

# 基于多模态大模型推理的无人机未知环境目标搜索

张雪波<sup>1,2</sup>, 马哲<sup>1,2</sup>, 张世勇<sup>1,2†</sup>, 王子玉<sup>1,2</sup>, 奚浩博<sup>1,2</sup>, 张智勇<sup>1,2</sup>, 袁明星<sup>1,2</sup>

(1. 南开大学人工智能学院, 天津 300350; 2. 南开大学机器人与信息自动化研究所, 天津 300350)

**摘要:** 面向未知环境下的零样本目标搜索任务, 提出一种多模态大模型推理与自主探索相融合的无人机导航方法. 首先, 针对多模态大模型难以处理三维数据的问题, 提出了一种空间-视觉逆映射方法, 通过构建具备显式三维坐标约束的场景图像作为多模态大模型输入, 赋予多模态大模型同时理解场景图像与定位关键区域的能力. 然后, 针对现有目标搜索方法泛化性差的问题, 设计了一种蕴含“辨识—评估—转移”逻辑的提示词, 引导无人机实现跨场景条件下的零样本目标搜索. 最后, 针对现有目标搜索方法存在显著仿真—真实差距的问题, 在无人机自主探索框架中引入几何—语义异步增益融合机制与动态评估策略, 实现“空间自主探索”与“语义规律利用”自适应平衡. 仿真结果表明, 在三类 Gazebo 场景中, 所提方法在路径长度、搜索时间及成功率等指标上均明显优于基线方法. 此外, 室外未知场景实验验证了所提方法在零样本目标搜索任务中的有效性.

**关键词:** 无人机; 目标搜索; 自主探索; 多模态感知; 物体目标导航; 零样本导航

中图分类号: TP273 文献标志码: A

DOI: 10.13195/j.kzyjc.2025.1335

引用格式: 张雪波, 马哲, 张世勇, 等. 基于多模态大模型推理的无人机未知环境目标搜索 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

## Target searching in unknown environments by an unmanned aerial vehicle based on multimodal large model inference

ZHANG Xue-bo<sup>1,2</sup>, MA Zhe<sup>1,2</sup>, ZHANG Shi-yong<sup>1,2†</sup>, WANG Zi-yu<sup>1,2</sup>, XI Hao-bo<sup>1,2</sup>, ZHANG Zhi-yong<sup>1,2</sup>, YUAN Ming-xing<sup>1,2</sup>

(1. College of Artificial Intelligence, Nankai University, Tianjin 300350, China; 2. Institute of Robotics and Automatic Information Systems, Nankai University, Tianjin 300350, China)

**Abstract:** This paper proposes a navigation method for unmanned aerial vehicles (UAVs) that integrates multimodal large model reasoning with autonomous exploration technique for tackling zero-shot target searching in unknown environments. First, to address the difficulty of multimodal large models in directly processing 3-D data, a space-vision inverse mapping approach is introduced. By constructing scene images with explicit 3D coordinate constraints as inputs, the multimodal large model is endowed with the capability to simultaneously understand scene imagery and localize key regions. Second, to overcome the poor generalization of existing target searching methods, a prompting strategy embedding a “recognition-evaluation-transfer” logic is designed to guide the UAV in performing zero-shot target searching across diverse scenarios. Finally, to mitigate the significant sim-to-real gap that existing methods are currently struggling with, a geometric-semantic asynchronous gain fusion mechanism and a dynamic evaluation strategy are incorporated into a UAV autonomous exploration framework, achieving an adaptive balance between “spatial autonomous exploration” and “semantic regularity exploitation”. Simulation results in three Gazebo environments show that the proposed method significantly outperforms baseline approaches in terms of path length, search time, and success rate. In addition, real-world experiments conducted in unknown outdoor environments demonstrate the effectiveness of the proposed method on zero-shot target searching tasks.

**Keywords:** unmanned aerial vehicles; target searching; autonomous exploration; multimodal perception; object-goal navigation; zero-shot navigation

收稿日期: 2025-12-25; 录用日期: 2026-05-13.

基金项目: 国家重点研发计划青年科学家项目 (2022YFB4701800); 国家自然科学基金项目 (62303249); 京津冀基础研究合作专项项目 (24JCZXCJ00390); 中国博士后科学基金项目 (2024M751526).

†通信作者. E-mail: zhangshiyong@nankai.edu.cn.

## 0 引言

未知环境目标搜索可抽象为在三维空间中确定一个或多个按某种特定规律分布的目标点坐标. 当这种分布规律不可知时, 目标分布则被视为完全随机分布, 需通过探索整个未知空间进行搜索; 当目标分布规律能够通过某种手段推断时, 可据此优化未知空间的探索流程, 从而提升搜索效率与成功率. 因此, 未知环境目标搜索本质上是在“空间自主探索”和“语义规律利用”之间不断进行权衡的过程.

### 0.1 聚焦于空间探索的目标搜索

目前, 关于三维未知空间探索的相关研究已经十分广泛. 针对无人机自主探索任务的主流方法主要分为基于采样的方法 (sampling-based approaches)<sup>[1-4]</sup> 与基于边界的方法 (frontier-based approaches)<sup>[5-7]</sup>. 基于采样的方法在已知空间采样视点并计算信息增益, 用于导航决策. 该方法优势是可以根据不同任务目的灵活设定信息增益计算公式. 基于边界的方法将已知空间与未知空间的边界 (frontier) 作为候选区域, 根据边界的价值进行运动规划, 能够更直接地完成对探索区域的拓展.

无人机自主探索任务主要关注未知空间的占据信息, 其目标是在尽可能短的时间内完成空间覆盖. 而以空间探索为主的目标搜索任务, 除了需要判断空间是否被物体或建筑占据外, 还需进一步识别其中是否存在兴趣目标. 基于物体语义的探索方法 (下文称为 Semantic 方法)<sup>[8]</sup> 将未知环境自主探索任务的目标拓展至具体物体检测与重建. 具体而言, 该方法同时构建截断符号距离场物体地图与添加有体素最近观察距离的八叉树占据背景地图, 在边界与物体附近进行视点采样, 然后分别使用三个射线投射在两个地图中计算目标物体增益、背景边界增益与未知区域信息熵, 三者融合后通过滑动窗口计算候选方位价值. Semantic 方法虽然关注目标物体, 但是仅当物体已经出现时才会被计算价值, 在其未见到目标物体时完全根据几何结构信息进行探索.

为了无缝地检测空间占据情况并对潜在目标进行精细地视觉识别, Star-Searcher 方法<sup>[9]</sup> 利用激光雷达快速获取周围空间的几何占据信息, 并通过相机对潜在目标区域进行精细检查. 具体而言, 该方法基于可见性视点簇设计了层次化规划策略, 一方面高效遍历全局视点簇, 另一方面在局部执行精细路径规划, 以保证对潜在目标区域的完整视觉检查. 同时, 提出历史感知全局路径规划机制, 在重规划时继承先前规划结果, 保持全局路径的一致性并减少不必

要的往复运动. 然而该方法没有物体性 (objectness) 认知, 即不会判断某个区域是否可能包含完整的、独立的物体, 或是判断某区域属于与物体相对应的墙壁、地面、天空等背景.

上述无人机自主探索与目标搜索算法通常仅依据几何结构信息规划探索行为, 未充分利用关于环境常见结构的先验认知来估计目标位置. 因此, 在面向目标搜索任务时, 这类方法难以对场景进行语义层面的理解与目标存在概率评估, 缺乏与目标相关物体和背景的显式关注.

### 0.2 聚焦于规律利用的目标搜索

与前述聚焦于空间探索的目标搜索不同, 本节关注另一类利用环境语义规律推断目标潜在位置的物体目标导航 (object-goal navigation) 研究<sup>[10-17]</sup>. 此类方法通过利用场景语义线索实现推理式导航, 提高目标搜索的效率.

现有物体目标导航通常在结构明确的室内环境中开展. 例如, 执行“寻找沙发”等任务时, 系统会根据“沙发—客厅”之间的语义关联优先选择导航区域. 先进方法<sup>[12]</sup> 通常由感知与语义理解、建图与语义记忆、探索与目标选择以及导航与执行控制子模块构成, 通过构建二维语义占据图并根据区域语义价值进行导航决策, 实现基于语义信息的引导式目标搜索. 然而, 这类语义规律利用方法在无人机未知环境目标搜索任务中面临三方面根本性限制: 第一, 环境泛化性受限. 现有方法通常依赖平面场景<sup>[13]</sup> 与固定区域划分假设, 如“卧室—厨房—客厅”等语义结构, 但在尺度大、边界模糊的室外三维环境中并不存在清晰区域划分, 难以直接迁移<sup>[14]</sup>. 第二, 语义表达闭集化. 大多依赖 YOLO 等有限类别的语义检测模型, 仅关注“物体性”语义, 难以处理开放词汇目标及在户外任务中具有关键意义的背景语义, 如河岸、道路延伸方向等. 第三, 忽视自由探索能力. 现有方法局限于二维运动空间中的导航决策, 缺乏语义信息与三维空间的联合表示, 不具备三维空间运动规划与探索能力.

为增强规律利用能力, 近期研究开始引入多模态大模型, 以实现开放词汇理解与高层语义推理<sup>[15-16]</sup>. 相关方法或利用语言模型生成描述以联系开放词汇与闭集标签, 或基于模型输出对场景区域进行语义价值评估并用于导航决策. 然而, 由于多模态大模型无法直接处理三维空间数据, 现有方法仍依赖二维投影式语义价值分配. 例如, 文献 [17] 首先利用 BLIP-2 (Bootstrapped Language - Image Pretraining -

2) 计算当前 RGB 观测与目标文本提示间的语义相似度, 经视角加权后投射至整个二维扇形视野当中, 形成语义价值图用于导航决策. 文献 [12] 在文献 [17] 基础上增加自适应探索策略来选取下一导航点. 上述方法语义价值的分配是在整个观测视野中进行的, 这种平面化投影方式容易在空间上产生语义定位误差, 尤其对关键目标物体的实际位置估计不准. 此外, 当应用于大尺度三维场景时, 二维投影式语义价值分配难以准确反映远近关系与空间结构, 使得语义信息在地图中的分布过于模糊, 限制复杂环境下的导航精度与泛化能力.

本文提出一种“多模态大模型推理”与“自主探索规划”有机融合的无人机目标搜索框架:

通过显式三维空间编码、开放词汇语义建图以及探索-利用自适应融合, 实现对目标分布规律的动态估计与利用. 所提框架突破了现有算法缺乏开放语义认知、难以匹配无人机运动特性的瓶颈, 显著提升了目标搜索效率与环境适应性. 主要贡献如下:

1) 针对多模态大模型难以处理三维数据的问题, 提出了一种空间-视觉逆映射方法, 构建具备显式三维坐标约束的场景图像作为多模态大模型输入, 增强多模态大模型的空间位置信息推理能力.

2) 针对现有目标搜索方法泛化性差、场景单一的问题, 设计了一种蕴含“辨识—评估—转移”逻辑的自然语言提示词, 实现目标语义线索的持续追踪与零样本目标更新.

3) 针对现有目标搜索方法存在显著仿真—真实差距的问题, 将目标线索推理与定位模块及无人机自主探索算法相融合, 继承后者在真实环境部署的可靠性. 融合通过几何-语义异步增益融合机制与动态评估策略完成, 使规划系统在“空间自主探索”与“语义规律利用”间自适应切换.

### 1 任务描述

未知环境目标搜索任务期望无人机在最短时间内搜索并确定三维未知空间中的目标位置. 在搜索过程中, 无人机根据机载传感器, 如激光雷达、RGB-D 相机以及惯性测量单元等, 在实现自身定位的同时进行目标的检测与定位. 本文做出假设如下:

- 1) 无人机系统具备精确的同时定位与建图能力;
- 2) 无人机系统可与部署在地面站或服务器的大模型进行通信;

3) 海量样本预训练的大模型参数中隐含了与环境结构及物体语义相关的目标位置分布规律.

### 2 方法介绍

针对未知环境下的目标搜索问题, 无人机需在缺乏地图与目标位置等先验信息的条件下, 通过视觉感知与语义推理估计目标的位置, 以实现高效自主搜索. 其核心为在不确定空间中推断目标位置分布, 并据此规划搜索路径.

如图 1 所示, 首先, 为便于空间建模与导航规划, 将搜索空间  $\Omega \subset \mathbb{R}^3$  均匀划分, 得到  $N$  个可探索区域<sup>[18]</sup> (Explorable Region of Interest, 简称 EROD):

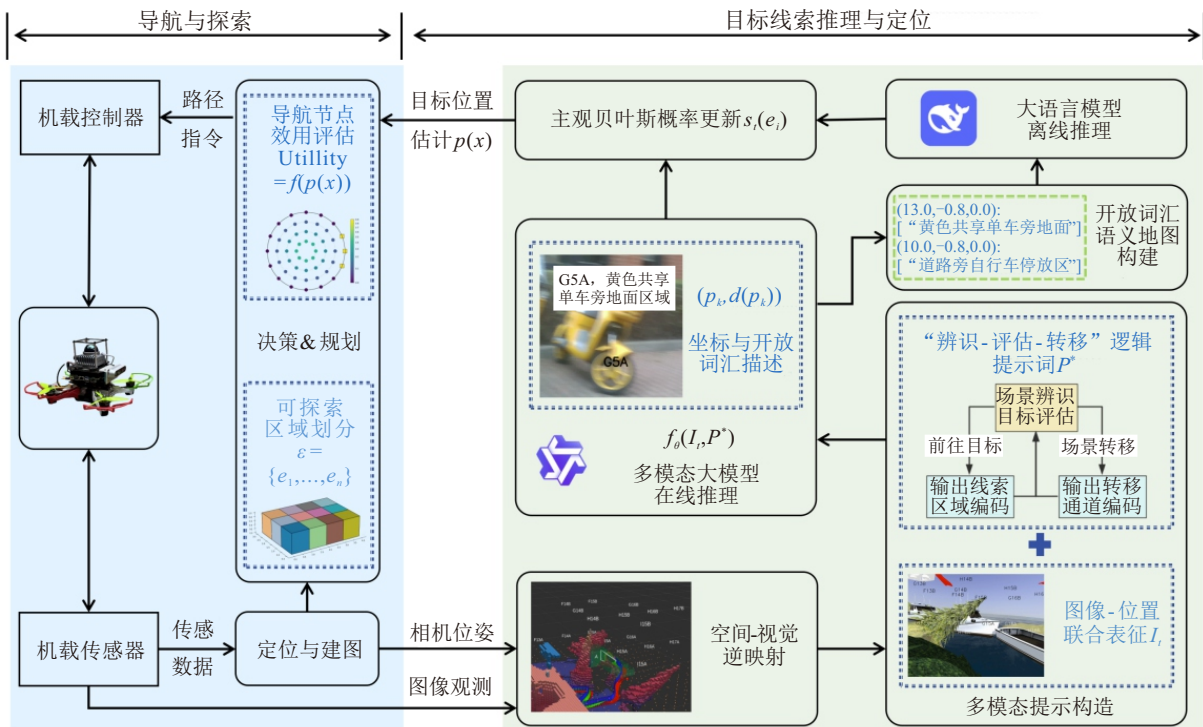


图1 方法结构框图

$$\mathcal{E} = \{e_i \mid e_i \subset \Omega, i = 1, \dots, N\}. \quad (1)$$

随后, 在每个 EROI 中均匀采样生成视点, 作为无人机的候选导航点. 同时, 为每个 EROI 赋予两个核心属性, 定义如下:

- 几何价值: 由探索算法 FSMP<sup>[19]</sup> 计算得出, 衡量区域在几何可视性与覆盖性上的重要性.
- 语义价值: 由多模态大模型在语义空间中进行推理得出, 用于反映区域内存在目标的概率.

目标线索推理与定位模块首先对无人机获取的前向彩色图像执行空间视觉逆映射, 得到具有空间坐标信息的图像  $I_t$ ; 随后将该编码图像与零样本结构化提示词  $P^*$  一同输入视觉语言大模型 (Vision Language Model, VLM)  $f_\theta$ . 模型输出目标相关区域的“<坐标, 开放语义>”对, 即:

$$f_\theta(I_t, P^*) \rightarrow \{(\mathbf{p}_k, l_k)\}, \quad (2)$$

其中  $\mathbf{p}_k \in \mathbb{R}^3$  为 EROI 中心坐标,  $l_k$  为目标相关语义描述. 基于此, 系统在线获得该坐标区域语义评分, 并构建开放词汇地图. 进而, 利用大语言模型 (Large Language Model, LLM) 对地图进行推理, 以评估大尺度区域内目标的潜在概率.

各个时刻的评估结果使用主观贝叶斯方法进行更新, 得到每个 EROI 的语义价值:

$$\hat{s}_t(e_i) = \text{Bayes}(s_{t-1}(e_i), \text{SemV}_\theta(e_i)). \quad (3)$$

$\hat{s}_t(e_i)$  表示时刻  $t$  对可探索区域  $e_i$  中存在目标的后验概率估计, 是对目标空间分布  $p(\mathbf{x})$  的近似.

在搜索决策阶段, 系统基于一个效用函数来评估所有候选的 EROI. 该函数综合考虑以下三个因素:

- 至 EROI 的路径代价;
- 该 EROI 的几何价值;
- 该 EROI 的语义价值.

系统选择效用值最高的 EROI 作为下一目标, 并由运动规划模块据此规划无人机运动.

## 2.1 目标线索感知与表示

### 2.1.1 区域划分与几何增益

如前所述, 探索空间  $\Omega \subset \mathbb{R}^3$  事先被均匀划分为一组 EROI, 同时, 在以各 EROI 中心为原点的柱坐标系中进行均匀采样, 生成其对应的候选视点集, 所有视点的偏航角均被设置为指向其所属 EROI 的中心. 为每个 EROI 及视点定义三种状态: 未激活、激活与失效. 初始状态下, 所有 EROI 与视点均为“未激活”. 当某个 EROI 内的体素被传感器观测后, 其状态将切换为“激活”. 随后, 系统将对所有处于“激活”状态的 EROI, 计算其无碰撞且非失效视点的信息增益. 具体地, 从该视点向外发射射线, 并累加所

有射线击中的“边界”体素所贡献的面积; 此处“边界”特指尚未被观测到的体素.

### 2.1.2 空间视觉逆映射

在开放场景下实时获取实例语义并进行定位, 是空间智能面临的一大挑战, 也是物体目标导航的关键前提. 传统方法通常将 RGB 图像通过识别与分割获得语义标签, 再映射至三维空间. 然而, 这类方法缺乏零样本特性且对场景理解不完全. 为此, 本文提出一种将三维空间坐标进行编码并逆向映射至 RGB 图像的空间表示方法. 该方法包括四个步骤: 坐标系转换, 针孔投影成像, 字符编码, 字符遮挡剔除.

#### 1) 坐标系转换

设 EROI 中心坐标点集:

$$P_\mathcal{E} = \left\{ \begin{array}{l} \mathbf{p} = (x, y, z)^\top, \\ x = x_{\min} + i\Delta, i = 0, 1, \dots, N_x, \\ y = y_{\min} + j\Delta, j = 0, 1, \dots, N_y, \\ z = z_{\min} + k\Delta, k = 0, 1, \dots, N_z \end{array} \right\}. \quad (4)$$

其中  $\Delta$  为 EROI 尺寸. 为简化表示, 本文默认采用各向同性划分, 即  $\Delta x = \Delta y = \Delta z = \Delta$ . 需要结合场景先验时, 该划分可自然扩展为各向异性设置 ( $\Delta x$ 、 $\Delta y$ 、 $\Delta z$  分别设定), 例如在楼层结构明显的场景中可取  $\Delta z$  为层高或期望高度层分辨率.

相机位姿由位置向量  $\mathbf{t}^w \in \mathbb{R}^3$  与姿态四元数  $\mathbf{q}$  共同表示. 从相机物理坐标系  $u$  到光学坐标系  $c$  的固定旋转为:

$$R_u^c = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}. \quad (5)$$

将  $\mathbf{q}$  转为旋转矩阵  $R_u^w \in SO(3)$ , 可得到从世界系  $w$  到光学系  $c$  的旋转变换:

$$R_w^c = R_u^c (R_u^w)^\top. \quad (6)$$

基于上述变换, 空间中任一 EROI 中心坐标点  $\mathbf{p}^w \in P_\mathcal{E}$  在光学坐标系  $c$  中的坐标为:

$$\mathbf{p}^c = R_w^c (\mathbf{p}^w - \mathbf{t}^w) = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}. \quad (7)$$

根据任务目的与场景范围, 设置以下超参数以筛选有效的 EROI 中心点: 点的深度须满足  $Z > 0$  (位于相机前方) 且与无人机距离  $d_{\min} < \|\mathbf{p}^c\| < d_{\max}$  (例如无人机前方 3 – 12 m 的范围) 时才被考虑.

#### 2) 针孔投影成像

设相机内参矩阵为:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}. \quad (8)$$

进而,  $\mathbf{p}^c$ 在相机图像中的像素坐标 $(u, v)$ 可由经典针孔投影模型获得:

$$u = \frac{f_x X}{Z} + c_x, v = \frac{f_y Y}{Z} + c_y. \quad (9)$$

若 $(u, v)$ 位于相机分辨率范围内, 则其对应的EROI中心被判定为可见候选点.

### 3) 字符编码规则

采用如**算法1**所示方式将EROI中心坐标映射到简短字符文本. 例如“A1A”表示世界坐标系下点 $\mathbf{p}^w = [x_{\min} \ y_{\min} \ z_{\min}]^T$ 处的EROI中心坐标点. 采用“字母-数字-字母”的混合编码形式, 目的是在图像上实现固定长度、紧凑且易解析的三维索引表达, 其中将中间维度用数字表示, 主要是考虑字母与数字混排可降低相邻编码的视觉混淆与误读概率.

#### 算法1 空间坐标字符编码

**输入:** EROI中心坐标点 $\mathbf{p} = (x, y, z)$ ; 网格分辨率 $\Delta$ ; 边界 $(x_{\min}, y_{\min}, z_{\min})$ .

**输出:** 字符编码 $\mathcal{H}(\mathbf{p})$ .

step 1: 计算索引

- a.  $i_x \leftarrow \left\lfloor \frac{x - x_{\min}}{\Delta} \right\rfloor$ ;
- b.  $i_y \leftarrow \left\lfloor \frac{y - y_{\min}}{\Delta} \right\rfloor + 1$ ;
- c.  $i_z \leftarrow \left\lfloor \frac{z - z_{\min}}{\Delta} \right\rfloor$ .

step 2: 映射为字符并拼接

- a.  $c_x \leftarrow \text{chr}(A + i_x)$ ;
- b.  $c_z \leftarrow \text{chr}(A + i_z)$ ;
- c.  $\mathcal{H}(\mathbf{p}) \leftarrow \langle c_x, i_y, c_z \rangle$ .

step 3: 输出 $\mathcal{H}(\mathbf{p})$ .

### 4) 字符遮挡剔除

为实现编码字符的视觉可区分性, 需对候选字符集 $E_T \subset \{\mathcal{H}(\mathbf{p}) \mid \mathbf{p} \in P_\varepsilon\}$ , 进行遮挡剔除. 其中字符 $\mathcal{H}(\mathbf{p}_i)$ 的框边界记为 $B_i$ . 首先按照字符对应坐标到无人机的欧氏距离对 $E_T$ 降序排列, 接着遍历此序列, 维护一个已接受字符集合 $A \subset \{\mathcal{H}(\mathbf{p}) \mid \mathbf{p} \in P_\varepsilon\}$ . 当且仅当字符 $\mathcal{H}(\mathbf{p}_i)$ 的边界与 $A$ 中所有字符的边界框均无交叠, 即:

$$B_i \cap B_j = \emptyset, \forall \mathcal{H}(\mathbf{p}_j) \in A. \quad (10)$$

将字符 $\mathcal{H}(\mathbf{p}_i)$ 加入 $A$ .

至此, 经过上述四个步骤, 得到时刻 $t$ 的具备三维坐标的场景彩色图像 $I_t$ .

如图2所示, 通过以上方法, 可在单帧图像中实时、一一对应地标注三维空间离散点编码, 并抑制文字重叠造成的视觉混淆. 该方法将区域坐标嵌入至彩色图像当中, 保持图像场景完整的同时又将图像所示区域与坐标相对应, 可为后续导航区域评估任务同时提供图像与位置信息表征, 实现基于视觉模型的开放场景实例语义实时在线获取与定位.

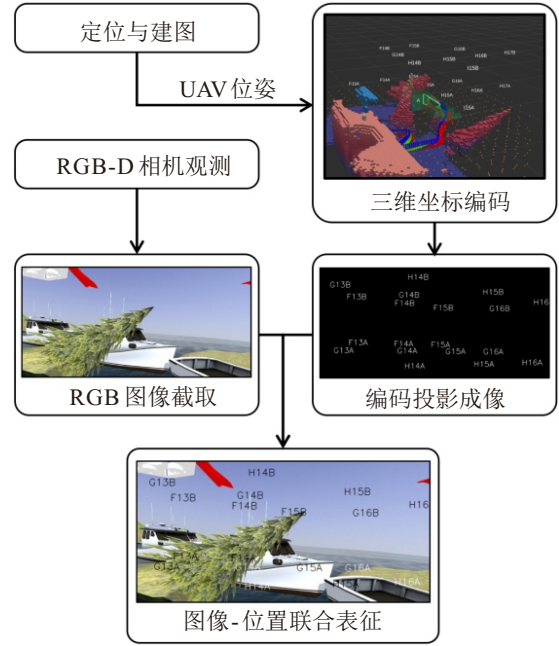


图2 Gazebo 仿真场景中的图像编码流程示例

### 2.1.3 视觉语义提取与定位

本文将视觉语言大模型形式化地定义为一个映射函数:

$$f_\theta : C \times L \rightarrow L. \quad (11)$$

其中 $C$ 表示彩色图像集合,  $L$ 为自然语言输出域. 据此, 在时刻 $t$ , 系统可将编码图像 $I_t$ 中的视觉语义信息, 通过VLM在特定提示词 $P^*$ 的引导下进行推理, 并转化为自然语言描述:

$$l_t \leftarrow f_\theta(I_t, P^*), l_t \in L. \quad (12)$$

此输出 $l_t$ 在结构上被定义为一个“<坐标编码, 开放语义>”对:

$$l_t = \langle \mathcal{H}(\mathbf{p}_k), d(\mathbf{p}_k) \rangle. \quad (13)$$

其中 $d(\mathbf{p}_k) \in L$ 为位置 $\mathbf{p}_k$ 的开放语义描述. 具体而言,  $l_t$ 的示例如下:

**例1** Code:G5A, Description: 黄色共享单车旁地面区域.

**例1** 对应于截取自室外真实场景实验图3. 实验以开放目标“骑行头盔”作为搜索任务验证对象.

彩色图像由无人机机载前向相机获取,故存在运动模糊.其中“G5A”等文字编码即为按照前述视觉逆映射方法构造的三维坐标编码,其同时反映场景的三维位置信息与图像的二维位置信息.根据实验参数

$$\Delta = 3m, [x_{\min} \ y_{\min} \ z_{\min}]^T = [-5.0m \ -20.0m \ 0.0m]^T, \text{可以计算出“G5A”}$$

所对应场景三维坐标为:

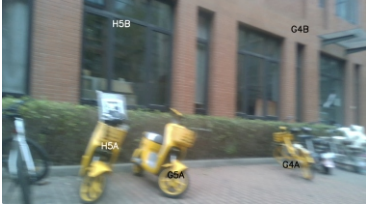
$$\mathbf{p}_t = \mathcal{H}^{-1}(G5A) = [13.0m \ -8.0m \ 0.0m]^T. \quad (14)$$


图3 室外场景无人机前置相机编码图像  $I_t$

编码“G5A”在图像中的二维位置可以在图3中体现,其与VLM所描述的“黄色共享单车旁地面区域”在视觉上对齐.这种从二维图像坐标到场景三维坐标一致对应关系,赋予了VLM理解场景语义并同时对其进行定位的能力.

在此框架下,“黄色共享单车旁地面区域”是VLM基于任务提示词  $P^*$  对  $I_t$  进行信息压缩后得到的开放语义描述,“G5A”为该语义描述在当前场景中的三维坐标  $\mathbf{p}_t$ ,两者共同组成  $t$  时刻的“<坐标,开放语义>”对  $l_t$ .在得到  $l_t$  后,系统即时地对坐标  $\mathbf{p}_t$  区域赋予高价值语义评分,并将  $l_t$  加入开放词汇语义地图.

### 2.1.4 区域开放词汇语义表示

为了显式利用场景感知的历史结果并获得细粒度语义表示,本文构建开放词汇语义地图.该地图一方面记录感知结果,另一方面为大语言模型(LLM)推理提供结构化的场景表达.

不同于物体类别固定的语义占据图,本方法的语义元素来自海量自然语言样本预训练的模型参数空间,其语义粒度可根据任务目标、环境复杂度及提示词灵活调整,具备可扩展性与任务自适应性.

在此表示框架中,  $M = \{(\mathbf{p}_k, d(\mathbf{p}_k))\}$  记录了每个空间位置及其语义描述.由于语义描述  $d(\mathbf{p}_k)$  来源于VLM,其内容可涵盖丰富信息,提供对于目标可能性区域的细致刻画.例如,在“骑行头盔”搜索任务中,系统可在不同视点获得如图4中的描述.

为了平衡语义表示的精度与搜索规划的计算代价,将相邻的若干EROI归并,成为更大尺度的区域单元  $r_n \in \mathcal{R}$ ,归并过程使用映射  $g: P_{\mathcal{E}} \rightarrow \mathcal{R}$  表示;当区域单元  $r_n$  内积累的语义观测数量  $U(r_n)$  达到阈

```

“(13.0, -8.0, 0.0)”:[
  “黄色共享单车旁地面区域”
],
“(10.0, -2.0, 0.0)”:[
  “建筑入口前通道,能向玄关区域”
],
“(4.0, -2.0, 0.0)”:[
  “道路旁停放自行车区域”
],
“(10.0, -8.0, 0.0)”:[
  “道路旁自行车停放区”,
  “道路旁停放区潜在头盔区域”
],

```

图4 开放词汇语义地图示例

值  $\theta$  时,系统触发推理步骤,将该区域单元内的语义描述集合  $\{d(\mathbf{p}_k) \mid g(\mathbf{p}_k) \in r_n\}$  组织成提示词  $P_{r_n}^*$ ,并输入LLM进行高层语义理解与目标相关性评估,从而得到区域单元语义评分  $s(r_n)$ .

算法2展示了  $t$  时刻更新开放词汇语义地图并借助其进行区域单元语义评分的简略流程.

### 算法2 开放词汇语义地图构建与区域评分

**输入:** <坐标,开放语义>对  $l_t = \langle \mathbf{p}_k, d(\mathbf{p}_k) \rangle$ ,更新阈值  $\theta$ .

**输出:** 坐标-语义映射  $M$  及评分  $s(r)$ .

step 1:  $M \leftarrow \emptyset$ ; 对每个  $r_n \in \mathcal{R}$ , 设  $U(r_n) \leftarrow 0$ .

step 2: 在时刻  $t$  执行:

a.  $M \leftarrow M \cup \{\langle \mathbf{p}_k, d(\mathbf{p}_k) \rangle\}$ ;

b.  $r_n \leftarrow g(\mathbf{p}_k)$ ;

c.  $U(r_n) \leftarrow U(r_n) + 1$ .

step 3: 若存在  $r_n$  使得  $U(r_n) \geq \theta$ , 则执行:

a.  $P_{r_n}^* \leftarrow \langle d(\mathbf{p}_k) : g(\mathbf{p}_k) \in r_n \rangle$ ;

b.  $s(r_n) \leftarrow \text{LLM}(P_{r_n}^*)$ ;

c.  $U(r_n) \leftarrow 0$ .

step 4: 输出结果  $(M, s(r))$ .

## 2.2 目标线索评估与运动决策

### 2.2.1 提示词与零样本更新

针对未知环境下的目标搜索任务,本文设计了一种以“场景辨识—场景评估—场景转移”为核心逻辑框架的自然语言提示词,并结合时刻  $t$  的编码图像共同作为VLM的输入,以高效追踪零样本目标相关线索或进行场景转移.同时,提示词支持基于模板替换的零样本目标更新,使系统在出现新目标时可快速生成对应提示词,无需模型微调或额外数据标注,具备开放性与即时适应能力.

#### 1) 提示词逻辑设计

如图5所示,遵循“辨识—评估—转移”推理逻辑设计视觉语言大模型提示词,帮助无人机模拟人

类在未知环境中基于上下文进行目标搜索的推理过程. 该提示词结构不依赖特定环境或目标样本预训练, 可通过语义层面的逻辑规则实现泛化. 其层次化推理逻辑可适配不同尺度和复杂度的任务场景, 从而实现目标的高效搜索与决策转移.

任务说明
你是UAV画面分析助手. 【目标】=【头盔】 以“辨识-评估”的循环范式, 定位[目标]最可能出现可应当转移的关键位置.
问题A
对场景图像进行辨识 当前场景类型是? 与[目标]是否匹配? 是否存在[目标]相关线索? 转移通道(门、窗等)?
问题B
评估: [前往线索]或[场景转移] 存在[目标]相关线索时直接前往. 场景类型与[目标]不匹配时进行转移.
输出约束
输出位置编码与描述 输出图像上距所选择区位置编码并进行简短的描述.
样本示例
示例1(场景摘要): 车库储物架上有明显头盔形状物体. 期望输出: Code: F6A, Description: 车库架上骑行头盔

图5 “辨识—评估—转移”提示词内容示例

## 2) 零样本目标更新

为了支持开放词汇目标搜索, 本文引入了一种零样本目标更新机制. 该机制通过语言模型的上下文理解能力, 将原提示词模板中的目标实体替换为新的语义对象, 无需任何再训练即可生成新的任务提示. 具体过程可形式化表示为:

$$P_{\text{new}}^* = \text{LLM}(P_{\text{old}}^*, T_{\text{new}}). \quad (15)$$

其中:  $P_{\text{old}}^*$  表示原始提示词模板, 包含完整的逻辑框架;  $T_{\text{new}}$  为新的目标描述;  $P_{\text{new}}^*$  为由 LLM 根据语义替换生成的目标自适应提示词.

在生成过程中, 模型执行以下隐式映射:

$$\begin{aligned} \text{Logic}(P_{\text{new}}^*) &= \text{Logic}(P_{\text{old}}^*), \\ \text{Target}(P_{\text{new}}^*) &= T_{\text{new}}. \end{aligned} \quad (16)$$

即保持原有提示词的逻辑结构与决策规则不变, 仅更新与目标相关的语义域. 此过程相当于在语言层面实现了目标概念的语义注入与结构保持.

### 2.2.2 分层语义价值更新

本文使用 VLM 对图像逐帧在线赋分反映视野内 EROI 的短时目标线索语义价值; 基于开放词汇语义地图的 LLM 赋分则由多帧观测结果在区域上累计得到, 反映相邻多个 EROI 的长时语义价值. 二者在时间尺度与空间范围上不同.

为了实现 VLM 与 LLM 不同时间与范围评分的融合, 本文采用主观贝叶斯方法对 EROI 的语义价值进行更新, 其核心作用是将来自不同时间的 VLM 局部评分与 LLM 区域评分融合为概率值  $s(e) \in [0, 1]$ .

### 1) 编码图像的在线评估赋分

在主观贝叶斯方法中,  $C(E|S)$  可用于刻画证据强度, 即在给定语义证据  $S$  时, 对事件  $E$  的支持强度. 本文将大模型对区域的语义评分归一到  $[-5, 5]$  区间, 并在后续更新式中通过尺度系数进行归一化, 因此取  $C(E|S) = 5$  作为强正证据的评分以保持一. 同时, 在提示词设计中约束: 仅当 VLM 判断当前位置存在“目标或强相关线索”时才输出坐标编码与描述; 因此当获得 VLM 输出  $\mathbf{p}_k$  时, 可将其视作对局部邻域  $\mathcal{N}_r(\mathbf{p}_t)$  所覆盖 EROI 集合的一次强正证据.

如 2.1.3 节所述, VLM 从时刻  $t$  的编码图像中推理出单个 <坐标, 开放语义> 对  $l_t = \langle \mathbf{p}_k, d(\mathbf{p}_k) \rangle$ ; 以  $\mathbf{p}_k$  为球心, 半径  $r$  的邻域所覆盖的  $m$  个 EROI 的集合为  $\mathcal{N}_r(\mathbf{p}_t) = \{e_1, \dots, e_m\}$ .

首先依据 EROI 的总数  $N$  初始化每个 EROI 的语义价值  $s(e_j), \forall e_j \in \mathcal{E}$ :

$$s(e_j) = \tilde{p} = 1/N. \quad (17)$$

根据主观贝叶斯更新公式并对  $C(E|S)$  按照数量  $m$  进行平均有:

$$\begin{aligned} \forall e_j \in \mathcal{N}_r(\mathbf{p}_t), \\ s(e_j) \leftarrow s(e_j) + \frac{\frac{1}{5}C(E|S)}{m}(1 - s(e_j)), \end{aligned} \quad (18)$$

代入常数:

$$C(E|S) = 5 \Rightarrow \frac{\frac{1}{5}C(E|S)}{m} = \frac{1}{m}, \quad (19)$$

于是更新规则简化为:

$$s(e_j) \leftarrow (1 - \frac{1}{m})s(e_j) + \frac{1}{m}, \forall e_j \in \mathcal{N}_r(\mathbf{p}_t). \quad (20)$$

即所有被模型选中的  $m$  个 EROI 的语义价值均匀地向 1 逼近一次加权平均.

### 2) 基于语义地图的评估赋分

由算法 2 得到  $k$  个 EROI 区域  $r_n = \{e_1, \dots, e_k\}$  的语义价值  $s(r_n)$  后; 据主观贝叶斯更新规则, 当该语义价值  $s(r_n)$  落入区间  $[-5, 0]$  时, 对该  $r_n$  内的所有 EROI 按如下方法进行概率增益更新:

$$s(e_j) \leftarrow \tilde{p} + [s(e_j) - \tilde{p}] \times \frac{1}{5}s(r_n) + 1] \times \frac{1}{k}. \quad (21)$$

当  $s(r_n) \in (0, 5]$  时,  $\forall e_j \in r_n$ :

$$s(e_j) \leftarrow s(e_j) + [1 - s(e_j)] \times \frac{s(r_n)}{5k}. \quad (22)$$

同时, 维持任意 EROI 语义价值不低于初始值:

$$s(e_j) \geq \tilde{p}, \quad \forall e_j \in r_n. \quad (23)$$

通过主观贝叶斯方法更新 EROI 中目标存在的先验概率, 系统在提升模型所推测高价值区域搜索优先级的同时, 仍保留对其他区域的探索, 从而在目标导向搜索与未知空间探索之间取得折中.

### 2.2.3 动态路径代价评估

在未知环境自主探索中, 机器人需要选择能够以最小路径代价换取最大信息增益的最优目标点. 文献 [17] 中将信息增益与路径代价固定组合构成效用函数以计算最优目标点, 本文在此基础上引入节点语义价值  $s(x_i)$ , 用于反映节点  $x_i$  所在 EROI 的语义价值. 根据节点得分动态决定是否考虑路径代价项, 使搜索过程在“空间自主探索”与“语义规律利用”之间自适应切换.

#### 1) 基于语义价值的动态路径代价调节

系统基于一个增量生成的路网图  $R = (V, E)$  进行搜索, 其中节点集  $V = \{x_i\}$  由自由空间采样, 边集  $E = \{(x_i, x_j)\}$  为节点间的无碰撞路径.

该路网图以后续探索的前沿区域为中心进行增量式构建, 并始终维持着稀疏与连通的拓扑特性.

对于机器人当前位置  $x_r$  与候选节点  $x_i$ , 定义路径  $\gamma = f(x_r, x_i)$ . 其对应的效用函数定义为:

$$U(\gamma) = \begin{cases} I(\gamma) s(x_i), & s(x_i) > \tau \\ I(\gamma) s(x_i) e^{-\lambda L(\gamma)}, & s(x_i) \leq \tau \end{cases}. \quad (24)$$

其中,  $I(\gamma)$ : 路径终点的几何价值;  $s(x_i)$ : 节点语义价值;  $\tau$ : 动态阈值,  $\tau = k \max_{x_j \in V} s(x_j)$ ,  $k$  为超参数;  $L(\gamma)$ : 从  $x_r$  到  $x_i$  的路径代价;  $\lambda$ : 路径代价系数.

该策略为启发式策略, 目的是提升高置信线索的响应速度. 当节点语义价值接近全局最优 ( $s(x_i) > \tau$ ) 时, 代表该节点区域高价值, 此时算法将暂时解除对路径代价的优化约束, 采取一种“贪心”策略, 以加速探索; 当节点语义价值处于中等水平时, 引入以路径长度为变量的指数惩罚项  $e^{-\lambda L}$ , 使决策在关注语义价值的同时, 兼顾路径效率.

如图 6 所示, 效用值  $U$  通过颜色深浅反映, 实心圆点表示语义价值小于动态阈值的候选节点, 距离圆心越远则路径代价越高、效用值越低; 而实心方块表示语义价值高于动态阈值的候选节点, 其效用值不受距离影响, 从而增强“语义规律利用”效果.

#### 2) 基于效用的图搜索过程

### 算法3 基于语义价值的效用驱动图搜索

**输入:** 路网图  $R = (V, E)$ , 起点  $x_r$ , 参数  $\lambda, k$ ; 几何价值  $I(\cdot)$ , 路径代价  $L(\cdot)$ , 语义价值  $s(\cdot)$ .

**输出:** 最优目标  $x^*$  与路径  $\gamma^* = f(x_r, x^*)$ .

step 1: 初始化变量:

- 计算  $s_{\max} \leftarrow \max_{x_i \in V} s(x_i)$ ;
- 设置阈值  $\tau \leftarrow k s_{\max}$ ;
- 初始化  $U^* \leftarrow 0, x^* \leftarrow x_r$ .

step 2: 调用 Dijkstra 算法, 从  $x_r$  出发在  $R$  上依次扩展节点  $(x_i, L_i)$ :

- 计算效用  $U_b(x_i) \leftarrow I(\gamma) s(x_i)$ ;
- 若  $s(x_i) \leq \tau$ , 则考虑路径代价:  $U(x_i) \leftarrow U_b(x_i) e^{-\lambda L(\gamma)}$ ; 否则,  $U(x_i) \leftarrow U_b(x_i)$ ;
- 若  $U(x_i) > U^*$ , 则更新最优值:  $U^* \leftarrow U(x_i), x^* \leftarrow x_i$ .

step 3: 输出最优目标节点  $x^*$  及其对应路径  $\gamma^* = f(x_r, x^*)$ .

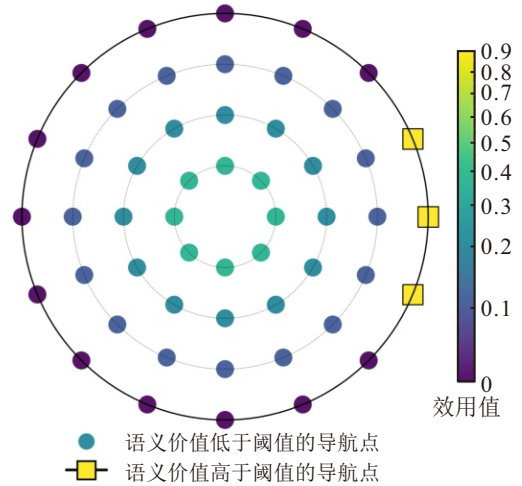


图6 基于语义价值的动态路径代价调节平面示意图

### 2.3 运动规划与执行

在确定最优目标点及对应路径  $\gamma^* = f(x_r, x^*)$  后, 系统进入运动规划与执行阶段. 由于基于路网图的路径通常由离散节点组成, 因此需在执行前进行轨迹平滑与动态优化<sup>[19]</sup>.

首先, 系统采用最小曲率约束的路径平滑算法, 将离散路径转化为连续可导的飞行轨迹, 确保在障碍环境下的可行性与安全性. 该过程可有效减少姿态变化, 提高无人机飞行的稳定性与能量效率.

随后, 基于时间最优分配方法对平滑轨迹进行动态标定, 根据无人机的最大速度、加速度约束确定每段轨迹的时间参数, 从而在满足动力学约束的前

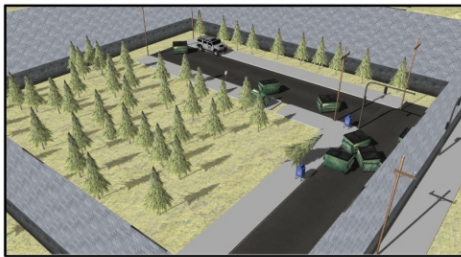
提下, 尽可能提升飞行效率。

最终, 由底层飞行控制器执行该轨迹。当环境信息或目标更新时, 系统进行重规划。

### 3 仿真对比与分析

#### 3.1 仿真设置

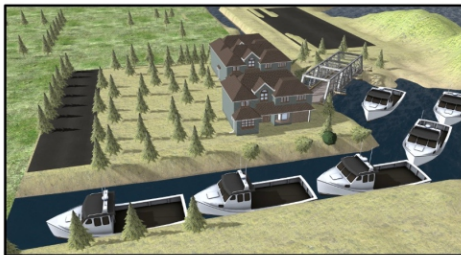
如图7所示, 本文在基于物理引擎的 Gazebo 11 平台上构建了三个典型开放场景, 用于评估所提出方法在不同地形结构、语义类型及空间尺度下的目标搜索性能。平台运行于 Ubuntu 20.04 + ROS Noetic 环境中, 无人机配置双目彩色-深度相机与同时定位与建图模块。任务开始时, 无人机不具备任何环境先验地图, 需要在未知空间中同时完成几何探索与语义线索捕获, 以尽可能短的时间定位目标并完成近距离观察。所有场景均设置有遮挡结构, 使任务起点与目标区域之间不存在直接视野。



(a) 道路车辆, 45 m\*48 m\*2 m



(b) 建筑车辆, 40 m\*40 m\*2 m



(c) 河流桥梁, 65 m\*47 m\*2 m

图7 Gazebo 仿真场景

#### 场景 (a): 搜索道路旁车辆 (物体性目标)

该场景用于评估系统在半结构化道路环境中, 能否从局部视觉观测中提取与“车辆”相关的多模态语义线索, 并据此判断道路拐角、停靠点、遮挡物周边等与“车辆”语义关联的区域, 从而实现对物体性目标的优先访问与快速确认。场景中树木、垃圾桶与倒伏设施形成多重遮挡, 可检验系统在弱结构场景中根据线索推理主动规避低价值区域的能力。

#### 场景 (b): 搜索建筑外车辆 (跨结构遮挡)

该场景用于评估系统在存在跨室内外的复杂环境中, 能否遵循“辨识—评估—转移”的推理循环, 推断并前往车辆可能出现的区域, 并在探索过程中赋予其更高访问优先级。

#### 场景 (c): 搜索跨河桥梁 (背景性目标)

该场景用于评估系统在大尺度自然环境中, 将河道边缘、道路延伸方向及航道交汇处识别为“跨河桥梁”高度相关线索区域, 从而优先访问这些区域以定位背景性目标。相较于前两类物体场景, 该场景检验系统利用开放词汇线索处理无明确边界、尺度较大的背景性结构目标的能力。

#### 3.2 模型与参数设置

视觉语言大模型 (VLM) 在实验中使用 qwen3-vl-plus 作为视觉-语言推理模块, 用于从机载图像中解析开放词汇语义并输出短结构化结果 (<坐标编码, 语义>), 进而更新开放词汇语义地图。VLM 采用联网请求、异步并发的方式: 真机实验中调用频率为 10 Hz, 仿真中为 5 Hz。

大语言模型 (LLM) 使用 qwen-plus 作为区域级语义推理模块, 用于对局部区域语义观测进行归纳总结与目标相关性评估。LLM 采用事件触发: 当同一区域开放词汇语义更新次数超过阈值  $N = 3$  时触发一次该区域的 LLM 查询。

为保证轨迹可行性与控制安全性, 真机与仿真均采用一致的运动学参数上限约束: 最大偏航角速度:  $2.0 \text{ rad/s}$ ; 最大偏航角加速度:  $2.0 \text{ rad/s}^2$ ; 最大线速度:  $1.0 \text{ m/s}$ ; 最大线加速度:  $2.0 \text{ m/s}^2$ ; 最大加速度:  $20.0 \text{ m/s}^3$ 。

#### 3.3 仿真对比

本文选取先进的无人机自主探索方法 FSMP<sup>[19]</sup> 以及目标搜索方法 Star-Searcher<sup>[9]</sup> 进行仿真对比; FSMP 方法有机地融合了基于边界和基于采样的策略, 实现对未知环境的快速全局探索; Star-Searcher 方法基于可见性视点簇设计了层次化规划策略, 一方面高效遍历全局视点簇, 另一方面在局部执行精细路径规划, 实现对潜在目标区域的完整视觉检查。

如表1所示, 在三类仿真场景中, 所提方法在路径长度、搜索时间、成功率与路径长度加权成功率 (Success weighted by Path Length, SPL) 方面均取得最优结果, 并在所有场景保持 100% 的成功率, 证明了所提方法在未知环境下目标搜索效率和场景泛化性能方面的优越性。

首先, 在“道路车辆”场景中, 虽然三种方法均

表4 三类仿真场景下各方法的性能对比

场景	方法	路径长度 (m)	搜索时间 (s)	成功率 (%)	SPL (%)
道路车辆(45 × 48 × 2m <sup>3</sup> )	所提方法	<b>96.86±13.46</b>	<b>113.85±11.12</b>	<b>100</b>	<b>65.4</b>
	FSMP <sup>[9]</sup>	183.95±71.54	202.48±73.51	<b>100</b>	37.3
	Star-Searcher <sup>[9]</sup>	283.04±71.83	450.58±171.30	60	15.9
建筑车辆(40 × 40 × 2m <sup>3</sup> )	所提方法	<b>38.93±3.92</b>	<b>49.27±3.72</b>	<b>100</b>	<b>77.4</b>
	FSMP <sup>[9]</sup>	156.43±84.61	173.98±88.15	<b>100</b>	29.8
	Star-Searcher <sup>[9]</sup>	207.76±181.66	238.99±261.66	80	32.4
河流桥梁(65 × 47 × 2m <sup>3</sup> )	所提方法	<b>158.14±26.99</b>	<b>183.50±30.56</b>	<b>100</b>	<b>38.4</b>
	FSMP <sup>[9]</sup>	208.95±79.86	233.50±83.99	<b>100</b>	31.7
	Star-Searcher <sup>[9]</sup>	212.41±56.63	467.97±173.03	60	20.0

能最终找到目标,但FSMP作为典型的纯未知环境自主探索方法,其规划目标是最大化空间覆盖,因此会在大量语义价值低的空旷区域进行无差别扩展,导致路径和时间显著增加。Star-Searcher则聚焦于对“可见物体表面”进行遍历式检查,其视点簇机制会对树木、杆件等所有物体表面进行同等细致扫描,但缺乏对“车辆可能出现区域”的语义关注重点,因而在无关物体上产生大量冗余飞行。相比之下,所提方法通过开放词汇语义线索将道路拐角、停车带等高价区域优先化,使平均路径长度减少约50%–60%,路径长度加权成功率提升至65.4%。

在“建筑车辆”场景中,FSMP仍维持100%成功率,但由于无法根据语义判断室内外区域的重要性,常在室内空间产生大量无效探索;Star-Searcher则因对多类建筑立面和遮挡物进行“表面遍历”而表现不稳定,成功率降至80%。所提方法可依据“室外–道路–停放区”等语义线索完成更具方向性的场景转移,使路径与时间显著缩短,并保持最高的路径长度加权成功率(77.4%)。

在“河流桥梁”场景中,FSMP与Star-Searcher都需要对大范围河岸进行覆盖或物体表面遍历,难以从语义层面推断“桥梁应位于道路跨河位置”这一结构规律,因此产生长距离翻找与高时间开销。所提方法能够基于开放词汇语义线索(如“河道边缘”“道路延伸方向”)快速收敛到潜在跨河通道区域,使平均路径与时间分别降低约20%以上,并取得显著更高的路径长度加权成功率(38.4%)。

总体而言,FSMP属于无语义目标的自主探索方法,更倾向于对未知空间进行无差别的探索扩展;Star-Searcher则偏向对所有物体表面进行“遍历式”检查,缺乏目标物体针对性。相比之下,本文利用多模态大模型提供的高价值语义线索,实现了“探索与利用”的协同规划策略,从而在三个场景中均取得最优性能。

### 3.4 消融分析

为定量评估系统各核心模块的贡献,在保持环境、传感器、底层探索与规划设置一致的前提下,设计了五组对照:(i)完整系统;(ii)权重衰减赋分对照:以局部视野投影并按视角衰减赋分替代所提方法;(iii)移除动态路径代价调节:移除2.2.3节中基于语义价值的动态路径代价调节机制;(iv)结构化提示词简化:用最简提示替换结构化提示词,仅要求视觉语言模型输出关键位置编码;(v)移除开放词汇地图评估:不使用LLM对所构建的开放词汇地图进行推理评估,仅依赖VLM即时输出。

由于“空间视觉逆映射”模块是其余所提模块的基础,难以独立移除,因此参考文献[12]采用“视野权重衰减赋分对照”间接反映其贡献。具体而言,将彩色图像直接输入VLM并由模型给出目标相关性得分,将该得分投影到当前视野的frontiers中,并按视角进行置信度衰减:光轴附近栅格权重较高,随角度偏移按 $\cos^2(\frac{\theta}{\theta_{\text{FOV}}/2} \cdot \frac{\pi}{2})$ 衰减,其中 $\theta$ 为偏移角度, $\theta_{\text{FOV}}$ 为视场角大小。

如表2所示,三类场景下各消融设置均保持100%成功率,说明各方法均具备完成目标搜索的基本能力,性能差异主要体现在路径长度、搜索时间与SPL。总体来看,“权重衰减赋分对照”在建筑车辆与河流桥梁场景中退化明显,表明显式三维坐标编码有助于缓解二维投影带来的语义定位偏差,增强语义线索与空间区域的对应关系。在道路车辆场景中,移除“开放词汇地图评估”造成最显著退化,说明历史语义观测累积与区域级推理能够提升搜索稳定性。

结构化提示词在道路车辆与建筑车辆场景中带来稳定增益,但其效果在河流桥梁场景中并不单调,表明提示设计仍受场景结构与线索分布影响。动态路径代价调节同样具有场景依赖性:其在道路车辆场景中提升搜索效率,而在建筑车辆和河流桥梁场

表5 三类仿真场景下不同模块消融设置的性能对比

场景	方法	路径长度 (m)	搜索时间 (s)	成功率 (%)	SPL (%)
道路车辆 (45 × 48 × 2m <sup>3</sup> )	(i) 完整系统(所提)	<b>96.86±13.46</b>	<b>113.85±11.12</b>	100	<b>65.4</b>
	(ii) 权重衰减赋分对照	115.29±13.10	132.93±12.97	100	54.7
	(iii) 移除动态路径代价调节	106.29±23.01	121.96±24.76	100	61.0
	(iv) 结构化提示词简化	117.28±32.65	137.85±31.53	100	56.0
	(v) 移除开放词汇地图评估	139.47±48.63	161.19±50.31	100	50.4
建筑车辆 (40 × 40 × 2m <sup>3</sup> )	(i) 完整系统(所提)	38.93±3.92	<b>49.27±3.72</b>	100	77.4
	(ii) 权重衰减赋分对照	60.03±14.48	73.19±16.63	100	51.9
	(iii) 移除动态路径代价调节	37.82±11.05	50.91±11.51	100	<b>83.1</b>
	(iv) 结构化提示词简化	56.73±38.61	71.42±40.66	100	66.6
	(v) 移除开放词汇地图评估	<b>37.70±2.47</b>	53.15±3.63	100	79.6
河流桥梁 (65 × 47 × 2m <sup>3</sup> )	(i) 完整系统(所提)	158.14±26.99	183.50±30.56	100	38.4
	(ii) 权重衰减赋分对照	265.70±177.20	295.79±186.68	100	31.5
	(iii) 移除动态路径代价调节	<b>120.49±13.71</b>	<b>137.31±14.88</b>	100	<b>49.9</b>
	(iv) 结构化提示词简化	142.70±28.97	164.63±30.90	100	43.0
	(v) 移除开放词汇地图评估	174.74±46.97	200.69±52.83	100	36.3

景中可能因语义误差造成绕行。

### 3.5 模型推理时延分析

表 3 给出了实验中的推理统计.VLM 与 LLM 的平均时延分别为 1.51s 与 0.35s. 本文在系统实现上将语义评估 (VLM/LLM) 与运动规划设计为异步低耦合: 规划线程无需等待语义线程返回即可输出导航目标点, 语义结果仅以概率权重形式影响候选区域价值. 因此, 大模型推理延迟不会阻塞运动规划与控制; 当语义更新明显滞后或短时缺失时, 系统将退化为自主探索继续进行空间覆盖式目标搜索。

表6 大模型推理统计与平均时延

模型	调用策略	平均输入	平均输出	平均时延 (s)
		Tokens	Tokens	
qwen3-vl-plus (VLM)	真机 1 0Hz; 仿真 5 Hz	1083.4	18.0	1.51
qwen-plus (LLM)	事件触发; N=3	381.4	1.1	0.35

## 4 实验验证

为了进一步验证所提出方法在真实环境中的可行性与鲁棒性, 本文在一套自主研发的无人机平台上进行了室外真机演示实验. 实验无人机搭载了 Mid360 激光雷达、IMU 以及 Intel RealSense D435i 深度相机, 能够同时获得点云、RGB-D 观测及姿态信息. 机载计算单元为 Intel CORE i7-1260P 处理器、32GB 内存, 并具备稳定的互联网连接, 用于实时访问多模态大模型进行语义推理. 实验搜索目标分别为: “骑行头盔” 与 “雨伞”, 分别对应物体性目标与可由背景结构衍生出的场景语义目标. 如图 8、图 9 所示, 实验场景为典型户外区域, 包含树木、建筑立面、停放车辆、共享单车等多种非结构化元素, 且目

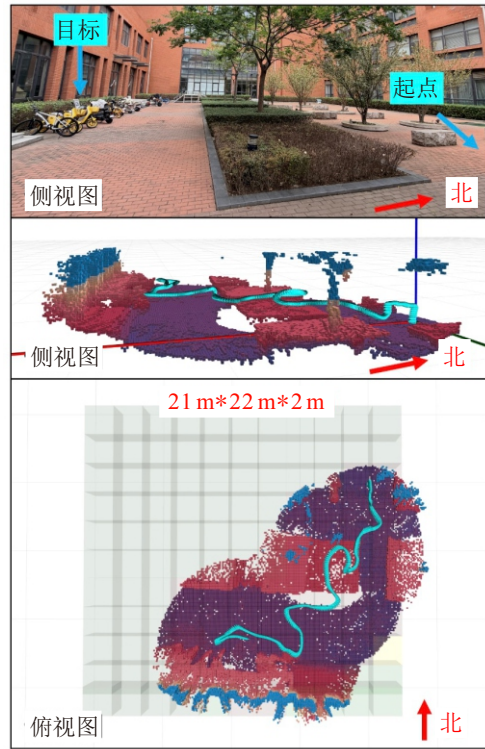


图8 搜索“骑行头盔”实验场景与结果

标均处于一定遮挡之下, 无人机在起飞位置无法直接观测到目标. 在搜索“骑行头盔”任务中, 无人机能够根据模型推理出的线索, 如“靠近单车、地面支架或骑行相关区域”, 优先访问道路边缘与单车聚集位置, 并最终成功抵近头盔目标. 在搜索“雨伞”任务中, 系统从模型生成的背景性线索 (如“靠近房屋入口、临时遮阳区域或路边结构”) 中推断出高价值区域, 实现对非标准化物体的有效检索. 实验结果表明, 所提出方法在真实环境中能够稳定运行, 并可在完全未知场景下依托多模态语义线索实现对目标的

快速搜索. 相比仅依赖几何信息的探索策略, 该方法能够显著减少无效飞行范围, 并在物体类别、外观变化及弱结构场景条件下保持较强的泛化性与实用性.

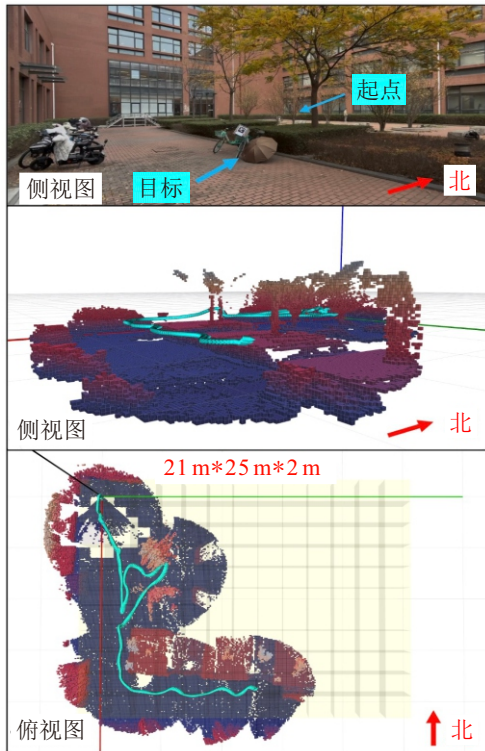


图9 搜索“雨伞”实验场景与结果

## 5 结论

本文提出一种融合多模态大模型推理与自主探索规划的无人机零样本目标搜索方法. 通过空间-视觉逆映射、开放词汇语义地图构建以及几何-语义价值融合与动态评估机制, 实现了对目标语义线索的在线定位与高效利用, 显著提升了未知环境下的搜索效率. 仿真与真机实验均验证了方法的有效性与泛化能力. 所提方法在任务与场景上的适用性与局限性总结如下:

1. 任务适用性: 所提方法适用于未知环境下的零样本单目标搜索, 尤其当目标分布与环境语义线索存在相关性时, 可通过开放词汇语义地图与区域级语义评估提高搜索效率. 对于背景性或弱边界目标 (如与道路、河道走向等相关的大尺度结构), 开放词汇线索能够提供比闭集物体类别更灵活的语义识别与推理.

2. 多目标与动态目标局限: 当前实现以区域内“目标存在概率”的累积评估为核心, 更适合静态或慢变化环境. 对于多目标搜索, 需要进一步引入目标后验维护、去重确认与任务分配策略; 对于强动态场景或移动目标, 历史语义证据可能造成决策滞后, 需引入时间衰减更新或动态环境表示以提升鲁棒性.

3. 场景适用性: 方法对室外开放、半结构化场景具有较强适应性; 在高度拥挤、强遮挡环境中, 由于视野受限, 单帧图像中可嵌入的空间坐标锚点数量可能减少, 从而降低 VLM 可用观测密度并使搜索效率下降. 室内狭窄空间亦可运行, 但通常需要更细的区域划分与更保守的飞行动力学约束.

4. 系统条件边界: 方法依赖可靠的定位与建图, 并依赖云端大模型提供语义评估. 在定位漂移或弱网条件下, 语义价值更新将退化, 此时系统将更多依赖几何探索模块完成覆盖式搜索.

未来工作将围绕模型轻量化、多目标搜索以及动态场景下的语义预测展开, 以进一步增强系统在复杂环境中的适应性与实用性.

## 参考文献 (References)

- [1] Cao C, Zhu H, Choset H, et al. TARE: A hierarchical framework for efficiently exploring complex 3D environments[C]. Robotics: Science and Systems XVII. Cambridge, 2021: 18.
- [2] Bircher A, Kamel M, Alexis K, et al. Receding horizon “next-best-view” planner for 3D exploration[C]. IEEE International Conference on Robotics and Automation. Stockholm, 2016: 1462-1468.
- [3] Respass V M, Devitt D, Fedorenko R, et al. Fast sampling-based next-best-view exploration algorithm for a MAV[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 89-95.
- [4] Zhang X T, Chu Y B, Liu Y S, et al. A novel informative autonomous exploration strategy with uniform sampling for quadrotors[J]. IEEE Transactions on Industrial Electronics, 2022, 69(12): 13131-13140.
- [5] Yamauchi B. A frontier-based approach for autonomous exploration[C]. Proceedings IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. Monterey, 2002: 146-151.
- [6] Zhang H, Wang S Y, Liu Y S, et al. EFP: Efficient frontier-based autonomous UAV exploration strategy for unknown environments[J]. IEEE Robotics and Automation Letters, 2024, 9(3): 2941-2948.
- [7] Zhou B Y, Zhang Y C, Chen X Y, et al. FUEL: Fast UAV exploration using incremental frontier structure and hierarchical planning[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 779-786.
- [8] Papatheodorou S, Funk N, Tzoumanikas D, et al. Finding things in the unknown: Semantic object-centric exploration with an MAV[C]. IEEE International Conference on Robotics and Automation. London, 2023: 3339-3345.
- [9] Luo Y M, Zhuang Z X, Pan N, et al. Star-searcher: A complete and efficient aerial system for autonomous target search in complex unknown environments[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4329-4336.

- [10] 陈铂垒, 康嘉绪, 钟萍, 等. 面向具身人工智能的物体目标导航综述[J]. 软件学报, 2025, 36(4): 1715-1757. (Chen B L, Kang J X, Zhong P, et al. Survey on object goal navigation for embodied AI[J]. Journal of Software, 2025, 36(4): 1715-1757.)
- [11] Sun J W, Wu J, Ji Z, et al. A survey of object goal navigation[J]. *IEEE Transactions on Automation Science and Engineering*, 2025, 22: 2292-2308.
- [12] Zhang M J, Du Y H, Wu C K, et al. ApexNAV: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion[J]. *IEEE Robotics and Automation Letters*, 2025, 10(11): 11530-11537.
- [13] Hong Y C, Wang Z, Wu Q, et al. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 15418-15428.
- [14] Xie Q T, Zhang T Y, Xu K D, et al. Reasoning about the unseen for efficient outdoor object navigation[J/OL]. 2025, arXiv: 2309.10103.
- [15] Lu J W, Wu Z Y, Xu X W, et al. SG-nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation[C]. Advances in Neural Information Processing Systems 37. Vancouver, 2024, DOI: 10.52202/079017-0171.
- [16] Dorbala V S, Mullen J F, Manocha D. Can an embodied agent find your “cat-shaped mug”? LLM-based zero-shot object navigation[J]. *IEEE Robotics and Automation Letters*, 2024, 9(5): 4083-4090.
- [17] Yokoyama N, Ha S, Batra D, et al. VLFM: Vision-language frontier maps for zero-shot semantic navigation[C]. IEEE International Conference on Robotics and Automation. Yokohama, 2024: 42-48.
- [18] Dong Q L, Xi H B, Zhang S Y, et al. Fast and communication-efficient multi-UAV exploration via voronoi partition on dynamic topological graph[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi, 2024: 14063-14070.
- [19] Zhang S Y, Zhang X B, Dong Q L, et al. FSMP: A frontier-sampling-mixed planner for fast autonomous exploration of complex and large 3-D environments[J]. *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 7504414.

## 作者简介

张雪波 (1984-), 男, 教授, 博士, 主要研究方向为机器人与人工智能, E-mail: [zhangxuebo@nankai.edu.cn](mailto:zhangxuebo@nankai.edu.cn);

马哲 (2001-), 男, 硕士生, 主要研究方向为无人机视觉语言导航, E-mail: [zhema@mail.nankai.edu.cn](mailto:zhema@mail.nankai.edu.cn);

张世勇 (1993-), 男, 助理研究员, 博士, 主要研究方向为多旋翼无人机环境感知与运动规划, E-mail: [zhangshiyong@nankai.edu.cn](mailto:zhangshiyong@nankai.edu.cn);

王子玉 (2003-), 男, 硕士生, 主要研究方向为无人机运动规划与目标跟踪, E-mail: [wziyu@mail.nankai.edu.cn](mailto:wziyu@mail.nankai.edu.cn);

奚浩博 (2000-), 男, 硕士生, 主要研究方向为无人机激光 SLAM, E-mail: [xihaobo@mail.nankai.edu.cn](mailto:xihaobo@mail.nankai.edu.cn);

张智勇 (2003-), 男, 硕士生, 主要研究方向为无人机建图, E-mail: [zhangzhiyong@mail.nankai.edu.cn](mailto:zhangzhiyong@mail.nankai.edu.cn);

袁明星 (1990-), 男, 副教授, 博士, 主要研究方向为智能机器人规划与控制, E-mail: [mxyuan@nankai.edu.cn](mailto:mxyuan@nankai.edu.cn).

## 科研团队简介

张雪波教授科研团队立足于南开大学人工智能学院, 长期专注于机器人领域, 致力于将前沿科技与民生产业需求紧密结合. 团队的研究重点涵盖地面与空中移动机器人、人机交互及智能操作等多个方向. 凭借一系列创新性成果, 团队已将其技术成功应用于灾难搜索救援、高原与极地科考、智能电力作业、医疗健康以及智能制造等关键领域, 取得了显著的社会与经济效益. 在科研攻坚的同时, 团队也高度重视学术梯队建设, 在培育杰出人才、营造创新科研氛围、拓展国际合作视野等方面积累了丰富经验.

课题组负责人张雪波教授, 南开大学教授、博导, 现任人工智能学院副院长、机器人与信息自动化研究所所长、天津市智能机器人技术重点实验室副主任. 入选全球 2% 顶尖科学家榜单, 并同时担任 IEEE/ASME Trans. on Mechatronics 以及 ASME Journal of Dynamic Systems, Measurement and Control 等国际期刊编委. 研究团队承担了国家重点研发计划课题、国家自然科学基金等 20 多项课题, 近五年主持经费 3000 多万. 在 IEEE 汇刊上发表了 50 余篇论文. 团队注重理论结合应用, 推动高海拔科考机器人、配网带电作业机器人、救援机器人等特殊服役环境下机器人技术的发展. 成果获天津市科技进步一等奖、天津市自然科学一等奖与二等奖、吴文俊人工智能自然科学一等奖、天津市教学成果一等奖、中国自动化学会教学成果一等奖.