiency and the gap between its semantic reasoning and human capabilities.

6 Conclusion

In this study, we introduced a Visual Object Search (AVOS) task for UAVs in urban environments. We formalized the AVOS task and introduced the CityAVOS benchmark. To tackle this task, we proposed an agentic method, namely PRPSearcher, which mimics human perception, reasoning, and planning through specialized semantic, cognitive, and uncertainty maps. The experimental results demonstrate PRPSearcher's significant advantages over existing methods. This work represents a substantial step towards enabling embodied UAV target search capabilities in complex city spaces. In the future, we will attempt to further improve PRPSearcher by incorporating collaborative human-agent or multi-agent strategies to handle more complex AVOS tasks.

Acknowledgments

We thank all co-authors and team members for their valuable contributions to this project. This work is supported by the National Natural Science Foundation of China (72501291, 62306329, 62173337) and the Natural Science Foundation of Hunan Province, China (Grant No. 2023JJ40676).

References

Al-Halah, Z.; Ramakrishnan, S. K.; and Grauman, K. 2022. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17031–17041.

Bousselham, W.; Petersen, F.; Ferrari, V.; and Kuehne, H. 2024. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3828–3837.

Cai, W.; Huang, S.; Cheng, G.; Long, Y.; Gao, P.; Sun, C.; and Dong, H. 2024a. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5228–5234. IEEE.

Cai, Z.; Cardenas, C. R.; Leo, K.; Zhang, C.; Backman, K.; Li, H.; Li, B.; Ghorbanali, M.; Datta, S.; Qu, L.; et al. 2024b. NEUSIS: A Compositional Neuro-Symbolic Framework for Autonomous Perception, Reasoning, and Planning in Complex UAV Search Missions. *arXiv preprint arXiv:2409.10196*.

Chaplot, D. S.; Salakhutdinov, R.; Gupta, A.; and Gupta, S. 2020. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12875–12884.

Dorbala, V. S.; Mullen, J. F.; and Manocha, D. 2023. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5): 4083–4090.

Döschl, B.; and Kiam, J. J. 2024. Say-REAPEx: An LLM-Modulo UAV Online Planning Framework for Search and Rescue. In *2nd CoRL Workshop on Learning Effective Abstractions for Planning*.

Du, H.; Li, L.; Huang, Z.; and Yu, X. 2023. Object-goal visual navigation via effective exploration of relations among historical navigation states. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2563–2573.

Gao, C.; Zhao, B.; Zhang, W.; Mao, J.; Zhang, J.; Zheng, Z.; Man, F.; Fang, J.; Zhou, Z.; Cui, J.; et al. 2024a. EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment. *arXiv preprint arXiv:2410.09604*.

Gao, Y.; Wang, Z.; Jing, L.; Wang, D.; Li, X.; and Zhao, B. 2024b. Aerial Vision-and-Language Navigation via Semantic-Topo-Metric Representation Guided LLM Reasoning. *arXiv preprint arXiv:2410.08500*.

Hou, Y.; Zhao, J.; Zhang, R.; Cheng, X.; and Yang, L. 2023. UAV swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Transactions on Intelligent Vehicles*, 9(1): 568–578.

Huang, J.; Xu, Y.; Wang, Q.; Wang, Q. C.; Liang, X.; Wang, F.; Zhang, Z.; Wei, W.; Zhang, B.; Huang, L.; et al. 2025. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*.

Kim, N.; Kwon, O.; Yoo, H.; Choi, Y.; Park, J.; and Oh, S. 2023. Topological semantic graph memory for image-goal navigation. In *Conference on Robot Learning*, 393–402. PMLR.

Li, P.; Wu, K.; Fu, J.; and Zhou, S. 2025. REGNav: Room Expert Guided Image-Goal Navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4860–4868.

Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15384–15394.

Liu, Y.; Chen, W.; Bai, Y.; Liang, X.; Li, G.; Gao, W.; and Lin, L. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*.

Mezghan, L.; Sukhbaatar, S.; Lavril, T.; Maksymets, O.; Batra, D.; Bojanowski, P.; and Alahari, K. 2022. Memory-augmented reinforcement learning for image-goal navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3316–3323. IEEE.

Mousavian, A.; Toshev, A.; Fišer, M.; Košecká, J.; Wahid, A.; and Davidson, J. 2019. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, 8846–8852. IEEE.

Nie, D.; Guo, X.; Duan, Y.; Zhang, R.; and Chen, L. 2025. WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation. *arXiv preprint arXiv:2503.02247*.

Qin, Z.; Wang, L.; Wang, Y.; Zhou, S.; Hua, G.; and Tang, W. 2025. RSRNav: Reasoning Spatial Relationship for Image-Goal Navigation. *arXiv preprint arXiv:2504.17991*.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18890–18900.

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; and Xu, X. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3): 1–21.

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

She, R.; and Ouyang, Y. 2021. Efficiency of UAV-based last-mile delivery under congestion in low-altitude air. *Transportation Research Part C: Emerging Technologies*, 122: 102878.

Wang, X.; Yang, D.; Wang, Z.; Kwan, H.; Chen, J.; Wu, W.; Li, H.; Liao, Y.; and Liu, S. 2024. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*.

Wu, C.; Ju, B.; Wu, Y.; Lin, X.; Xiong, N.; Xu, G.; Li, H.; and Liang, X. 2019. UAV autonomous target search based on deep reinforcement learning in complex disaster scene. *IEEe Access*, 7: 117227–117245.

Wu, P.; Mu, Y.; Wu, B.; Hou, Y.; Ma, J.; Zhang, S.; and Liu, C. 2024. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*.

Wu, Y.; Nie, M.; Ma, X.; Guo, Y.; and Liu, X. 2023. Co-evolutionary algorithm-based multi-unmanned aerial vehicle cooperative path planning. *Drones*, 7(10): 606.

Xing, L.; Fan, X.; Dong, Y.; Xiong, Z.; Xing, L.; Yang, Y.; Bai, H.; and Zhou, C. 2022. Multi-UAV cooperative system for search and rescue based on YOLOv5. *International Journal of Disaster Risk Reduction*, 76: 102972.

Yin, H.; Xu, X.; Zhao, L.; Wang, Z.; Zhou, J.; and Lu, J. 2025. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19057–19066.

Yu, B.; Kasaei, H.; and Cao, M. 2023. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3554–3560. IEEE.

Zhang, J.; Li, M.; Xu, Y.; He, H.; Li, Q.; and Wang, T. 2025. StrucGCN: Structural enhanced graph convolutional networks for graph embedding. *Information Fusion*, 117: 102893.

Zhao, N.; Lu, W.; Sheng, M.; Chen, Y.; Tang, J.; Yu, F. R.; and Wong, K.-K. 2019. UAV-assisted emergency networks in disasters. *IEEE Wireless Communications*, 26(1): 45–51.

Zhao, Y.; Chen, B.; Wang, X.; Zhu, Z.; Wang, Y.; Cheng, G.; Wang, R.; Wang, R.; He, M.; and Liu, Y. 2022. A deep reinforcement learning based searching method for source localization. *Information Sciences*, 588: 67–81.

Zhao, Y.; Xu, K.; Zhu, Z.; Hu, Y.; Zheng, Z.; Chen, Y.; Ji, Y.; Gao, C.; Li, Y.; and Huang, J. 2025. Cityeqa: A hierarchical llm agent on embodied question answering benchmark in city space. *arXiv preprint arXiv:2502.12532*.