**中图法分类号**:TP183; TP391.4　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-29

**论文引用格式**:Zheng Zhouyi, Guo Chenrui, Shan Chun, Zhang Lei, Wei Wei. A Survey of Unmanned Aerial Vehicle Vision-Language Navigation from the Perspective of Embodied Intelligence[J/OL]. Journal of Image and Graphics, XXXX:1-29. DOI: 10.11834/jig.260203. (郑周一, 郭宸瑞, 单淳, 张磊, 魏巍. 具身智能视角下无人机视觉语言技术研究进展[J/OL]. 中国图象图形学报, XXXX:1-29. DOI: 10.11834/jig.260203.)
[DOI:10.11834/jig.260203]

# 具身智能视角下无人机视觉语言技术研究进展

郑周一[1,2],郭宸瑞[3,4],单淳[1,2],张磊[1,2],魏巍[1,2]

1. 西北工业大学计算机学院,西安　710129; 2. 西北工业大学空天地海一体化大数据应用技术国家工程实验室,西安　710129; 3. 宇航智能控制技术全国重点实验室,北京　100854; 4. 北京航天自动控制研究所,北京　100854

**摘　要**: 无人机视觉语言导航(aerial vision-language navigation, AVLN)是融合计算机视觉、自然语言处理与无人机控制的空中具身智能前沿方向,旨在使无人机依据自然语言指令在非结构化三维环境中实现自主导航。针对传统导航难以理解高层语义、无法适应复杂空间约束的局限,本文系统梳理无人机视觉语言导航的研究脉络与技术体系。首先归纳通用仿真、真实场景重建、虚拟场景建模三类仿真平台的特点,对比主流数据集在场景复杂度、指令语义与动作表征上的差异;其次从感知表征、推理范式、记忆存储、具身控制四个核心模块,剖析跨模态对齐、大模型推理、长程记忆与连续控制的关键技术;最后总结该技术在城市巡检、灾害救援、智能物流与精准农业中的落地应用,指出仿真到现实迁移、多机协同、长程鲁棒性等挑战。本文全面呈现领域研究现状,为空中具身智能的进一步发展提供参考。

**关键词**: 无人机;视觉语言导航;具身智能;跨模态对齐;大模型

# A Survey of Unmanned Aerial Vehicle Vision-Language Navigation from the Perspective of Embodied Intelligence

Zheng Zhouyi[1,2], Guo Chenrui[3,4], Shan Chun[1,2], Zhang Lei[1,2], Wei Wei[1,2]

*1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China; 2. National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Northwestern Polytechnical University, Xi'an 710129, China; 3. State Key Laboratory of Aerospace Intelligent Control Technology, Beijing 100854, China; 4. Beijing Aerospace Automatic Control Institute, Beijing 100854, China*

**Abstract**: Unmanned aerial vehicle vision-language navigation stands as a core research direction of aerial embodied intelligence, which combines computer vision, natural language processing, robot control and unmanned system technology. It endows unmanned aerial vehicles with the ability to understand human natural language instructions and perform autonomous navigation, path planning and task execution in unstructured, GPS-denied and dynamically changing three-dimensional environments. Different from traditional unmanned aerial vehicle navigation methods that rely on satellite positioning, inertial navigation or manual waypoint setting, unmanned aerial vehicle vision-language navigation establishes a direct mapping from high-level semantic instructions to low-level continuous flight actions, so that unmanned aerial vehicles can complete complex tasks such as target search, obstacle avoidance, long-distance cruising and scene reasoning only through visual observation and language understanding. This technology effectively breaks the limitations of traditional

**收稿日期**:2026-04-14;**修回日期**:2026-06-14

**基金项目**:国家自然基金(62372379,62472359)

**Supported by**: the National Natural Science Foundation of China (62372379 and 62472359)

navigation in semantic interaction and environmental adaptability, and has important theoretical value and engineering application prospects in low-altitude economy, urban security, emergency rescue, industrial inspection and precision agriculture. In recent years, with the rapid development of multimodal large models, transformer architectures and embodied artificial intelligence, unmanned aerial vehicle vision-language navigation has gradually evolved from early modular pipelines to end-to-end learning frameworks, and then to highly interpretable reasoning systems driven by large language models and vision-language models. However, compared with ground robot vision-language navigation, unmanned aerial vehicle vision-language navigation still faces many unique challenges caused by aerial movement characteristics. First, unmanned aerial vehicles move with six degrees of freedom, and frequent changes in height, pitch, yaw and roll lead to unstable visual input, large differences in object scales and serious geometric distortion, which increases the difficulty of cross-modal alignment between vision and language. Second, three-dimensional spatial topology is more complex, and spatial prepositions such as "above", "between", "along" and "across" require stronger geometric reasoning and spatial awareness, which cannot be satisfied by two-dimensional image matching alone. Third, long-range navigation tasks bring problems such as massive visual information, cumulative positioning errors and semantic forgetting, which put forward higher requirements for efficient memory mechanism and environmental representation. Fourth, the discrete semantic decision space is difficult to match with the continuous physical control space of unmanned aerial vehicles, resulting in a large gap between simulation training and real-world deployment. Aiming at the above problems, this paper systematically summarizes the research progress of unmanned aerial vehicle vision-language navigation from the perspective of embodied intelligence. First, three types of simulation platforms are reviewed: general robot simulation platforms, simulation platforms based on real scene reconstruction, and large-scale virtual simulation platforms built by game engines. These platforms provide safe, low-cost and high-efficiency test environments for algorithm verification, data generation and model training, and effectively narrow the domain gap between simulation and reality. Second, mainstream datasets are compared and analyzed from the dimensions of scene scale, instruction complexity, action space definition and environmental authenticity. The development of datasets reflects the trend of unmanned aerial vehicle vision-language navigation moving from indoor structured scenes to outdoor large-scale urban scenes, from short simple instructions to long sequential reasoning instructions, and from discrete actions to continuous six-degree-of-freedom control. Third, the core technical framework is divided into four modules: perception representation, reasoning paradigm, memory storage and embodied control. The perception representation focuses on visual feature extraction, geometric alignment, multimodal fusion and world model construction. The reasoning paradigm mainly includes cross-modal attention mechanism, transformer-based pretraining and large-model-driven chain-of-thought reasoning. The memory storage evolves from implicit temporal memory to explicit semantic maps and topological graphs. The embodied control develops from modular trajectory planning to end-to-end vision-language-action generation with safety constraints. In terms of practical applications, unmanned aerial vehicle vision-language navigation has been widely explored in urban infrastructure inspection, disaster emergency search and rescue, intelligent logistics distribution and precision agricultural monitoring. It can reduce manual participation, improve operation efficiency and enhance safety in high-risk, high-complexity and high-efficiency-demanding scenarios. However, there are still key bottlenecks restricting large-scale deployment, such as simulation-to-real transfer, generalization in unseen environments, real-time performance of onboard computing, safety and stability in dynamic environments, and collaborative navigation of multiple unmanned aerial vehicles. Finally, this paper prospects the future development trends of unmanned aerial vehicle vision-language navigation, including stronger world model and predictive reasoning, better generalization and robustness based on multimodal large models, safer and more reliable physical control, more efficient human-machine interaction and collaborative intelligence, as well as deeper integration with low-altitude digital economy and unmanned system ecology. This review aims to provide a complete and clear technical route for researchers in the fields of embodied intelligence, computer vision, natural language processing and unmanned aerial vehicle systems, promote the breakthrough of key technologies, and accelerate the practical and industrialization process of unmanned aerial vehicle vision-language navigation.

**Key words**: unmanned aerial vehicle; vision-language navigation; embodied intelligence; cross-modal alignment; large model

**论文引用格式：**Zheng Zhouyi, Guo Chenrui, Shan Chun, Zhang Lei, Wei Wei, Wang Wei. A Survey of Unmanned Aerial Vehicle Vision-Language Navigation from the Perspective of Embodied Intelligence[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260203.（郑周一，郭宸瑞，单淳，张磊，魏巍. 具身智能视角下无人机视觉语言技术研究进展. 中国图象图形学报. DOI: 10.11834/jig260203.）

## 0 引 言

人工智能的演进正经历着从“数字孪生”到“具身智能”的历史性飞跃。随着大规模预训练模型在自然语言处理与计算机视觉领域取得突破性进展，研究重心已开始从虚拟世界的被动信息处理，转向物理实体的自主感知、推理与交互。具身智能作为连接数字算法与物理现实的桥梁，已被公认为通往通用人工智能的必经之路(Savva 等，2019)。在这一技术浪潮中，无人机凭借其超越地面机器人的高机动性与三维空间覆盖能力，成为具身智能研究中最具挑战性与应用潜力的载体之一。特别是随着全球“低空经济”概念的提出与战略化部署，无人机在城市智慧物流、精准农业监测、复杂灾难救援以及电力基础设施巡检等领域的应用呈现出指数级增长态势。在这些实际任务中，无人机不仅需要具备基础的避障与路径规划能力，更需要能够理解人类复杂且多样的自然语言指令，并在非结构化、通信受限且动态变化的环境中自主执行导航任务。

传统的无人机导航技术主要依赖于全球导航卫星系统、惯性测量单元以及基于同步定位与地图构建的空间表征技术(Cadena 等，2016)。然而，这些方法在面对“飞往那栋冒烟建筑的二层窗户并搜索幸存者”这类具有高度语义抽象性的任务时，往往显得无能为力。这是因为传统导航范式缺乏对环境语义的深度理解以及视觉信息与语言描述之间的跨模态对齐机制。因此，无人机视觉语言导航应运而生，它旨在构建一个能够将高层语义指令直接映射为底层飞行控制序列的智能认知系统。无人机视觉语言导航的研究不仅涉及计算机视觉中的目标检测与语义分割，还涵盖了自然语言处理中的指令拆解与上下文理解，更触及了机器人学中的动力学约束与连续控制，是一个典型的多模态交叉前沿领域。

从相关概念的层级关系来看，具身智能是无人机视觉语言导航所属的总体研究范式，其核心在于智能体通过感知、认知、决策和行动与物理或仿真环境形成闭环交互(Savva 等，2019)。视觉语言导航则是具身智能在导航任务中的典型体现，强调智能体依据自然语言指令和视觉观测，在环境中完成目标定位、路径选择与动作执行(Anderson 等，2018)。在此基础上，世界模型进一步关注智能体对环境状态、动态演化和动作后果的内部建模，使其能够在部分可观测、长时序和不确定环境中进行前瞻性推理与规划(Ha 和 Schmidhuber，2018)。近年来兴起的视觉-语言-动作模型(vision-language-action，VLA)则试图将视觉观测、语言目标与动作生成统一到同一模型框架中，使智能体能够从多模态输入直接生成可执行动作，从而推动具身智能由“感知—规划—控制”的模块化范式向端到端闭环决策范式发展(Brohan 等，2023)。因此，具身智能、视觉语言导航、世界模型与视觉语言动作模型并非彼此割裂的概念，而是形成了从研究范式、任务定义、环境建模到动作生成的递进关系。对于无人机视觉语言导航而言，具身智能提供总体理论框架，视觉语言导航定义语言条件下的空间导航任务，世界模型支撑三维环境记忆与未来状态预测，视觉-语言-动作模型则为多模态感知到连续飞行动作的统一映射提供新的技术路径。

相较于地面机器人视觉语言导航，无人机视觉语言导航并非仅是将智能体从地面平台迁移到空中平台，而是在任务属性、感知方式、空间推理和动作执行机制上均发生了显著变化。首先，在运动形态上，地面智能体通常受限于二维平面或近似二维拓扑结构，其动作空间多表现为前进、转向、停止等离散或低维控制；而无人机具备升降、横移、偏航、俯仰等高自由度运动能力，导航过程呈现明显的三维空间特征。由此带来的问题并不是简单的类别级语义对齐失效，而是目标尺度、观测角度、相对方位和高度层次会随着飞行姿态与视点变化在时序上快速改变，进而使语言指令中涉及空间关系、运动方向和目标位置的表达更难被稳定定位到视觉观测与空间状态中。其次，在空间推理层面，地面视觉语言导航更多依赖房间、道路、走廊等相对稳定的拓扑结构，而无人机视觉语言导航需要同时理解水平位置、高度

关系、遮挡结构和可飞行空域。例如,对于“飞到建筑物上方”“穿过两栋楼之间”“沿河道向前搜索”等指令,无人机不仅需要识别目标类别,还需要建立目标之间的三维相对位置、可达空间和高度约束关系,这对几何常识推理和三维场景建模提出了更高要求(Hong 等,2020;Krantz 等,2020)。再次,在长程任务层面,无人机通常面向城市街区、灾害现场、农田或园区等大范围开放环境,飞行距离长、观测信息冗余且地标稀疏,容易出现局部视觉证据丢失、累积定位误差和语言子目标遗忘等问题。因此,如何构建面向长时序任务的语义地图、拓扑记忆或世界模型,是区别于短程地面导航的重要挑战。最后,在动作执行层面,现有视觉语言导航方法多基于离散动作建模,而真实无人机必须在连续控制空间中满足动力学约束、安全距离、飞行稳定性和能耗限制。高层语言指令、离散语义决策与底层连续飞行控制之间存在明显鸿沟,这也是无人机视觉语言导航从仿真环境迁移到真实低空场景时面临的关键瓶颈(Loquercio 等,2021)。因此,无人机视觉语言导航的核心问题可以概括为:如何在动态三维观测条件下,将自然语言指令稳定映射为具有空间一致性、长程记忆能力和物理可执行性的飞行动作序列。

无人机视觉语言导航经历了从模块化流水线到端到端黑盒学习,再到如今认知驱动范式的转变。早期研究者倾向于采用分而治之的策略,通过视觉识别与语言解析的简单级联来实现导航,但这种方式极易产生误差扩散。随着深度强化学习与模仿学习的兴起,研究者提出了跨模态注意力机制,尝试构建从感知到动作的直接映射模型(Wang 等,2019;Fried 等,2018)。这一阶段的工作虽然在特定数据集中表现出色,但在面对未见环境时泛化性较差。随后,Transformer 架构的引入引发了跨模态预训练的革命。模型如 VLN-BERT(Majumdar 等,2020)、Airbert(Guhur 等,2021)等通过在海量图像-文本对上进行掩码预测与匹配训练,赋予了智能体更强的场景表征能力,显著提升了导航的成功率。近年来,大语言模型与视觉语言模型的爆发为具身智能注入了新的活力。如 LM-Nav(Shad 等,2023)等工作的出现,标志着无人机开始具备利用大语言模型(large language model, LLM)作为“大脑”进行任务逻辑拆解、风险预测甚至自我纠错的能力,实现了从单纯的“模式匹配”向“逻辑推理”的迈进(Ahn 等,2022;Driess 等,2023)。

尽管国内外学术界已涌现出若干关于具身智能或地面导航的综述文章(Gu 等,2022;Lei 等,2024;Zhang 等,2025),但在无人机这一细分领域,仍缺乏系统性且具有前瞻性的深度总结。现有的部分工作往往局限于算法性能的横向对比,未能深刻揭示三维空间物理约束与跨模态语义理解之间的内在耦合规律。针对这一现状,本文旨在通过解构具身认知闭环,从感知表征、推理范式、记忆存储与具身控制四个核心维度对无人机视觉语言导航方法进行全面综述。本文通过深入挖掘上述四个环节在实际应用中的关键技术瓶颈,试图为构建更具鲁棒性与通用性的空中具身智能体提供理论支撑。

全文的组织结构安排如下。首先,本文对无人机视觉语言导航的仿真进行综述。随后,本文对主流的无人机视觉语言导航数据集进行汇总评析,从指令的语言多样性、环境的几何复杂度以及动作空间的定义方式等维度解析现有数据集的优缺点。在此基础上,本文进行无人机视觉语言导航核心技术分析,详细阐述感知表征中的跨模态对齐技术、推理范式中的大模型驱动机制、记忆存储中的语义图构建方法以及具身控制中的连续空间映射策略。接着,本文探讨了无人机视觉语言导航在复杂灾难救援、城市精细化巡检与智能物流等典型实际场景中的落地挑战。最后,本文针对从仿真迁移到现实的跨域鸿沟、多无人机协同视觉语言导航的通信约束以及长距离复杂任务下的逻辑鲁棒性等前沿课题进行展望,旨在为未来的研究工作提供指导与启发,推动无人系统在自动化与智能化领域的长远发展。

## 1 仿真平台

无人机视觉语言导航算法的研发、验证与迭代高度依赖可控、可复现、低成本的实验环境。真实无人机飞行试验受硬件成本高昂、空域监管严格、飞行安全风险高、环境不可控等因素制约,难以支撑大规模、高频次、长时序的算法训练与测试。在此背景下,高保真仿真平台成为无人机视觉语言导航研究的核心基础设施,能够为多模态感知、跨模态对齐、语义推理、连续控制等关键技术提供闭环验证环境,是连接算法设计与工程落地的重要桥梁。

无人机视觉语言导航本质是多模态具身智能任

务，需要同时完成视觉观测、语言指令理解、空间定位与动作执行，对训练数据的规模、多样性、标注精度均提出极高要求。现实场景中，大规模多模态标注数据难以通过人工采集获得，不仅成本高、周期长、一致性差，还易受光照、气象、地形、遮挡等不可控因素影响，导致数据分布单一、泛化能力不足。高保真仿真环境可通过程序化生成无限场景，实现传感器数据、语义标签、空间位姿、深度信息、障碍物分布等数据的全自动精准标注，显著降低数据获取成本，缓解具身智能领域数据稀缺与标注困难的问题。同时，仿真平台可灵活构建极端光照、复杂气象、动态干扰、通信受限等非结构化场景，全面测试算法在极端条件下的鲁棒性，为模型优化提供可靠依据。

依据场景来源、构建方式与仿真目标差异，现有无人机视觉语言导航仿真平台可划分为三大类：通用仿真平台、基于真实场景重建的仿真平台、基于虚拟场景建模的仿真平台。三类平台分别面向基础算法验证、真实域适配、大规模泛化训练等不同研究需求，共同构成从基础控制、跨模态对齐到高层语义推理的全链路仿真支撑体系，为无人机视觉语言导航的技术演进提供重要支撑。

### 1.1 通用仿真平台

通用仿真平台是面向机器人领域的模块化、标准化仿真环境，具备跨平台、可扩展、多物理引擎兼容等特点，可为无人机提供基础动力学仿真、传感器模拟、环境交互与控制接口。其核心设计思路是解耦机器人硬件形态与导航算法逻辑，通过标准化传感器模型、执行器模型与物理引擎，使研究者能够在统一框架下快速验证算法，支持与ROS（robot operating system）、YARP（yet another robot platform）、Pocolibs无缝对接，便于仿真策略向真实无人机系统迁移。

早期经典通用平台为机器人仿真奠定了重要基础。Gazebo（Koenig 等，2004）作为开源社区广泛使用的仿真工具，集成 Open Dynamics Engine 等物理引擎，可稳定实现户外三维环境物理交互模拟，支持多机协同与传感器仿真。Webots（Michel 等，2004）具备快速原型开发能力，内置丰富的商用无人机与移动机器人模型，支持用户使用高级编程语言直接编写可部署至真实硬件的控制代码。USARSim（unified system for automation and robot simulation）（Carpin 等，2007）依托虚幻引擎实现高真实感渲染，开创了游戏引擎用于机器人仿真的先例，成为国际机器人竞赛常用平台。MORSE（modular open robots simulation engine）（Echeverria 等，2011）基于 Blender 建模与 Python 脚本实现组件化设计，支持不同抽象级别仿真，可直接提供场景语义标签，简化视觉处理流程。CoppeliaSim（Rohmer 等，2013）具备分布式控制、多物理引擎切换与高度可扩展性，适用于多精度、多任务联合仿真。

随着具身智能对视觉真实感与物理精度需求不断提升，面向无人机与自动驾驶的专用高保真仿真平台快速发展。AirSim（Shah 等，2018）针对无人机与无人车设计，提供精细的空气动力学模型与接近摄影级的视觉渲染，支持多旋翼、固定翼等多种机型仿真。CARLA（Dosovitskiy 等，2017）基于虚幻引擎构建大规模开源城市场景，提供语义分割、深度图、法向量等伪传感器数据，支持多样化气象与光照变化，适用于城市低空导航与地标的视觉语义学习。NVIDIA Isaac Sim 依托 Omniverse 平台，利用极限光线追踪（ray tracing texel eXtreme，RTX）与 PhysX 物理引擎，实现硬件加速仿真，支持大规模并行训练与合成数据生成。此外，Google Earth（Gorelick 等，2017）、GTAV（grand theft auto V）（Ji 等，2025）等地理信息系统与游戏场景也被用于超大规模视觉数据获取，弥补传统模拟器场景规模有限的问题。面向大模型与具身交互的新型框架如 PromptCraft（Vemprala 等，2023）进一步提供标准化提示工程接口，显著降低自然语言驱动的无人机导航研究门槛。

不同通用仿真平台在物理一致性、视觉保真度、计算开销、易用性之间存在明显权衡。Gazebo、Webots 在动力学建模与底层控制反馈上可靠性高、资源占用低，但视觉真实感较弱；基于虚幻引擎的 AirSim、CARLA 视觉效果接近真实拍摄，跨模态对齐效果更优，但配置复杂、计算开销大，不利于大规模并行训练。MORSE 灵活性强，但处理高帧率图像流时受限于 Python 执行效率。FastSim（Abderehman 等，2022）等轻量化模拟器牺牲部分视觉真实感，换取极高的仿真速度，适合快速策略迭代。PyBullet（Panerati 等，2021）封装 Bullet 物理引擎，计算高效、物理精度充足，成为深度强化学习在无人机导航中应用的主流选择。总体而言，通用仿真平台为无人机视觉语言导航提供了基础实验环境，支撑从简单动作映射到复杂推理的全流程研发。

### 1.2 基于真实场景重建的仿真平台

基于真实场景重建的仿真平台以现实物理世界为蓝本，通过摄影测量、三维点云重构、卫星遥感影像、城市地理数据等手段构建高保真数字孪生环境，能够精准还原真实场景的几何结构、空间拓扑关系、地表纹理细节与语义分布特征，摆脱传统人工建模的主观性与局限性。相较于通用仿真平台，该类平台可天然呈现真实世界中存在的视觉噪声、地标稀疏性、复杂遮挡关系、非规则建筑形态与动态环境变化，从而有效缩小仿真与现实之间的域间隙，显著提升无人机视觉语言导航模型在真实场景中的泛化能力与部署鲁棒性，是推动算法从实验室走向工程应用的关键支撑，从主流基于真实场景重建的仿真平台获取的图像数据如图1所示。

在无人机视觉语言导航领域，面向真实环境的仿真平台近年来快速发展，逐步实现从小规模室内场景向大规模城市场景、从静态环境向动态环境、从单视角观测向多模态感知的升级。AVDN（aerial vision-and-dialog navigation）（Fan 等，2023）率先基于真实卫星图像构建俯视视角航空导航环境，依托xView 数据集（Lam 等，2018）实现连续空间下的真实视觉观测，同时支持人机对话交互与人工注意力标注，为对话式无人机导航任务提供了重要的仿真基础与数据支撑。EmbodiedCity（Gao 等，2024）面向开放城市场景，基于虚幻引擎高保真复刻真实城市商业区三维结构，融合真实路网、建筑模型、动态行人与车流模拟，提供第一人称多传感器感知接口与连续动作控制能力，将具身智能导航从室内环境成功拓展至大规模真实城市场景。CityNav（Lee 等，2024）依托真实城市点云数据重建剑桥、伯明翰等真实城区场景，构建了目前规模最大的真实场景无人机视觉语言导航数据集，覆盖大范围长程导航任务，为地理空间感知与语言引导的航空导航提供了高难度、高真实度的测试基准。OpenFly（Gao 等，2025）进一步整合多渲染引擎与三维重建技术，实现从真实场景到仿真环境的自动化构建，并支持大规模导航轨迹与语言指令自动生成，重点解决真实复杂场景下视觉表征与自然语言指令的对齐难题，为通用化无人机视觉语言导航模型提供了可扩展、高效率的实验平台。

整体而言，基于真实场景重建的仿真平台将无人机导航环境从“人工设计”推向“真实复刻”，迫使智能体在非理想观测条件、模糊语言描述与复杂空间结构中学习鲁棒的导航策略，标志着无人机视觉语言导航向实用化迈出关键一步。随着神经辐射场（neural radiance fields，NeRF）、三维高斯泼溅（3D Gaussian Splatting）等实时高精度三维重建技术不断成熟，未来真实场景重建平台将在更低计算开销下实现更高的视觉保真度、几何精度与交互实时性，进一步推动空中具身智能从仿真环境快速落地到复杂真实场景中。


图1 从基于真实场景重建的仿真平台采集的图像数据

Fig. 1 Image data collected from simulation platform based on real-world scene reconstruction

### 1.3 基于虚拟场景建模的仿真平台

基于虚拟场景建模的仿真平台依托商用游戏引擎、物理仿真插件与程序化生成技术，构建大规模、高自由度、高视觉保真度的虚拟实验环境。与真实场景重建平台不同，该类平台并非直接采集现实世界数据，而是依据地形规则、建筑布局、物体分布、光照气象等设计规范生成开放式数字场景，具有环境可控、扩展性强、标注全自动、训练成本低等突出优势，能够为无人机视觉语言导航模型提供海量、多样、高覆盖度的训练数据，显著提升模型在未知环境中的泛化性能，是当前大规模具身智能算法研发的主流技术路线。从主流基于虚拟场景建模的仿真平台采集的图像数据如图2所示。

虚拟场景建模平台能够灵活突破现实物理约束，快速生成城市、郊区、山地、园区、建筑群等多样化场景，并可自由调节光照、天气、遮挡、动态障碍物等条件，特别适合对无人机视觉语言导航算法进行极限条件测试与鲁棒性验证。同时，虚拟平台可自动输出深度图、语义分割掩码、三维位姿、目标包围

框、拓扑关系等高质量标注信息，完全免除人工标注成本，有效解决多模态具身智能数据稀缺、标注困难、一致性差等核心问题。

在无人机视觉语言导航领域，面向空中智能体的虚拟仿真平台已形成较为完善的技术体系。AerialVLN（vision-and-language navigation for UAVs）（Liu 等，2023）基于虚幻引擎 4 与 AirSim 构建了包含 25 种典型城市大类场景的大规模户外环境，覆盖市中心、工业区、公园、村庄等，平均路径长度达 661.8 米，单条指令平均包含 9.7 个参考目标，是 R2R（room-to-room）数据集（Anderson 等，2018）的 2.6 倍，任务复杂度显著高于传统地面导航数据集，支持前向、转向、升降、平移 4 自由度动作空间，重点针对无人机三维空间导航与长程推理能力进行评估。MetaUrban（Wu 等，2024）提出组合式程序化场景生成方案，通过街道块布局、功能区划分、障碍物分布、动态实体填充等环节，实现无限不重复城市场景生成，大幅提升智能体在开放、非结构化环境中的空间适应能力与语义理解泛化性。GRUtopia（Wang 等，2024）基于 NVIDIA Isaac Sim 搭建城市级大规模具身交互场景，集成海量建筑、植被、道路、人流、车流等资产，为无人机提供高复杂度语义对齐、空间拓扑推理与长时序记忆决策的仿真环境。

为进一步提升虚拟环境真实性与动态交互能力，UnrealZoo（Zhong 等，2024）在仿真场景中引入多样化人类与动物模型，增强动态目标感知与复杂交互任务的仿真真实性，使无人机能够在搜救、监测等任务中学习动态语义理解能力。OpenUAV（unmanned aerial vehicle）（Wang 等，2024）实现对无人机六自由度（six degrees of freedom，6-DoF）。运动的完整建模，支持连续姿态控制与高精度轨迹仿真，更贴近真实飞行动力学特性，为视觉语言导航到连续控制指令的端到端学习提供支撑。

总体来看，基于虚拟场景建模的仿真平台为无人机视觉语言导航提供了高可用、高效率、高扩展性的“数字练兵场”。从 AerialVLN 定义无人机专用导航仿真环境，到 MetaUrban 与 GRUtopia 实现大规模城市生态构建，再到 UnrealZoo 与 OpenUAV 完善动态要素与飞行物理特性，虚拟仿真技术正推动无人机导航从简单几何感知向高级常识推理加速演进。未来，随着生成式人工智能与实时渲染技术深度融合，虚拟场景将具备更强物理真实性、动态交互性与环境多样性，为低空智能体在复杂现实世界中执行长程、安全、高鲁棒性导航任务提供更加坚实的支撑。


图 2　从基于虚拟场景建模的仿真平台采集的图像数据

Fig. 2　Image data collected from simulation platform based on virtual scene modeling

### 1.4　仿真平台横向比较与发展脉络

从平台功能与研究适配性的角度看，通用仿真平台、真实场景重建平台和虚拟场景建模平台分别对应无人机视觉语言导航发展的不同阶段与需求侧重点，其横向比较如表 1 所示。

通用仿真平台主要服务于无人机动力学建模、传感器仿真和底层控制验证，其优势在于接口成熟、物理引擎稳定、易于与真实飞控系统或机器人操作系统连接，适合开展避障控制、轨迹跟踪、强化学习策略训练等基础研究。然而，该类平台通常缺乏面向自然语言任务的语义标注、复杂地标关系和高层交互机制，难以完整支撑视觉—语言—动作闭环任务。

基于真实场景重建的仿真平台则更强调场景几何结构、地理空间关系和视觉纹理的真实性，能够在一定程度上缓解仿真环境与真实低空环境之间的域间隙。AVDN、EmbodiedCity、CityNav、OpenFly 等平台通过卫星影像、真实城市点云、三维重建或多渲染引擎构建接近真实世界的导航环境，使无人机能够在地标稀疏、遮挡复杂、尺度变化明显的场景中学习语言指令与空间结构之间的对应关系。因此，该类平台更适合评估长距离导航、真实场景泛化、城市级空间推理和仿真到现实迁移能力。但其不足在于场景构建成本较高，动态事件、极端天气和可交互物体的可控性相对有限，难以像程序化虚拟环境一样大规模生成多样化任务。

基于虚拟场景建模的仿真平台则突出可扩展性、可控性和自动标注能力。AerialVLN、MetaUrban、GRUtopia、UnrealZoo、OpenUAV 等平台能够通过游戏引擎、程序化生成和物理仿真插件快速构建大规模、多类型、可重复的训练场景，并自动提供深度图、语义分割、目标位置、轨迹状态等监督信息，适合进行大规模模型训练、长程任务生成和极端环境鲁棒性测试。其局限在于，虚拟场景虽然具有较高视觉保真度，但其物体分布、语言描述和动态交互模式仍可能与真实低空环境存在差异，模型在真实城市、灾害现场或农业环境中的泛化能力仍需通过真实数据或实飞实验进一步验证。

总体而言，无人机视觉语言导航仿真平台的发展呈现出从“控制验证工具”向“具身智能闭环环境”演进的趋势。早期平台更关注无人机能否稳定飞行、避障和完成航迹跟踪；随后，真实场景重建平台开始强调真实几何结构、视觉纹理和地理空间语义；近年来，面向无人机视觉语言导航的专用平台进一步引入自然语言指令、目标搜索、连续动作控制和大模型交互，使仿真平台逐渐具备评估“语言理解—视觉感知—空间推理—飞行控制”完整闭环的能力。未来平台建设需要进一步融合三类平台优势：既保持通用平台的物理一致性和控制接口，又提升真实场景平台的视觉与地理真实性，同时利用虚拟场景平台的可扩展性和自动标注能力，从而支撑更接近真实低空应用的无人机视觉语言导航研究。

**表1　无人机视觉语言导航仿真平台横向比较**

**Table 1　Horizontal Comparison of UAV Vision-Language Navigation Simulation Platforms**

| 平台类型 | 代表平台 | 主要优势 | 主要局限 | 适用问题 |
|---|---|---|---|---|
| 通用仿真平台 | Gazebo、Webots、AirSim、CARLA、Isaac Sim | 物理建模成熟，接口标准化，便于控制算法验证 | 语言任务、语义交互和复杂地标关系建模不足 | 飞控验证、避障、轨迹跟踪、强化学习训练 |
| 真实场景重建平台 | AVDN、EmbodiedCity、CityNav、OpenFly | 场景几何和视觉纹理更接近真实世界，利于泛化评估 | 构建成本高，动态交互和任务可控性有限 | 城市级导航、真实域迁移、长程空间推理 |
| 虚拟场景建模平台 | AerialVLN、MetaUrban、GRUtopia、UnrealZoo、OpenUAV | 场景可扩展、标注自动化、任务生成效率高 | 虚拟分布与真实低空环境仍存在差异 | 大规模训练、极端场景测试、语言任务生成 |

## 2　数据集

数据集是驱动无人机视觉语言导航算法创新、模型训练与性能评估的核心基础，其场景规模、环境复杂度、指令语义丰富度、动作空间定义方式直接决定了研究的深度与泛化能力。随着具身智能从二维平面导航向三维空间具身交互演进，无人机视觉语言导航数据集也逐步从小规模、结构化、短指令场景，向大规模、非结构化、长时序、真实域方向快速发展。与地面机器人视觉语言导航数据集相比，无人机数据集更加强调三维空间感知、高度层次变化、大范围拓扑推理与连续飞行控制，更贴合空中智能体的实际运动特性与任务需求。

当前，主流无人机视觉语言导航数据集已在仿真环境与真实场景、室内小范围与室外大规模、简单指令与复杂逻辑推理等维度形成多样化布局，为跨模态对齐、空间推理、长程记忆、连续控制等关键技术提供了标准化评测基准。本章从任务场景的感知复杂度、指令逻辑的语义维度、具身动作的表征方式三个核心层面，系统梳理与对比主流无人机视觉语言导航数据集，分析各类数据集的构建思路、场景特点、语言范式与适用任务，为研究者选取合适评测基准、设计更贴合实际需求的数据集提供全面参考。根据表2所示，现有无人机视觉语言导航数据集在场景类型、语言表示、动作表征和环境来源上呈现出明显分化。AVDN、WebUAV-3M 等数据集更强调真实环境下的视觉观测与对话式或指令式导航，适合评估模型在真实遥感或航拍视角下的跨模态理解能力；AerialVLN、CityNav、OpenFly 等数据集主要面向大尺度室外仿真场景，具有较强的任务可控性和轨迹生成能力，适合研究长程路径跟随、城市地标识别和空间关系推理；UAV-ON、UAV-VLPA、UAV-VLN 等数据集进一步强化开放目标搜索、路径规划和连续动作控制，更接近真实无人机任务中的“语言理解—目标定位—飞行执行”闭环；SpatialSky-Bench、UrbanVideo-Bench 等问答类数据集则更侧重空间理解、视频推理和三维场景认知能力评估。总体来看，

表2中的数据集已从早期单一指令跟随逐渐扩展到目标搜索、空间问答、开放任务推理和连续控制等多种任务形态，但仍普遍存在真实飞行数据不足、多源传感器覆盖有限、动态环境和多机协同任务缺乏等问题。

### 2.1 任务场景的感知复杂度分析

无人机视觉语言导航数据集的场景复杂度，集中体现为感知尺度、环境动态性、观测视角与传感器模态四个维度的持续升级，并呈现出从室内受限空间向室外广域非结构化场景迁移的清晰趋势。早期无人机导航相关数据集多聚焦于结构化室内环境，感知目标以障碍物规避与局部几何定位为主，而随着 AVDN、CityNav、AerialVLN 等数据集的发布，研究重心已全面转向城市级、野外、大尺度、强干扰的真实化场景。

从感知视角与尺度变化来看，无人机具备独特的高空俯瞰与连续变高能力，导致目标物体在观测序列中存在显著的尺度伸缩、视角畸变与拓扑变形。WebUAV-3M 与 AerialVLN 通过引入大范围高度变化，构建了包含近地面、低空、高空多视点切换的视觉序列，对模型的跨尺度特征匹配与空间定位能力提出更高要求。从环境动态性与非结构化程度来看，AeroVerse 与 UrbanVideo-Bench 在场景中加入动态车流、行人、树木晃动、光照突变等真实干扰因素，要求智能体在高动态背景下保持语义理解与导航决策的稳定性。

在多模态感知融合方面，近期数据集逐步突破单一红绿蓝色彩模式（red-green-blue，RGB）图像限制，向深度、光流、点云、语义分割等多源信息协同方向发展。例如 UAV-Flow 与 AeroDuo 将光流信息与深度估计引入导航任务，用以补偿无人机高速运动带来的运动模糊与姿态抖动，提升了在复杂纹理、弱纹理区域的感知鲁棒性。整体而言，无人机视觉语言导航数据集的场景设计正不断贴近真实低空飞行环境，感知维度从二维平面扩展到三维空间，观测条件从理想受控转向复杂非结构化，推动导航模型从“被动识别”向“主动感知”升级。

**表2 主流无人机视觉语言导航数据集**

**Table 2 Mainstream Unmanned aerial vehicle Vision-Language Navigation Datasets**

| 数据集 | 年份 | 场景类型 | 语言表示 | 动作表征 | 环境来源 |
|---|---|---|---|---|---|
| AVDN（Fan 等，2023） | 2023 | 室外 | 对话 | 连续 | 现实 |
| WebUAV-3M（Zhang 等，2023） | 2023 | 室外 | 指令 | 连续 | 现实 |
| AerialVLN（Liu 等，2023） | 2023 | 室外 | 指令 | 连续 | 仿真 |
| OpenUAV（Wang 等，2024） | 2024 | 室外 | 指令 | 连续 | 仿真 |
| CityNav（Lee 等，2024） | 2024 | 室外 | 指令 | 连续 | 仿真 |
| AeroVerse（Yao 等，2024） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| OpenVLN（Liu 等，2025） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| UAV-VLPA*（Sautenkov 等，2025） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| UAV-VLN（Saxena 等，2025） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| UAV-Flow（Wang 等，2025） | 2025 | 室外 | 指令 | 连续 | 现实 |
| AeroDuo（Wu 等，2025） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| UAV-ON（Xiao等，2025） | 2025 | 室外 | 指令 | 连续 | 仿真 |
| SpatialSky-Bench（Zhang 等，2025） | 2025 | 室外 | 问答 | 连续 | 仿真 |
| UrbanVideo-Bench（Zhao 等，2025） | 2025 | 室外 | 问答 | 连续 | 混合 |
| OpenFly（Gao 等，2025） | 2026 | 室外 | 指令 | 连续 | 仿真 |
| InDoorUAV（Liu 等，2026） | 2025 | 室内 | 指令 | 连续 | 仿真 |
| FreeFly-Thinking（Zhou 等，2026） | 2026 | 室外 | 指令 | 连续 | 仿真 |

### 2.2 指令逻辑的语义维度划分

自然语言指令是连接人类意图与无人机自主行为的核心桥梁,其语义抽象程度、时序结构与推理深度,直接决定了导航任务的智能化水平。当前无人机视觉语言导航数据集的指令设计,已从简单的空间指向描述,逐步演进为具备长程规划、细粒度属性约束与高阶常识推理的复杂语义表达。依据指令的逻辑复杂度与认知要求,可将其划分为三级演进范式。

第一阶段为基础具身指令映射,以简单、直接、单步的空间动作为主。指令多为原子化方位描述,仅需完成语言符号与局部动作的粗略对齐,代表性工作如 UAV-VLN,指令结构清晰、推理链短,主要用于验证基础跨模态匹配与定位能力。

第二阶段为长程时序推理与细粒度属性识别,指令包含多段连续子任务、中间地标与目标属性描述。此类指令要求智能体理解先后顺序、相对方位、颜色、形状、高度等复合信息,并具备跨帧语义记忆能力。AerialVLN 与 CityNav 是该阶段的典型代表,其指令包含多层空间关系与多目标依赖,更贴近真实任务中的复杂人工指令。

第三阶段为高阶逻辑推理与隐式意图理解,指令不再给出显式路径,而是提出任务目标,需要智能体自主完成任务拆解、常识推理与决策规划。以 OpenFly、FreeFly-Thinking 为代表的前沿数据集,引入模糊任务式指令,依赖大语言模型的思维链完成从“模糊意图”到“可执行导航动作”的转化,标志着无人机视觉语言导航进入具身认知新阶段。

整体来看,指令语义的演进过程,本质是从“告诉无人机怎么走”到“告诉无人机做什么”再到“告诉无人机任务目标”的升级,充分反映了领域从模式匹配向常识推理、从被动执行向主动认知的发展趋势。

### 2.3 具身动作的表征方式对比

具身动作的表征方式直接决定无人机导航模型与真实飞行控制的匹配程度,是连接语义推理与物理执行的关键环节。当前无人机视觉语言导航数据集的动作定义,正呈现出从离散拓扑导航向连续空间控制、从简化运动向 6-DoF 全自由度运动快速升级的趋势,更贴合空中具身智能的真实动力学特性。

早期数据集为降低学习难度,多采用离散拓扑动作表征。例如早期 OpenUAV 相关工作将导航空间简化为预设图节点与路径点,智能体仅需在有限候选动作中选择,虽提升了训练效率,但严重丢失无人机升降、俯仰、横滚、平移等三维运动特性,与真实飞行控制存在显著差异。

为贴近实际物理约束,主流数据集逐步转向连续动作表征,直接输出三维坐标、速度、角速度或姿态增量,更符合无人机飞控系统执行逻辑。AeroVerse 构建了面向航天具身智能的统一基准套件,支持前后、左右、升降、转向等完整连续控制接口,可输出高精度位姿与多模态观测数据,为场景感知、空间推理、导航探索、任务规划与运动决策五大任务提供统一连续动作支撑。UAV-ON 面向开放世界目标导航任务,采用完全参数化连续动作空间,包含平移、旋转、升降等物理可执行指令,摒弃传统瞬移式交互,要求智能体在无全球定位系统(global positioning system, GPS)与全局地图条件下完成安全避障与目标搜索,显著提升仿真到现实的迁移潜力。

前沿研究进一步探索多粒度、解耦式动作表征,兼顾高层规划与底层控制。UAV-VLPA 提出视觉-语言-路径-动作一体化框架,将自然语言指令解析为目标航路点,结合旅行商问题实现全局路径优化,并通过 A*算法完成局部避障与轨迹细化,形成“语言理解——全局规划——局部控制”的分级执行机制。SpatialSky-Bench 构建面向无人机空间智能的评测体系,通过目标定位、距离估计、高度感知、空间关系推理等 13 项细粒度任务,强化模型对三维空间的精准理解,为连续控制提供可靠感知支撑。FreeFly-Thinking 采用双头视觉-语言-动作架构,同步输出思维链推理文本与连续三维路点,在提升可解释性的同时实现高精度连续空间控制。

总体而言,动作表征从离散到连续、从简化到全自由度、从耦合到解耦的演进,显著提升了无人机视觉语言导航的物理合理性与落地可行性,为实现安全、平滑、可控的空中具身智能飞行奠定基础。

综上,现有无人机视觉语言导航数据集的发展脉络可以概括为三个方面:其一,场景规模由室内或局部受限空间扩展到城市、园区、灾害现场等室外广域环境,推动模型从局部视觉匹配走向大范围空间推理;其二,语言形式由简单方位指令发展为包含多地标、多属性、多步骤和隐式意图的复杂任务描述,促使模型从跨模态对齐走向高层语义推理;其三,动作表征由离散拓扑动作转向连续三维控制和多粒度

路径—动作联合建模，使数据集更加贴近真实无人机飞行约束。尽管如此，现有数据集仍主要依赖仿真环境，真实低空飞行中的风场扰动、传感器噪声、通信延迟、安全约束和多无人机协同交互尚未被充分覆盖。未来数据集建设需要进一步面向真实应用闭环，强化真实飞行数据、多源感知、动态目标、多机协同和安全控制标注。

## 3 评价指标

在评估指标方面，当前主流的视觉语言导航评价体系主要围绕导航精度与任务完成率展开，常用指标包括导航误差（navigation error，NE）、成功率（success rate，SR）、OSR：最优成功率（oracle success rate，OSR）以及路径长度加权成功率（success rate weighted by path length，SPL）。导航误差用于衡量无人机最终位置与目标位置之间的平均欧式距离；成功率指无人机进入目标点指定阈值范围（如3米）内的任务占比，直接反映导航结果的准确性。最优成功率是更宽松的指标，只要轨迹中任意一点进入目标成功阈值即判定为成功，用于评估目标的潜在可达性。路径长度加权成功率结合路径效率与任务完成情况，通过对比实际路径与最优路径长度对成功率进行加权，鼓励更短、更高效的路径规划。这些指标相互补充，为无人机视觉语言导航任务性能提供了多维度量化依据。

随着世界模型被逐步引入无人机视觉语言导航领域，采用弗雷歇起始距离（Fréchet inception distance，FID）、DreamSim、学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）评估生成画面的语义保真度，用均方误差（mean squared error，MSE）、结构相似性（structural similarity，SSIM）、峰值信噪比（peak signal-to-noise ratio，PSNR）衡量像素级重建精度。针对场景描述、任务规划等需要文本输出的复杂任务，相关研究引入大语言模型，并且开始关注语言生成内容质量，借助GPT-4为不同下游任务定制评价标准，同时引入双语评估研究（bilingual evaluation understudy，BLEU）、语义命题图像标题评价（semantic propositional image caption evaluation，SPICE）等传统文本生成指标，全面衡量模型的整体语言能力。

## 4 方　法

无人机视觉语言导航作为典型的空中具身智能任务，其技术体系以“感知—推理—记忆—控制”的完整认知闭环为核心骨架，贯穿从原始传感器输入到最终飞行执行的全流程，如图3所示。为实现自然语言指令到三维空间自主导航的端到端映射，现有研究普遍将整体框架解构为感知表征、推理范式、记忆存储、具身控制四大相互耦合、逐层递进的核心模块。其中，感知表征负责将无人机动态视觉观测与语言指令进行跨模态特征对齐，构建统一的空间语义表达；推理范式承担高层意图解析、空间逻辑判断与导航决策生成，是连接感知与控制的关键中枢；记忆存储用于对历史观测、空间拓扑与指令信息进行持久化编码，支撑长程导航中的信息保持与误差修正；具身控制则将抽象决策转化为符合无人机动力学约束的连续、安全、可执行动作，最终完成物理世界的闭环交互。本章围绕上述四大模块，系统梳理关键技术路线、典型方法、演进脉络与核心挑战，并且讨论多机协同语言导航的相关内容，全面呈现无人机视觉语言导航的技术全貌。

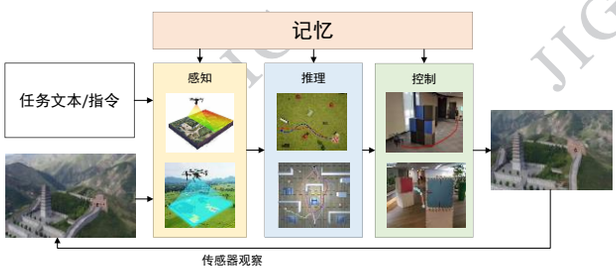


图3　无人机视觉语言导航认知闭环

Fig. 3　Cognitive Closed-loop for UAV Vision-Language Navigation

### 4.1 感知表征

在无人机视觉语言导航的链路中，感知表征作为连接原始物理环境与高层逻辑决策的桥梁，其核心任务在于将机载传感器捕获的高维、非结构化视觉流转化为具备空间逻辑、语义一致性且能与自然语言指令深度对齐的特征表示。由于无人机处于具有六自由度（6-DoF）的开放三维空间，其感知表征不仅要处理航拍视角下剧烈的尺度变化和小目标识别难题，还必须克服二维视觉输入与三维动作规划空

间之间的几何失配。目前的感知表征方法主要呈现出从单一图像特征提取向结构化空间建模及几何对齐增强演进的趋势。无人机视觉语言导航感知表征方法对比如表3所示。

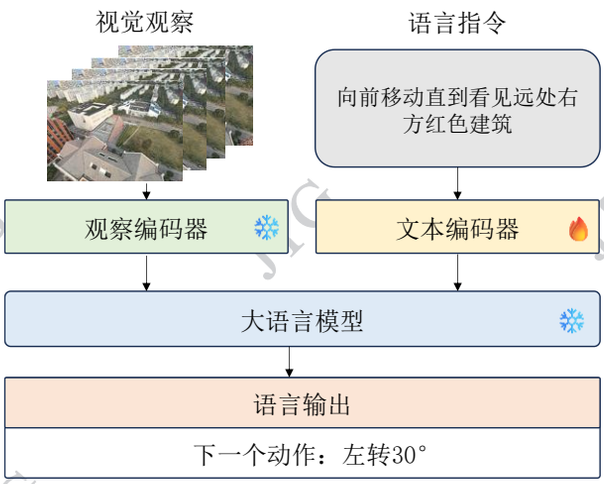


图4　Navid结构

Fig. 4　The Architecture of Navid

在现有的研究路线中，一类方法侧重于通过构建显式的空间结构来增强模型对大尺度环境的记忆与定位能力。例如，语义-拓扑-度量表征（semantic-topo-metric representation，STMR）通过将指令相关的地标语义掩码投影到顶视图地图中，并进一步转化为结构化矩阵，为大语言模型提供了具备度量信息的空间模板；而AeroDuo则利用高空视角构建全局正射投影图，通过跨高度的视觉协同来消除视频序列中的时空模糊性。与此类显式建模不同，另一种路径倾向于在不改变模型架构的前提下，通过视觉提示工程（Visual Prompting）提升感知的精度。ViSA（Tong 等，2026）框架利用标记集（set-of-mark，SoM）技术在图像上覆盖数值ID，将复杂的场景转化为可引用的结构化视觉表征，从而显著降低了多模态模型在空间推理中的“幻觉”现象。

随着端到端预训练技术的发展，感知表征正朝着几何对齐与预测性表征的方向深耕。SpatialFly（Li 等，2026）提出了几何引导的对齐机制，通过将全局结构线索引入语义词元，实现了二维视觉与三维轨迹决策空间的重参数化对齐。AutoFly（Sun 等，2026）则通过引入伪深度编码器，使模型在纯RGB输入下也能感知环境的深度信息与避障约束。与此同时，NaVid（Zhang 等，2024）等方法追求极简的纯视频流感知，通过大规模数据预训练让模型自发学习时空表征，摆脱了对地图、里程计或深度计的依赖，其结构如图4所示。更具前瞻性的研究如AirScape（Zhao 等，2025），将感知提升到了“世界模型”的高度，其表征不仅包含当前状态，更整合了运动意图导向的未来视觉预测。

综合来看，无人机视觉语言导航的感知表征正从单纯的视觉编码转向一种融合了几何先验、结构化提示与时空演化逻辑的综合性表达。这种演进不仅增强了模型对复杂地标关系的识别能力，也为后续的推理范式和具身控制提供了更加鲁棒且具备空间常识的输入特征，实现了从“看见环境”到“理解空间”的根本跨越。

**表3　感知表征方法对比**

**Table 3　Comparison of perception representation methods**

| 技术路线 | 核心思想 | 优点 | 缺点 | 代表工作 |
|---|---|---|---|---|
| 单一视觉编码 | 特征提取 | 简单快速 | 缺乏空间信息 | 卷积神经网络 |
| 跨模态对齐 | 视觉-语言特征融合 | 语义匹配强 | 几何理解弱 | VLN BERT、Airbert |
| 几何增强表征 | 物理先验 | 空间定位准 | 计算量大 | SpatialFly、AutoFly |
| 世界模型 | 环境预测与 动态建模 | 鲁棒性强 | 训练复杂 | AirScape |

### 4.2　推理范式

在无人机视觉语言导航的链路中，推理范式承载着将高层语义指令转化为底层动作序列的逻辑中枢功能，其核心定义在于代理如何建立多模态输入与三维连续空间运动之间的因果关联。传统的推理范式如Chen 等人（2024）提出的基于视觉-语言-控制的Transformer模型（vision-language-control transformer，VLCT），多依赖交叉注意力机制在文本词元与视觉块之间进行隐式特征映射，这种“黑盒”映射方式虽然在简单场景下有效，但在处理长程复杂的

城市级导航时，往往缺乏对空间拓扑结构的显式理解。无人机视觉语言导航推理范式对比如表4所示。

为了克服隐式推理的局限，研究者开始引入外部先验知识与多尺度表征来增强推理的逻辑性。例如，CityNav引入地理语义图作为空间导航的脚垫，而Hong等(2024)提出的基于多提示的视觉语言导航(vision-language navigation with multi-prompts, VLN-MP)则通过视觉提示(Visual Prompts)将纯文本推理转化为更具确定性的模态对齐推理。随着大语言模型(LLM)与多模态大模型(vision-language model, VLM)的崛起，推理范式经历了从“反应式映射”向“显式逻辑规划”的质变。FlightGPT (Cai等, 2025)与FreeFly-Thinking分别通过思维链(chain-of-thought, CoT)技术，诱导模型在输出控制参数前先生成中间层级的推理步骤，如地标定位、方位判定及意图拆解，显著提升了决策的可解释性，其结构如图5所示。同时，NavAgent(Liu等, 2024)强调了多尺度视图融合在推理中的作用，而GeoNav(Xu等, 2026)则通过模拟人类“从粗到细”的认知模式，构建双尺度空间表征以实现跨越公里级的长距离地理推理。此外，针对安全性的考量，基于安全控制屏障的自适应安全动作屏蔽机制(adaptive safety movement assurance, ASMA)(Sanyal等, 2025)等方法将推理输出与场景感知的安全约束(如控制障碍函数)相结合，确保逻辑路径在物理层面的可靠执行。

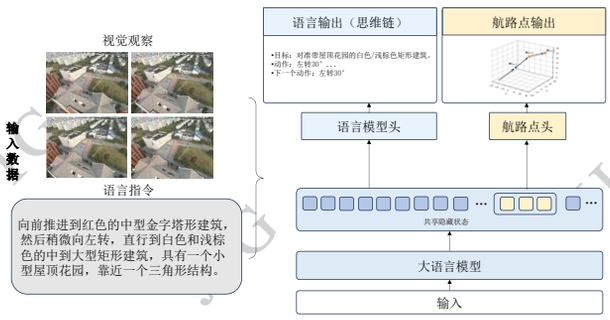


图5　FreeFly-Thinking结构

Fig. 5　The Architecture of FreeFly-Thinking

在上述基础上，近年大模型驱动的无人机视觉语言导航推理范式进一步从单一的语言规划或视觉问答式决策，向认知模块化、方向关系建模和实时决策解耦等方向发展。早期方法多将大语言模型或视觉语言模型作为统一决策器，直接根据当前观测和语言指令生成下一步动作或候选航点；而新近研究则更加关注复杂长程任务中的子目标拆解、空间关系消歧、过程监控和推理稳定性，使推理模块逐渐从“单步反应式决策”转向“多模块协同决策”。

在认知模块化推理方面，FineCog-Nav(Shao等, 2026)将无人机视觉语言导航过程拆解为语言处理、感知、注意力、记忆、想象、推理和决策等细粒度认知模块。与依赖单一大模型直接输出动作的方式不同，该方法通过角色化提示和结构化输入输出协议，使不同基础模型分别承担指令解析、子目标提取、场景理解、视觉注意、状态想象和动作选择等功能，从而形成由多个认知环节协同构成的零样本导航框架。该类方法的意义在于，它将长程导航中的隐式推理过程显式化，有助于提升复杂指令执行过程的可解释性，并缓解大模型在无人机三维场景中对全局状态、阶段目标和历史行为理解不足的问题。与此同时，FineCog-Nav构建了AerialVLN-Fine，对AerialVLN中的部分轨迹进行句子级指令—轨迹对齐和精细化地标标注，为细粒度评估无人机导航推理能力提供了新的参考。

在空间关系推理方面，LookasideVLN(Ning等, 2026)进一步指出，城市级无人机视觉语言导航中的指令理解不能仅依赖地标名称匹配，还需要充分利用“左转”“右侧”“沿着”“经过”等方向性语言线索。并且，LookasideVLN从历史导航经验中检索候选地标的位置与描述，并对当前观测、语言指令和方向性地标信息进行联合推理。该类方法表明，在无人机长程导航中，方向词和相对空间关系不仅是辅助信息，而且可以作为约束路径选择和消除地标歧义的重要推理线索。

在实时决策与过程监控方面，OnFly(Zheng等, 2026)面向机载零样本无人机视觉语言导航提出了共享感知的双智能体推理框架。该方法认为，单流视觉语言模型决策容易将高频目标生成和低频任务进度监控耦合在同一推理过程中，导致响应速度、监控稳定性和决策一致性之间出现冲突。因此，OnFly将推理过程划分为高频决策智能体和低频监控智能体：前者根据实时观测生成飞行目标，后者负责跟踪长程任务进度、判断终止或恢复信号。两类智能体共享视觉感知特征，但维护独立的推理上下文和KV-cache，从而在减少重复计算的同时提高实时导航的稳定性。该思路说明，大模型推理在真实

无人机系统中不仅要关注语义理解能力，还需要考虑机载计算、推理频率、长程监控和飞行安全之间的协调关系。

综合来看，无人机视觉语言导航的推理范式正朝着高度可解释、具备地理空间觉知以及层级化规划的方向演进。这种演进不仅体现在从离散动作到 6-DoF 连续轨迹预测的跨越，更体现在模型能够利用大语言模型与视觉语言模型的常识推理、空间消歧和任务拆解能力处理模糊指令，并通过空间图结构、认知模块、方向关系建模或双智能体机制弥补视觉观测的局部性与单流决策的不稳定性。总体而言，近年的大模型推理范式正在从“语言指令—视觉观测—动作输出”的直接映射，发展为面向复杂三维环境的分层认知决策框架，推动无人机视觉语言导航在动态且不确定的真实世界环境中实现更具鲁棒性、可解释性和可部署性的具身决策。

表 4　推理范式对比

Table 4　Comparison of reasoning paradigms

| 阶段 | 方法 | 特点 | 可解释性 | 适用任务 |
|---|---|---|---|---|
| 传统推理 | 交叉注意力、Transformer | 端到端黑盒 | 差 | 短程简单导航 |
| 增强推理 | 多尺度融合、外部先验 | 空间感知提升 | 中 | 复杂场景 |
| 大模型推理 | LLM+CoT、层级规划 | 逻辑推理强 | 高 | 长程、模糊指令 |

### 4.3　记忆存储

在无人机视觉语言导航中，记忆存储被定义为对时空观测信息与指令特征进行结构化编码与持久化保留的机制。其核心任务是打破即时感知的局限，通过在时序上整合历史视觉特征、在空间上构建环境表征，为智能体提供具有全局一致性的上下文信息，从而解决长程导航中的指令对齐偏差与状态漂移问题。无人机视觉语言导航记忆存储机制对比如表 5 所示。

从具体实现方法来看，记忆存储的演进呈现出从隐式表征向显式地图转换的趋势。早期方法如 Singla 等（2020）主要依赖隐式神经记忆，通过长短期记忆网络（long short-term memory，LSTM）及其隐藏状态来捕获时序关联，辅以时间注意力机制在历史深度观测中提取关键特征。这类方法处理速度快，但面对超长路径时易出现信息遗忘且缺乏空间几何约束。随后，显式空间地图逐渐成为主流，如 Zhao 等（2025）和 Zhang 等（2026）提出的网格化存储方案，通过构建基于鸟瞰图（bird′s-eye view，BEV）的语义地图或三维占据栅格，将 VLM 提取的语义得分反投影至空间网格中，生成包含“吸引力”、“已探索区域”和“障碍物”的多维记忆图谱，这种方式增强了决策的可解释性并能精确处理无人机特有的高度维度。CityNavAgent（Zhang 等，2025）则进一步将记忆抽象为拓扑图，通过记录历史轨迹节点与全景观测，将连续的物理空间离散化为可搜索的图结构，极大地缩减了长程探索的决策空间，具体工作流如图 3 所示。此外，针对连续导航中的误差累积问题，Tang 等（2026）引入了增强型滤波器记忆，将历史高置信度观测作为锚点存储在记忆库中，通过贝叶斯状态估计递归地修正智能体的位姿漂移。

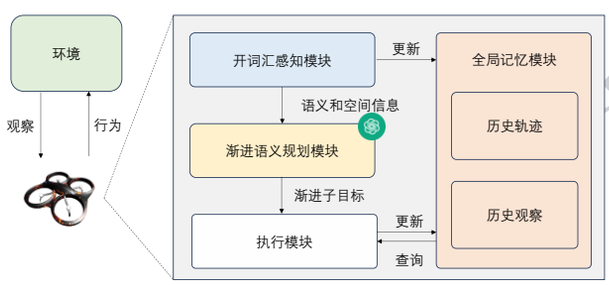


图 6　CityNavAgent 工作流程

Fig. 6　The overall workflow of CityNavAgent

随着大语言模型和视觉语言模型进入无人机视觉语言导航，记忆存储的功能也从单纯保存视觉特征或空间地图，进一步扩展为支持语言检索、任务分解、进度监控和安全执行的认知支撑结构。与传统地图记忆强调“空间在哪里”不同，大模型驱动的记忆机制更关注“当前指令进行到哪一步”“历史观测中哪些地标与当前语言相关”“哪些关键帧能够支撑长程任务判断”等问题。因此，近期研究开始将记忆划分为地标知识库、层级任务记忆和混合关键帧记忆等不同形式，以适应城市级长程导航、零样本推理和机载实时部署的需求。

在层级任务记忆方面，FineCog-Nav（Shao 等，2026）将记忆模块嵌入到认知式导航流程中，通过层级设计将长程轨迹中的局部细节、阶段目标和全局指令状态分别存储，既能保留连续导航过程中的关键上下文，又能通过摘要压缩减少冗余信息对决策的干扰。对于无人机视觉语言导航而言，这种记忆形式适合处理长指令、多地标和多阶段任务，有助于缓解“已完成子目标遗忘”和“当前动作与总体指令脱节”等问题。

在机载实时记忆方面，OnFly（Zheng 等，2026）针对零样本无人机视觉语言导航中的长程进度监控问题，提出将混合记忆用于低频进度监控智能体，使其能够在长程导航过程中判断任务是否完成、是否需要恢复或重新规划。该设计说明，真实无人机系统中的记忆机制不仅要考虑语义完整性，还需要同时满足机载计算延迟、缓存稳定性和安全监控的工程约束。

通过对比可以发现，记忆存储已由单纯的视觉特征堆栈，转向深度融合语义规划、空间结构、任务进度与机载部署约束的综合体系。早期的时序记忆侧重于局部动态避障，显式地图记忆更强调对环境几何与拓扑结构的整体把握，而近期大模型驱动的记忆机制则进一步突出任务层级性和实时监控稳定性。FineCog-Nav 的多层级记忆体现了从轨迹历史保存向子目标状态管理的转变，OnFly 的混合关键帧记忆则体现了从离线长程上下文建模向机载实时监控的转变。未来的发展趋势在于，将大规模预训练模型的开放词汇知识更高效地植入动态更新的记忆库中，同时结合异步计算、关键帧压缩和语义检索机制，解决高分辨率空间存储与实时飞行控制之间的矛盾。这种从“感知即记忆”到“记忆指引决策”的范式转变，使无人机能够在复杂城市环境中通过回溯历史知识、识别任务阶段和检索相关地标，实现更高效的逻辑推理与路径寻迹。

**表 5　记忆存储机制对比**

**Table 3　Comparison of memory storage mechanisms**

| 记忆类型 | 实现方式 | 优点 | 缺点 | 适用场景 |
|---|---|---|---|---|
| 隐式时序记忆 | LSTM、注意力 | 轻量、速度快 | 易遗忘、无空间结构 | 短程导航 |
| 显式语义地图 | BEV、栅格地图 | 空间清晰、可解释 | 存储开销大 | 城市大场景 |
| 拓扑图记忆 | 节点 - 边结构 | 搜索高效、抗漂移 | 构建复杂 | 长程导航 |
| 滤波增强记忆 | 卡尔曼滤波修正 | 定位稳定、误差小 | 依赖先验 | 连续控制 |

### 4.4　具身控制

在无人机视觉语言导航的链路中，具身控制被定义为将高层感知表征与推理范式输出的抽象语义指令或空间导航意图，转化为无人机在三维物理世界中可执行的连续动作序列或底层控制信号的过程。这一环节不仅要处理从符号语义到物理动力学的跨模态映射，还必须在严格的实时性约束下，满足无人机六自由度运动的平滑性与安全性要求。

对比当前主流的方法，具身控制范式正经历从模块化解耦向端到端动作生成的演进。传统的模块化控制方法如 VLA-AN（Wu 等，2025），通常将控制链路分为“轨迹规划”与“跟踪控制”两个阶段，模型首先预测局部路标点或速度矢量，再由模型预测控制这样的传统控制器进行轨迹平滑与物理执行，如图 7 所示，这种方式虽然具备极强的物理可解释性，但在应对动态环境时存在模块间误差累积的问题。相比之下，近期涌现的端到端模型如 AerialVLA（Xu 等，2026）和 UAV-Track VLA（Zhang 等，2026），则尝试打破语义与动作的屏障，通过将连续控制信号（如位移增量、欧拉角或推力）直接离散化为数值词表中的 Token，利用大语言模型的推理能力直接生成动作块，显著提升了系统的响应敏捷度。然而，端到端生成的随机性也带来了安全风险，因此 NavRL（Xu 等，2025）等研究引入了基于速度障碍法或几何排斥场的安全屏蔽机制，在神经网络输出与物理执行器之间增加一层“安全护栏”，以修正可能

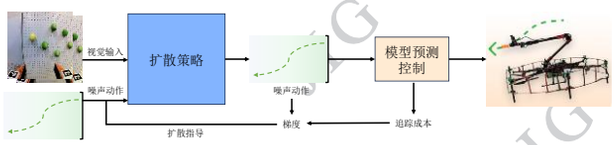


图 7　VLA-AN 工作流程

Fig. 7　The overall workflow of VLA-AN

导致碰撞的非安全指令。此外,针对跨设备部署的动力学差异,UMI-on-Air (Gupta 等,2025)等先进方法提出了具身感知引导策略,利用底层控制器的跟踪代价反馈来修正高层动作策略,实现了不同无人机硬件间的灵活适配。

在开放词汇目标理解与连续控制结合方面,VLFly(Zhang 等,2025)进一步展示了大模型从语言定位到无人机连续飞行控制的完整链路。该方法首先利用大语言模型将高层自然语言指令转化为结构化提示,以增强语言目标与视觉表征之间的语义一致性;随后通过视觉语言模型在候选图像池中进行跨模态目标检索,确定与指令最相关的目标图像;最后基于当前第一视角观测、历史视觉帧和目标图像生成可导航航点,并将航点进一步转化为线速度、角速度等连续控制命令,实现实时飞行执行。与依赖固定离散动作空间的传统视觉语言导航方法不同,VLFly 直接面向无人机连续运动控制,减少了地面视觉语言导航方法迁移到空中平台时的动作空间不匹配问题。同时,该方法不依赖外部定位或主动测距传感器,而是主要利用机载单目视觉实现目标定位和轨迹生成,体现了开放词汇语义理解、轻量视觉感知和连续控制之间的结合趋势。

除严格意义上的语言指令跟随型无人机视觉语言导航外,部分研究也开始探索视觉语言模型在无地图自主导航与低成本避障控制中的作用。例如,VLM-Nav(Sarker 等,2026)利用单目 RGB 图像生成深度图,再由视觉语言模型分析障碍物分布并给出避障相关反馈,最后结合相对航向角、左右距离传感器信息和近邻障碍检测结果输入导航模型,预测无人机的下一步动作。该方法并不依赖外部自然语言指令,因此更接近视觉语言模型辅助的无地图无人机自主导航。但它从另一个侧面说明,视觉语言模型不仅可以用于目标定位和语义推理,也可以作为低成本无人机控制链路中的环境理解模块,为障碍识别、路径选择和实时避障提供辅助信息。对于资源受限的机载平台而言,这类方法为利用单目视觉替代昂贵传感器、提升未知环境泛化能力提供了有益参考。

总体来看,无人机视觉语言导航的具身控制正在从单纯的离散动作选择,逐步发展为面向三维连续空间的语义—几何—动力学协同控制。一方面,视觉语言动作模型类方法尝试将视觉、语言和动作统一到同一模型框架中,推动无人机控制从模块化规划走向端到端动作生成;另一方面,VLFly、OnFly 等工作表明,即使在大模型参与决策的框架下,连续航点生成、语义—几何校验、安全屏蔽和局部轨迹规划仍然是确保真实飞行可执行性的关键环节。未来的具身控制研究需要在模型智能性与物理可靠性之间取得平衡:既要利用大模型的开放词汇理解、任务分解和空间推理能力,又要引入可验证的安全约束、实时控制机制和跨平台动力学适配策略。只有将高层语义决策与底层飞行控制稳定耦合,无人机视觉语言导航才能真正从仿真评测走向复杂低空环境中的可靠部署。

### 4.5 多机协同语言导航

随着无人机视觉语言导航任务从单目标、短距离导航扩展到大范围巡检、灾害搜救、环境监测和低空物流等复杂场景,单架无人机在视场覆盖、续航时间、任务效率和系统鲁棒性方面逐渐暴露出能力边界。多机协同语言导航旨在使多架无人机共同理解同一自然语言任务,并在共享语义目标、空间约束和安全规则的基础上完成分布式感知、协同规划与联合执行。与单机无人机视觉语言导航相比,多机协同语言导航并不是简单地增加无人机数量,而是需要进一步解决语言任务分解、角色分配、跨机信息共享、协同决策、通信受限推理以及多机安全避碰等问题。因此,该方向可以被视为无人机视觉语言导航从“单智能体理解与执行”向“群体智能协商与执行”的进一步扩展。

从技术基础来看,近年来“大模型 + 多无人机/机器人协同”的相关研究为多机协同语言导航提供了重要参考。UAV-CodeAgents (Sautenkov 等,2025) 构建了一个基于 LLM/VLM 的多智能体无人机任务生成框架,通过推理与行动(reasoning and acting, ReAct)范式将自然语言任务解析、卫星图像理解、像素级语义 grounding 和多无人机轨迹生成结合起来,使智能体能够在“观察—描述—推理—决策—执行”的循环中动态修正任务计划。该系统由空域管理智能体和 UAV agent 组成,前者负责解释自然语言指令、分析卫星图像并生成空间定位的任务计划,后者根据分配航点执行飞行、采集图像,并将语义摘要和置信度反馈给管理智能体,从而支持协同任务更新和动态重规划。该工作虽然更偏向多无人机任务规划,而非严格意义上的无人机视觉语言

导航，但其展示了大模型在语言驱动任务分解、地理空间目标定位和多智能体协同规划中的作用。

RALLY（Wang 等，2025）则从群体智能导航角度探索大语言模型对无人机集群协同决策的增强作用。该方法通过结构化自然语言实现无人机之间的语义通信和协同推理，并结合动态角色异构机制，使不同智能体能够根据任务状态进行角色切换和个性化决策。同时，RALLY 将大语言模型离线先验与多智能体强化学习的在线策略结合，用于提升多目标覆盖、收敛速度和泛化能力。与传统多智能体强化学习主要依赖数值状态或固定通信协议不同，该类方法说明自然语言可以成为多机协同中的高层语义通信媒介，为未来多无人机语言导航中的任务协商和角色分配提供了启发。

在群体控制与人机交互方面，FlockGPT（Lykov 等，2024）和 SwarmGPT-Primitive（Vyas 等，2024）等工作进一步展示了语言模型在无人机集群编队和安全运动规划中的潜力。FlockGPT 利用大语言模型将用户的自然语言描述转化为目标几何结构，并结合符号距离函数等空间表达方式，引导无人机群在不同目标形态之间平滑过渡，实现自然语言驱动的集群队形控制。SwarmGPT-Primitive 则将大语言模型的高层编排能力与安全运动原语相结合，通过安全过滤器在可行性、碰撞约束、下洗气流和执行器限制等方面对生成轨迹进行修正，使非专业用户能够通过语言交互生成可部署的无人机群体动作。虽然这类工作主要面向编队控制、集群表演或群体运动规划，而不是导航指令跟随任务，但它们为多机协同语言导航提供了两个关键启示：一是自然语言可以作为人类与无人机群之间的直观交互接口；二是大模型生成的高层群体策略必须与底层安全约束和可执行运动规划相结合。

在直接面向多无人机协同视觉语言导航的研究中，AeroDuo 是目前较具代表性的工作使两架无人机在不同高度协同完成导航：高空无人机负责宽视场环境理解和目标推理，低空无人机负责精细导航和目标定位。为支持该任务，AeroDuo 构建了 HaL-13k 数据集，包含 13，838 条高低空协同示范轨迹，并为每条轨迹配备面向目标的语言指令，同时设置未见地图和未见目标验证集，以评估模型在新环境和新目标上的泛化能力。在方法上，AeroDuo 采用双无人机协同框架，高空无人机集成多模态大语言模型进行目标推理，低空无人机则使用轻量化多阶段策略完成导航与目标定位，两者仅交换少量坐标信息以降低通信开销。该工作较为直接地将多机协同引入无人机视觉语言导航任务，说明不同高度、不同视角的无人机可以通过角色分工弥补单机视野受限和长程导航不稳定的问题。

综合来看，现有“多无人机 + 语言/大模型 + 协同规划/控制”研究已经为多机协同语言导航提供了初步技术基础。其中，UAV-CodeAgents 强调多智能体任务规划与地理空间定位，RALLY 强调语言化语义通信和角色自适应，FlockGPT 与 SwarmGPT-Primitive 展示了自然语言对无人机群体行为和安全运动原语的编排能力，而 AeroDuo 则进一步将双无人机协同引入无人机视觉语言导航任务本身。需要注意的是，当前真正直接面向多无人机协同视觉语言导航的研究仍然较少，多数相关工作仍停留在任务规划、编队控制、集群表演或通用多智能体协作层面，尚未充分解决自然语言指令下的多机协同感知、跨视角语义对齐、共享记忆构建、通信受限决策和安全协同控制等核心问题。

因此，多机协同语言导航仍是无人机视觉语言导航领域中有待深入探索的重要方向。未来研究需要进一步构建标准化的多无人机语言导航任务定义、仿真平台、数据集和评测指标，并重点关注以下问题：如何将一条自然语言指令自动拆解为多个可执行子任务，如何根据无人机高度、视角、传感器、电量和位置状态进行动态角色分配，如何在带宽受限条件下共享语义地图和任务进度，如何通过跨机记忆减少重复搜索和目标遗漏，以及如何将多机避碰、空域约束和故障替换机制嵌入语言驱动的协同决策链路。只有将单机的视觉语言理解能力扩展为多机的协同认知与联合执行能力，才能真正支撑灾害救援、城市巡检、环境监测和智能物流等大范围低空任务中的“一条指令、多机协同、全域执行”。

### 4.6 技术路线横向比较与演进脉络

从技术发展脉络看，无人机视觉语言导航方法总体经历了从模块化规划、学习式策略、跨模态预训练，到大模型驱动具身决策的演进过程。其横向比较如表6所示。

早期方法通常采用模块化框架，将语言解析、视觉识别、路径规划和飞行控制分开处理。这类方法具有较强可解释性和工程可控性，便于显式加入障

碍物约束、飞行动力学约束和安全边界，但其各模块之间容易出现误差累积，并且难以处理开放自然语言指令和复杂语义关系。

随后，模仿学习和强化学习方法推动无人机视觉语言导航从规则驱动转向数据驱动。该类方法通过专家轨迹或交互奖励学习从视觉—语言输入到动作输出的映射，在特定仿真环境中能够获得较好的导航性能，适合处理局部避障、路径跟随和短程目标搜索任务。然而，学习式方法通常依赖大规模训练数据，跨场景泛化能力不足，在未见环境、长距离导航和真实飞控扰动条件下容易出现策略失效。

跨模态预训练方法进一步提升了视觉与语言之间的语义对齐能力。基于 Transformer、跨模态注意力和视觉语言预训练的模型能够在图像区域、地标实体和指令词元之间建立更稳定的对应关系，适合解决复杂地标识别、空间关系理解和语言定位问题。但此类方法多数仍侧重表征学习，对三维几何关系、连续动作可执行性和长程空间记忆的建模不足，难以单独支撑真实低空环境中的完整导航闭环。

近年来，大语言模型和视觉语言模型的引入使无人机视觉语言导航进入认知驱动阶段。大模型方法能够对自然语言任务进行层级拆解、常识推理和风险判断，并通过思维链、工具调用或视觉提示机制提升复杂任务的可解释性。与传统端到端策略相比，大模型驱动方法更适合处理模糊指令、开放目标和任务规划问题，是当前无人机视觉语言导航的重要发展方向。然而，大模型方法也带来实时推理开销大、输出不稳定、物理约束弱和安全验证困难等问题。因此，未来技术路线不应简单追求端到端大模型替代全部模块，而应将大模型的语义推理能力与传统规划控制的安全性、可验证性相结合，形成“高层语言认知—中层空间规划—底层安全控制”的分层闭环架构。

总体而言，不同技术路线并非相互替代，而是面向不同问题层次形成互补关系。模块化规划方法适合工程安全约束明确的任务；学习式策略适合在仿真环境中进行大规模策略优化；跨模态预训练方法适合增强视觉语言定位能力；记忆与地图增强方法适合长程导航和空间一致性维护；大模型方法则适合开放任务理解、任务分解和复杂语义推理。无人机视觉语言导航未来的关键趋势，是将上述路线融合为可解释、可泛化、可验证、可部署的视觉—语言—动作系统，使无人机能够在真实低空环境中完成长程、安全、交互式的自主导航任务。

**表6　无人机视觉语言导航技术路线横向比较**

**Table 6　Horizontal Comparison of Technical Routes for UAV Vision-Language Navigation**

| 技术路线 | 核心思想 | 主要优势 | 主要局限 | 适用场景 |
|---|---|---|---|---|
| 模块化规划方法 | 语言解析、视觉识别、路径规划和控制分模块实现 | 可解释性强，易加入安全约束 | 模块误差累积，开放语言理解能力弱 | 工程巡检、已知环境飞行 |
| 模仿学习 / 强化学习方法 | 从专家轨迹或交互奖励中学习导航策略 | 可学习复杂策略，适合仿真训练 | 数据需求大，跨场景泛化弱 | 短程导航、局部避障、策略优化 |
| 跨模态预训练方法 | 通过 Transformer 和注意力机制对齐视觉与语言 | 语义匹配能力强，适合地标 grounding | 三维几何和连续控制建模不足 | 地标导航、指令跟随、目标搜索 |
| 记忆与地图增强方法 | 构建语义地图、拓扑图或时序记忆 | 适合长程导航和空间一致性维护 | 建图误差和记忆更新复杂 | 城市级导航、长距离任务 |
| 大模型驱动方法 | 利用 LLM/VLM 进行任务理解、推理和规划 | 开放语言理解强，可解释性较高 | 实时性、安全性和稳定性不足 | 模糊指令、开放目标、多步骤任务 |
| 分层融合方法 | 高层大模型推理 + 中层规划 + 底层安全控制 | 兼顾语义理解与物理可执行性 | 系统复杂度高，端到端优化困难 | 真实低空部署、复杂应用闭环 |

## 5　应　用

随着人工智能与机器人技术的深度融合，无人机视觉语言导航已从实验室环境下的理论探索逐步走向多元化的实际落地场景。得益于其卓越的空间机动性、非接触式作业能力以及日益增强的边缘感知效率，无人机在处理复杂、高危及动态变化的现实

任务中展现出传统地面机器人难以比拟的优势。与传统的基于全球定位系统路径点或同步定位与地图构建的几何导航方案不同，视觉语言导航范式赋予了无人机理解自然语言指令并与物理世界进行语义交互的能力，实现了从“坐标驱动”向“语义驱动”的范式转变。

本章将重点论述无人机视觉语言导航技术在提升工业效率、保障社会安全、优化物流配送及赋能智慧农业等方面的具体应用价值。在城市巡检与工业监测中，无人机通过语义理解与高精度感知，实现了对关键基础设施的自主诊断与安全监管；在灾害响应与应急搜索中，其利用多模态探测技术在极端环境下执行高效的人类目标搜救任务；在城市物流与安全交付中，无人机与地面车辆协同工作，有效解决了山地或拥堵城市环境下的“最后一公里”交付难题；最后，在精准农业与环境监测中，无人机结合多光谱成像与深度学习技术，为农作物的细粒度健康监测与精准干预提供了闭环解决方案。无人机视觉语言导航应用场景如图3所示。

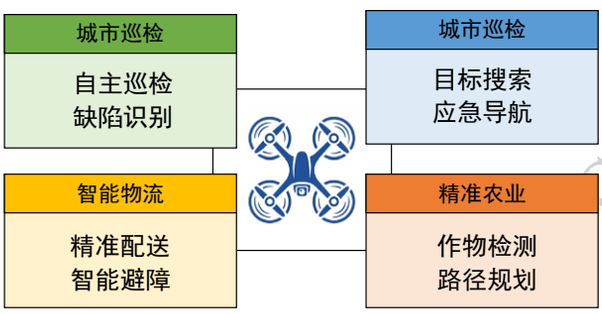


图8　无人机视觉语言导航应用场景

Fig. 8　Application Scenarios of UAV Vision-Language Navigation

### 5.1　城市巡检与工业监测

无人机凭借机动灵活、视角覆盖范围广、非接触式作业和超视距飞行等优势，已成为城市基础设施巡检与工业安全监测中的重要技术手段。相比传统直升机巡检或人工攀爬维护模式，无人机能够在高压线路、桥梁、塔架、厂区管廊、建筑外立面等高危或难以抵近区域执行快速巡查，降低人工风险并提升巡检效率。现有相关研究虽然多数仍属于无人机视觉感知或自主巡检范畴，但它们为无人机视觉语言导航在城市巡检与工业监测中的落地提供了关键技术支撑。以电力巡检为例，研究者通常利用激光雷达、可见光相机等传感器对输电线路进行扫描，构建高精度三维点云模型，并结合智能控制算法生成巡检路径；同时，针对高压传输线图像容易受到飞行振动、光照变化和环境噪声影响的问题，相关工作提出了图像预处理与中值滤波增强方法，以提升线路和缺陷目标的识别稳定性（Du 等，2022）。

随着边缘计算与轻量化检测模型的发展，无人机工业监测进一步从离线数据采集转向边缘侧实时感知与即时告警。例如，针对 10 kV 配电网带电作业这一高危场景，研究人员提出基于特征增强模块的 YOLOv8 实时智能检测算法，并将其部署于边缘计算设备，实现对作业人员绝缘头盔、披肩、手套等个人防护装备的高精度识别（Duan 等，2025）。该类研究能够在无人机巡检过程中实时发现设备缺陷、违规操作或安全隐患，为后续的语义理解、任务重规划和人机交互提供可靠感知输入。此外，工业监测也逐渐从“看见异常”走向“接近异常并执行操作”。例如，LOCATOR 利用双固态激光雷达融合感知实现对架空电缆的厘米级自主定位与对齐导航，并结合软体抓取机构完成挂载式物理交互，为无人机执行线路维护、异物清除和近距离检测等任务提供了技术基础（Iversen 等，2020）。因此，上述工作虽然不完全等同于无人机视觉语言导航，但分别在高精度感知、缺陷识别、边缘推理、路径生成和自主对齐控制等方面构成了无人机视觉语言导航应用闭环的底层支撑。

近年来，一些面向城市和工业场景的 无人机视觉语言导航工作进一步体现了这种闭环趋势。Sky-VLN 将视觉语言导航与非线性模型预测控制结合，使无人机能够根据自然语言指令和第一视角图像在城市环境中导航，并通过空间描述、历史路径记忆和避障控制提升导航稳定性（Li 等，2025）。CoDrone 提出了云—边—端协同的无人机自主导航框架，将机载轻量模型、边缘深度估计模型和云端视觉语言模型结合起来，用于提升复杂环境下的感知和安全导航能力（Chen 等，2025）。

此外，Herron 等（2025）提出的层级式智能体框架面向工业视觉巡检任务，由头代理进行高层规划，工作代理控制具体无人机执行动作，从而实现“规划—推理—执行—评估”的闭环过程，可用于读取压力表、检查设备状态等工业任务。HUGE-Bench（huge-scale UAV geospatial common sense and embodied

navigation benchmark)则将建筑巡检、道路巡检、区域测绘和避障穿越等任务纳入高层无人机视觉语言动作评测框架,强调短指令下的多阶段执行和碰撞安全评估,为城市巡检类任务提供了更贴近真实应用的评测方式(Guo 等,2026)。

因此,城市巡检与工业监测中的无人机视觉语言导航价值,不只是提高缺陷检测精度,而是让无人机能够根据人的自然语言意图主动完成巡检任务。例如,巡检人员可以发出"沿 3 号输电线路向北巡检,重点检查靠近变电站一侧的绝缘子""绕到厂房西侧管廊上方,检查疑似锈蚀的阀门区域"等指令。无人机需要理解指令中的目标、方位和安全要求,结合视觉、深度或点云信息定位目标区域,并在飞行过程中调整高度、角度和距离,最终完成异常复查和结果反馈。

综上,已有无人机巡检研究为无人机视觉语言导航提供了视觉识别、边缘推理、目标定位和安全控制等基础能力;而 SkyVLN、CoDrone、层级式智能体框架和 HUGE-Bench 等近期工作,则进一步展示了自然语言理解、视觉语义定位、任务规划和安全执行在无人机巡检闭环中的作用。通过这种方式,无人机可以从"被动航拍采集平台"转变为"可交互、自主决策的空中巡检智能体",从而更直接地支撑无人机视觉语言导航在城市基础设施和工业监测中的应用价值。

### 5.2 灾害响应与应急搜索

无人机在灾害响应与应急搜索任务中展现出了极高的应用价值,特别是在人员难以进入的高危环境下,能够显著缩短搜救时间并降低救援人员风险(Alpiste 等,2021)。快速定位幸存者是提高生存率的关键,尤其是在灾后"黄金时间"内,无人机凭借其机动灵活、覆盖范围广和可搭载多源传感器等优势,已成为应急救援体系中的重要空中感知平台。相比传统人工搜索或地面车辆巡查方式,无人机能够在雪崩现场、森林火灾、地震废墟、洪涝灾区和山地失联区域等复杂地形中执行快速侦察与目标搜索任务,弥补人力搜救在可达性、视野范围和响应速度方面的局限性。

现有相关研究首先从目标感知与多模态探测层面为无人机灾害搜救提供了重要技术支撑。在视觉感知方面,深度学习方法显著提升了无人机对遇险人员和异常目标的识别能力。针对雪崩搜救场景,基于卷积神经网络的视觉系统能够从航拍图像中近实时检测滑雪者或其他遇险目标,为快速定位幸存者提供了技术基础(Bejiga 等,2016)。Li 等(2023)进一步展示了利用目标检测算法在不同环境下定位失踪人员的有效性,并分析了光照强度、遮挡条件等因素对搜救效率的影响。针对植被茂密、目标尺度小且背景干扰强的森林环境,轻量化深度学习架构也被用于无人机端视觉识别,在保证推理速度的同时提升林间受困人员检测能力,为复杂地形下的自主搜索与航迹调整提供了感知支撑(Yong 等,2018)。

除了视觉目标检测,多模态探测技术也为灾害搜救中的盲区搜索提供了重要补充。在雪崩、泥石流或建筑坍塌等灾害场景中,受困者可能被积雪、植被或废墟遮挡,单纯依靠可见光图像难以获得有效线索。Wolfe 等(2015)提出利用无人机搭载接收器探测受困者手机信号的方法,使无人机能够在缺乏直接视觉证据的情况下,通过无线电信号引导搜索方向并估计受害者位置。在完成搜索后,具备运输能力的无人机还可以携带急救包等物资,通过自动化投放系统将其送达幸存者手中,实现了从搜寻到物资投送的救援闭环(Jo 等,2017)。此类工作说明,灾害响应中的无人机搜救并不只是单帧图像检测问题,而是涉及视觉、无线电、地理环境、障碍物分布和任务时限等多源信息融合的复杂决策过程。这些感知、检测和信号定位研究虽然尚未完全形成视觉语言导航闭环,但它们为无人机视觉语言导航在灾害响应中的目标发现、环境理解和安全路径生成提供了必要的底层能力。

在此基础上,近年来视觉语言模型和大语言模型的引入,使无人机灾害搜救逐渐从"人工设定航点+被动目标检测"转向"语言任务驱动+主动视觉探索+安全导航执行"的闭环模式。例如,VLM-RRT(Ye 等,2025)将视觉语言模型引入路径规划过程,使模型能够根据环境快照提供方向性引导,从而将采样偏向更可能存在可行路径的区域,提升复杂环境中的规划效率和路径质量。该类方法特别适用于森林火灾、坍塌区域和障碍密集灾区等高风险场景,因为搜救任务不仅要求无人机找到目标,还要求其在火线、障碍物和禁飞区域之间快速生成安全航迹。FlySearch (Pardyl 等,2025)则从评测角度构建了面向无人机的三维户外视觉语言搜索环境,要求无人

机根据自然语言描述在城市或森林场景中寻找火灾、失踪人员、垃圾堆等目标，并通过连续视觉观测和文本动作指令完成探索。这类基准进一步说明，灾害搜救场景中的关键能力已经从静态目标识别扩展为长时序、主动式、语义驱动的空中搜索。UAV-VLRR（Yasheer 等，2026）则更进一步面向应急搜索与救援任务，将视觉语言模型和大语言模型的场景解释能力与带避障约束的非线性模型预测控制相结合，使无人机能够根据自然语言任务和航拍图像识别目标点与障碍物，并快速、安全地执行飞行响应任务。

因此，在灾害响应与应急搜索中，无人机视觉语言导航的实际应用价值主要体现在完整任务闭环的形成。救援人员可以通过自然语言下达高层任务，例如“沿河道向上游搜索被洪水围困的人员，优先检查屋顶和桥梁附近区域”“从山谷入口开始搜索穿红色衣服的失联人员，避开浓烟和火线区域”“绕过坍塌建筑，从东侧进入废墟上方低速巡查并寻找挥手或闪光信号”。无人机首先需要解析语言指令中的搜索目标、空间范围、优先级和安全约束；随后结合机载相机、热红外、深度、无线电或地理信息等多源观测，识别可能存在受困者或危险源的候选区域；再利用视觉语言推理、空间记忆和路径规划确定下一步搜索方向；最后通过连续飞行控制、安全避障和动态重规划完成抵近观测、目标确认与位置回传。与传统无人机搜救系统相比，无人机视觉语言导航不仅能够“看见”目标，还能够理解救援人员意图、主动选择搜索策略，并根据实时环境变化调整飞行轨迹。

综上，灾害响应与应急搜索为无人机视觉语言导航提供了最具现实迫切性的应用场景之一。已有无人机目标检测、多模态探测和自主路径规划研究，为该方向提供了视觉识别、信号定位和安全控制基础；而基于视觉语言模型和大语言模型的新近工作则进一步证明，自然语言指令、视觉场景理解和无人机连续导航可以被组织为统一的具身任务闭环。未来，随着多源传感器融合、长程空间记忆、动态风险建模和多机协同搜索能力的提升，无人机视觉语言导航有望在灾害现场形成“人类指挥—空中搜索—智能决策—安全执行—结果反馈”的高效应急响应体系，从而更直接地支撑无人机视觉语言导航的实际应用价值。

### 5.3 城市物流与安全交付

无人机在城市物流与末端配送中的应用已成为解决“最后一公里”交付难题的重要技术路径。相比传统地面物流模式，无人机辅助交付具有空域机动性强、响应速度快、受地面交通拥堵影响小等优势，尤其适用于高密度城市社区、山地城市、偏远区域以及突发交通阻断条件下的快速配送（Li 等，2023）。然而，城市低空物流并不是简单的点到点飞行任务，而是同时涉及复杂建筑环境理解、动态障碍规避、交付目标定位、用户交互、安全着陆和空域约束等多重因素。因此，城市物流与安全交付为无人机视觉语言导航提供了具有代表性的应用场景：系统不仅要完成路径规划，还需要理解用户自然语言需求，并在视觉观测基础上自主定位交付目标、选择飞行视角和执行安全交付动作。

现有无人机物流研究首先从路径优化、车机协同和自动化搬运等方面为无人机视觉语言导航的应用落地提供了技术支撑。针对复杂地形，特别是山地城市中的配送难题，研究者提出了无人机与地面车辆联合配送的协同模式。Liu 等（2022）面向山地城市复杂地理约束，构建了车机联合路径优化模型，并通过启发式算法寻求配送路径、时间成本与运行成本之间的平衡。在该体系中，无人机作为地面车辆配送能力的延伸，可以跨越道路难以覆盖的地形障碍，完成点对点的快速投递。这类工作虽然主要关注路径优化和资源调度，但为无人机视觉语言导航中“全局任务分配—局部空中导航—终端交付执行”的分层任务框架奠定了基础。

除了室外大尺度配送，室内工业物流与供应链管理中的无人机自动搬运研究也为安全交付提供了重要参考。Awasthi 等（2023）探讨了在工业车间或仓库环境中利用无人机集群进行自动化包裹搬运的可行性，并结合机器人操作系统和仿真环境展示了无人机在与人类共存空间中执行物料流转任务的能力。该类研究强调多机协同、安全避障和高效搬运，对于城市终端配送中的楼宇周边飞行、狭窄空间穿行和多无人机任务调度具有启发意义。不过，这些工作仍主要依赖预设任务、固定目标或几何路径规划，尚未充分体现自然语言指令理解、视觉语义定位和动态交互决策在实际交付中的作用。

随着多模态大模型和具身智能技术的发展，城市物流场景开始从“路径优化驱动的自动配送”转向

“语言交互驱动的视觉语言导航闭环”。LogisticsVISTA 提出了面向三维终端配送的综合框架，将无人机、无人车和无人船纳入统一的物流服务体系，并利用大语言模型和大多模态模型支持任务规划、主动感知、自主决策和人机交互(Tian 等，2024)。该框架的意义在于，它不再把无人配送车辆视为单纯的运输执行器，而是将其建模为能够理解用户需求、感知环境状态并动态调整交付策略的具身智能体。例如，当用户因临时原因无法前往原定交付点时，系统可以根据用户自然语言描述重新理解目标位置，引导无人机执行环境感知、障碍规避和精准降落等操作。这种模式使城市物流从“固定航点配送”扩展为“用户意图理解—环境感知—路径重规划—安全交付”的交互式服务过程。

更进一步地，LogisticsVLN 将视觉语言导航直接引入低空终端配送任务，提出面向窗口级交付的无人机视觉语言导航系统(Zhang 等，2025)。与多数面向长距离、大目标的无人机视觉语言导航研究不同，该工作聚焦住宅楼等精细化交付场景，要求无人机根据用户自然语言请求，在没有预建地图的情况下定位具体楼层和目标窗口。其系统流程包括请求理解、楼层定位、目标对象识别、视角选择、动作决策和深度辅助安全控制等环节：首先由大语言模型解析用户请求，提取目标楼层和窗口周边显著物体；随后利用视觉语言模型估计楼层位置并引导无人机上升到合适高度；当无人机到达目标楼层后，再结合目标识别、视角选择和深度信息，围绕建筑物搜索目标窗口并完成安全抵近。这一流程较为完整地呈现了城市物流场景中无人机视觉语言导航的核心闭环，即“自然语言请求—视觉语义定位—三维空间定位—飞行动作决策—安全交付执行”。

因此，在城市物流与安全交付中，无人机视觉语言导航的应用价值不应仅理解为提高配送速度或降低物流成本，而应体现在其对复杂用户需求和开放城市环境的语义适应能力上。用户可以用自然语言提出更接近真实生活的交付需求，例如“请把药品送到三楼带蓝色花盆的窗户旁”“将文件送到办公楼南侧二层露台，避开人群密集区域”“把外卖送到小区东门内侧的白色快递柜附近”。无人机需要从这些指令中解析交付物、目标楼层、视觉参照物、空间方位和安全约束；然后结合机载 RGB-D 观测、建筑外立面特征、深度估计和障碍物检测结果，逐步定位目标交付点；在执行过程中，还需要根据临时障碍、风场扰动、用户位置变化或禁飞区域动态调整航迹。相比传统基于坐标或航点的配送方式，无人机视觉语言导航能够把用户语言、视觉感知、空间推理和物理控制统一到一个连续交互过程中，从而更好地适应真实城市物流中的不确定性。

综上，城市物流与安全交付为无人机视觉语言导航提供了高度契合的实际应用场景。Li 等(2023)、Liu 等(2022)和 Awasthi 等(2023)等研究从无人机物流体系、车机协同路径优化和自动化搬运等方面提供了配送效率、安全控制和协同调度基础；LogisticsVISTA 和 LogisticsVLN 等近期工作则进一步把基础模型、自然语言交互、视觉语义定位和连续飞行动作结合起来，初步展示了低空终端配送中的无人机视觉语言导航闭环范式。未来，随着多模态感知、精细化空间定位、低空空域管理和人机交互机制的进一步成熟，无人机有望从“自动化配送工具”发展为能够理解用户意图、主动感知环境并安全完成交付的空中物流智能体，从而更直接地支撑无人机视觉语言导航在低空经济中的应用价值。

### 5.4 精准农业与环境监测

精准农业与环境监测是无人机技术的重要应用方向。现有研究已广泛利用无人机搭载 RGB、多光谱、热红外、激光雷达等传感器，实现作物长势评估、病虫害识别、水分胁迫诊断、产量预测和精准喷洒等任务。例如，Shendryk 等(2020)通过融合激光雷达与多光谱数据预测甘蔗生物量和叶片含氮量；Zhou 等(2021)结合图像纹理特征与植被指数诊断冬小麦水分胁迫；Kerkech 等(2020)利用深度学习语义分割方法实现葡萄园病害区域识别；Rajagopal 等(2023)则结合边缘计算和轻量化模型，对腰果树早期真菌感染进行检测并触发精准喷洒。这些工作虽然主要属于农业视觉感知与智能作业范畴，但为无人机视觉语言导航在农业场景中的目标识别、异常定位和任务执行提供了重要技术支撑。

与传统“预设航线采集—离线图像分析”的无人机农业应用不同，无人机视觉语言导航更强调自然语言指令驱动下的主动巡查与闭环决策。例如，农业管理者可以直接发出“沿西侧果园第三排巡查叶片发黄区域”“飞到温室北侧检查长势异常的植株”“沿湖岸线搜索水体颜色异常区域”等任务指令。无人机需要解析语言中的空间范围、目标对象和异常

属性，结合视觉、多光谱、深度或点云观测识别候选区域，并通过路径规划和视角调整完成近距离复查、异常定位和结果反馈。

近年来，面向农业无人机自主飞行的仿真平台也为该类闭环任务提供了支撑。例如，Agri-fly 构建了面向农业环境的无人机仿真系统，包含飞行动力学建模、传感器合成、冠层下自主飞行栈和高保真农业场景生成工具，可用于果园、葡萄园、温室和垂直农场等场景中的自主避障、视角选择和作物空间关系分析（Zha 等，2024）。这说明农业无人机任务已逐渐从单纯的图像采集和目标检测，扩展到复杂植被结构中的自主飞行、主动观测和空间推理。

因此，精准农业与环境监测中的无人机视觉语言导航应用价值在于将作物监测、异常识别、空间搜索和精准作业统一到“语言指令—视觉感知—空间定位—自主导航—结果反馈”的闭环中。通过这种方式，无人机不再只是被动采集遥感数据的平台，而可以根据农业管理者的自然语言意图主动选择巡查区域、调整飞行视角、定位异常目标并辅助后续干预，从而为智慧农业和生态环境监测提供更加灵活、智能和可交互的低空具身智能方案。

## 6 未来展望

尽管无人机视觉语言导航在仿真环境、数据集建设、核心算法与应用探索中已取得显著进展，但面向真实低空场景的规模化、鲁棒性、通用性部署仍面临诸多瓶颈。结合具身智能、多模态大模型、低空经济与无人系统技术的发展趋势，无人机视觉语言导航未来将朝着更强环境理解、更高决策智能、更安全物理执行、更高效人机协同、更深产业融合的方向持续演进，重点发展趋势可归纳为以下五个方面。

### 6.1 面向三维空间的世界模型与前瞻性具身推理

未来感知与推理将从“被动响应”升级为主动构建与前瞻推演的世界模型范式。通过融合三维几何先验、时序动态信息与环境常识，无人机可在飞行过程中实时构建具备预测能力的空中世界模型，实现对未观测区域、障碍变化、运动趋势与指令隐含意图的提前推理。同时，基于大模型的思维链、工具调用与自我纠错机制将进一步深化，使无人机能够处理模糊指令、复杂任务约束与开放环境常识推理，从“按步骤执行”提升至“自主理解、自主规划、自主处置”的认知级导航。

### 6.2 强泛化的跨域迁移与少样本适应能力

缩小仿真到现实的域间隙、提升未知环境泛化性是未来核心突破方向。依托领域自适应、合成数据生成、神经渲染重建与轻量化预训练，模型将在视觉纹理、物理动力学、光照气象、动态干扰等维度更贴近真实世界。同时，基于统一多模态大模型的跨场景、跨任务迁移将成为主流，使无人机在未见过的城区、野外、楼宇间等场景中，实现少样本语言导航，大幅降低定制化训练成本。

### 6.3 高安全、高可靠、高动态的具身控制与机载部署

面向真实飞行的安全性与机载实时性，未来具身控制将向安全约束嵌入、动力学感知、端到端可验证控制发展。通过将控制障碍函数、碰撞规避、故障容错与应急处置机制嵌入语言－动作生成链路，实现语义决策与物理安全的统一。同时，结合模型压缩、量化、知识蒸馏与机载边缘计算，推动大模型导航算法上机载、上实时、上稳飞，满足低空飞行低延迟、高可靠、强抗扰的关键需求。

### 6.4 多机协同与语言级集群智能

单一无人机的能力边界将向多机协同、群体智能扩展。未来研究将重点突破面向多无人机的语言指令解析、分布式空间记忆、协同任务分配、通信受限下的协同推理与安全避航技术，实现“一条指令、多机协同、全域执行”的集群语言导航。同时，多机间可通过语言进行信息交互、任务协商与故障替换，形成具备高鲁棒性、高覆盖能力与复杂任务执行能力的空中智能群体。

### 6.5 低空经济深度融合与标准化体系构建

长期来看，无人机视觉语言导航将成为低空数字经济与智能空天系统的核心使能技术。随着低空网络、数字孪生城市、空天地一体化监管与飞行服务体系逐步完善，语言导航将深度融入城市巡检、应急救援、智能物流、低空观光、环境监测等规模化场景。同时，面向安全、交互、算力、接口的行业标准与评测基准将逐步建立，推动技术从实验室走向产品化、标准化与产业化，最终实现人机自然交互、空中智能体自主可信运行的低空生态。

## 7 结 论

无人机视觉语言导航作为空中具身智能的核心前沿方向，深度融合计算机视觉、自然语言处理与无人机自主控制技术，突破了传统导航依赖卫星定位、缺乏高层语义理解、难以适应复杂非结构化环境的局限，实现了从高层自然语言指令到底层三维连续飞行动作的直接映射，为人机自然交互与低空智能自主作业开辟了全新路径。

本文系统梳理了无人机视觉语言导航的研究脉络，从仿真平台、数据集、核心技术、落地应用与未来挑战等维度，完整呈现了该领域从模块化流水线、端到端学习到大模型驱动认知推理的技术演进历程。当前，依托高保真仿真环境与大规模多模态数据集，感知表征、跨模态对齐、长程记忆与具身控制等关键技术持续突破，已在城市巡检、灾害救援、智能物流、精准农业等场景展现出显著应用价值。

但面向真实低空场景的规模化落地，仿真与现实迁移、多机协同语言导航、长程鲁棒性、机载实时推理与动态环境安全控制等瓶颈仍待攻克。未来，随着三维世界模型、多模态大模型、神经渲染与边缘计算的深度融合，无人机视觉语言导航将朝着更强泛化性、更高安全性、更优人机协同与集群智能方向发展，深度融入低空经济与空天地一体化智能体系，最终实现人机自然对话、无人机自主执行的空中具身智能新生态，为低空智能装备的自主化、通用化与产业化奠定坚实基础。

## 参考文献(References)

Abderehman M, Patidar J, Oza J, Nigam Y, Khader TMA and Karfa C. 2022. FastSim: A Fast Simulation Framework for High-Level Synthesis. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41: 1371-1385 [DOI: 10.1109/TCAD.2021.3090339]

Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances[EB/OL]. [2026-04-09]. https://arxiv.org/pdf/2204.01691.pdf

Alpiste I, Golcarenarenji G, Wang Q and Alcaraz Calero J M. 2021. Search and rescue operation using UAVs: a case study. Expert Systems with Applications, 178: 114937 [DOI: 10.1016/j.eswa.2021.114937]

Anderson P, Wu Q, Teney D, Bruce J, Johnson M and Sünderhauf N, et al. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE. 2018: 3674-3683 [DOI: 10.1109/CVPR.2018.00387]

Anderson P, Wu Q, Teney D, Bruce J, Johnson M and Sünderhauf N, et al. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 3674-3683 [DOI: 10.1109/CVPR.2018.00387]

Awasthi S, Gramse N, Reining C and Roidl M. 2023. UAVs for Industries and Supply Chain Management [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2212.03346.pdf

Bejiga M B, Zeggada A and Melgani F. 2016. Convolutional neural networks for near real-time object detection from UAV imagery in avalanche search and rescue operations. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE: 4183-4186 [DOI: 10.1109/IGARSS.2016.7729174]

Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Frey N, et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control[EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2307.15818

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, et al. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics, 32(6): 1309-1332 [DOI: 10.1109/TRO.2016.2624754]

Cai H X, Dong J H, Tan J J, Deng J C, Li S H, Gao Z F, et al. 2025. FlightGPT: Towards Generalizable and Interpretable UAV Vision-and-Language Navigation with Vision-Language Models//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL: 6659 - 6676 [DOI: 10.18653/v1/2025.emnlp-main.338]

Carpin S, Lewis M, Wang J, Bagnell J A, Moore K and Stentz A. 2007. Usarsim: A robot simulator for research and education//Proceedings of the IEEE International Conference on Robotics and Automation. Rome: IEEE, 2007: 1400-1405 [DOI: 10.1109/ROBOT.2007.363204]

Chen P Y, Ouyang T, Luo K, Hong W J, Chen X. 2025. CoDrone: Autonomous Drone Navigation Assisted by Edge and Cloud Foundation Models[EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2512.19083

Chen Z, Li J Y, Fukumoto F, Liu P and Suzuki Y. 2024. Vision-Language Navigation for Quadcopters with Conditional Transformer and Prompt-based Text Rephraser//Proceedings of the 5th ACM International Conference on Multimedia in Asia. Tainan: ACM: 1-7 [DOI: 10.1145/3595916.3626450]

Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et

al. 2023. PaLM-E: an embodied multimodal language model [EB/OL]. [2026-04-09]. https://arxiv.org/abs/2303.03378

Du Q W, Dong W Z, Su W and Wang Q. 2022. UAV Inspection Technology and Application of Transmission Line. In: 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE). Dalian: IEEE: 1 - 5 [DOI: 10.1109/ICISCAE55891.2022.9927674]

Duan H B, Shi F R, Gao B, Zhou Y Y and Cui Q S. 2025. A novel real-time intelligent detector for monitoring UAVs in live-line operation on 10 kV distribution networks. Intelligence & Robotics, 5(1): 70-87 [DOI: 10.20517/ir.2025.05]

Echeverria G, Lemaignan S, Degroote A, Berger T, Escande A and Kheddar A. 2011. Modular open robots simulation engine: Morse// Proceedings of the IEEE International Conference on Robotics and Automation. Shanghai: IEEE, 2011: 46-51 [DOI: 10.1109/ICRA.2011.5980355]

Fan Y, Chen W, Jiang T, Liu H, Li J, Zhang L, et al. 2023. Aerial vision-and-dialog navigation. Findings of the Association for Computational Linguistics, 2023: 3043-3061 [DOI: 10.18653/v1/2023.findings-acl.190]

Fried D, Hu R, Cirik V, Rohrbach A, Andreas J and Morency L-P. 2018. Speaker-follower models for vision-and-language navigation [EB/OL]. [2026-04-09]. https://arxiv.org/pdf/1806.02724

Gao C, Zhao B, Zhang W, Mao J, Zhang J, Zheng Z, et al. 2024. EmbodiedCity: A benchmark platform for embodied agent in real-world city environment[EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2410.09604.pdf

Gao Y, Li C, You Z, Liu J, Li Z and Chen P, et al. 2025. OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2502.18041

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Salama A, Thau D, et al. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202: 18-27 [DOI: 10.1016/j.rse.2017.06.031]

Gu J, Stefani E, Wu Q, Thomason J and Wang X E. 2022. Vision-and-language navigation: a survey of tasks, methods, and future directions//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: Association for Computational Linguistics: 7606-7623 [DOI: 10.18653/v1/2022. acl-long.524]

Guhur P-L, Tapaswi M, Chen S, Laptev I and Schmid C. 2021. Airbert: In-domain Pretraining for Vision-and-Language Navigation// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montréal: IEEE/CVF: 1634-1643 [DOI: 10.1109/ICCV48922.2021.00166]

Guo J Y, Chen Z Y, Li Z W, Wang H, Liu Y and Zhang L, et al. 2026. HUGE-Bench: A Benchmark for High-Level UAV Vision-Language-Action Tasks[EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2603.19822

Gupta H, Guo X F, Ha H, Pan C E, Cao M Q, Lee D J, et al. 2025. UMI-on-Air: Embodiment-Aware Guidance for Embodiment-Agnostic Visuomotor Policies [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2510.02614.pdf

Ha D and Schmidhuber J. 2018. World Models[EB/OL]. [2026-05-01]. https://arxiv.org/pdf/1803.10122

Herron E, Lee X Y, Sin G, Gonzalez Diaz T, Farahat A and Gupta C, et al. 2025. A HIERARCHICAL AGENTIC FRAMEWORK FOR AUTONOMOUS DRONE-BASED VISUAL INSPECTION [EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2510.00259

Hong H D, Wang S, Huang Z, Wu Q and Liu J J. 2024. Why Only Text: Empowering Vision-and-Language Navigation with Multimodal Prompts//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Ljubljana: IJCAI: 839-847 [DOI: 10.24963/ijcai.2024/93]

Hong Y C, Rodriguez-Opazo C, Qi Y K, Wu Q and Gould S. 2020. Language and Visual Entity Relationship Graph for Agent Navigation [EB/OL]. [2026-04-09]. https://arxiv.org/pdf/2010.09304.pdf

Hong Y, Qi Y, Rodriguez-Opazo C, Wu Q and Gould S. 2021. VLN-BERT: a recurrent BERT architecture for vision and language navigation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE/CVF: 1643-1653 [DOI: 10.1109/CVPR46437.2021.00169]

Iversen N, Schofield O B and Ebeid E. 2020. LOCATOR - Lightweight and Low-Cost Autonomous Drone System for Overhead Cable Detection and Soft Grasping. In: 2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). IEEE: 1-6 [DOI: 10.1109/SSRR50563.2020.9292591]

Ji Y, He B, Tan Z, Liu Y, Chen F and Li J, et al. 2025. Game4loc: A uav geo-localization benchmark from game data//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia: AAAI Press, 39: 3913-3921 [DOI: 10.1609/aaai.v39i4.32409]

Jo D and Kwon Y. 2017. Development of Rescue Material Transport UAV (Unmanned Aerial Vehicle). World Journal of Engineering and Technology, 5(4): 720-729 [DOI: 10.4236/wjet.2017.54060]

Kerkech M, Hafiane A and Canals R. 2020. Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. Computers and Electronics in Agriculture, 179: 105446 [DOI: 10.1016/j. compag. 2020. 105446]

Koenig N and Howard A. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Sendai: IEEE, 2004: 2149-2154 [DOI: 10.1109/IROS. 2004.

1389727]

Krantz A, Gokaslan D, Batra D, Lee S and Maksymets O. 2020. Beyond the nav-graph: vision-and-language navigation in continuous environments//Proceedings of the European Conference on Computer Vision. Glasgow: Springer: 109-125 [DOI: 10.1007/9788-3-030-58604-1_7]

Lam D, Kuzma R, McGee K, Dooley S, Laielli M, Klaric M, et al. 2018. xView: Objects in context in overhead imagery [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/1802.07856.pdf

Lee J, Miyanishi T, Kurita S, Sakamoto K, Azuma D, Matsuo Y, et al. 2024. CityNav: A Large-Scale Dataset for Real-World Aerial Navigation[EB/OL].[2026-04-10]. https://arxiv.org/pdf/2406.14240.pdf

Lei Y J, Xu K, Guo Y L, Yang X, Wu Y W, Hu W, et al. 2024. Comprehensive survey on 3D visual-language understanding techniques. Journal of Image and Graphics, 29(6): 1747-1764 [DOI: 10.11834/jig.240029]

Li T S, Huai T Y, Li Z, Gao Y C, Li H A and Zheng X H, et al. 2025. SkyVLN: Vision-and-Language Navigation and NMPC Control for UAVs in Urban Environments[EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2507.06564

Li W, Li J, Wang Z, Chen H, Zhang H, Liu Y. 2026. SpatialFly: Geometry-Guided Representation Alignment for UAV Vision-and-Language Navigation in Urban Environments [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2603.21046.pdf

Li X, Tupayachi J, Sharmin A and Martinez Ferguson M. 2023. Drone-Aided Delivery Methods, Challenge, and the Future: A Methodological Review. Drones, 7: 191 [DOI: 10.3390/drones7030191]

Lin P C, Sun G, Liu C X, Li F Z, Ren W H and Cong Y. 2025. OpenVLN: Open-world Aerial Vision-Language Navigation [EB/OL]. [2026-04-10]. https://arxiv.org/abs/2511.06182.pdf

Liu S, Zhang H, Qi Y, Wang Y, Zhang X, Zhao J, et al. 2023. AerialVLN: Vision-and-language navigation for UAVs[EB/OL].[2026-04-10]. https://arxiv.org/pdf/2308.06735.pdf

Liu W S, Li W, Zhou Q, Die Q and Yang Y. 2022. The optimization of the 'UAV-vehicle' joint delivery route considering mountainous cities. PLOS ONE, 17(3): e0265518 [DOI: 10.1371/journal.pone.0265518]

Liu X, Liu Y, Qiu H S, Yang Q R and Lian Z H. 2026. IndoorUAV: Benchmarking Vision-Language UAV Navigation in Continuous Indoor Environments[EB/OL].[2026-04-10]. https://arxiv.org/abs/2512.19024.pdf

Liu Y Z, Yao F L, Yue Y C, Xu G L, Sun X and Fu K. 2024. NavAgent: Multi-scale Urban Street View Fusion For UAV Embodied Vision-and-Language Navigation[EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2411.08579.pdf

Loquercio A, Kaufmann E, Ranftl R, Müller M, Koltun V and Scaramuzza D. 2021. Learning high-speed flight in the wild. Science Robotics, 6(59): 3-26 [DOI: 10.1126/scirobotics.abg5810]

Lykov A, Karaf S, Martynov M, Serpiva V, Fedoseev A, Konenkov M and Tsetserukou D. 2024. FlockGPT: Guiding UAV Flocking with Linguistic Orchestration[EB/OL].[2026-05-06]. https://arxiv.org/pdf/2405.05872

Michel O. 2004. Webots: Professional mobile robot simulation. International Journal of Advanced Robotic Systems, 1: 39-42 [DOI: 10.5772/5618]

Ning Y W, Zhao G L, Qin Y P, Liu S, Liu Y, Lin L and Li G B. 2026. LookasideVLN: Direction-Aware Aerial Vision-and-Language Navigation[EB/OL].[2026-05-06]. https://arxiv.org/pdf/2604.17190

Panerati J, Schoellig A P, Watterson M, Turpin M, Fitch R and Barfoot T D. 2021. Learning to fly—a gym environment with pybullet physics for reinforcement learning of multi-agent quadcopter control// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021: 7512-7519 [DOI: 10.1109/IROS51168.2021.9636161]

Pardyl A, Matuszek D, Przebieracz M, Cygan M, Zieliński B, Wolczyk M. 2025. FlySearch: Exploring how vision-language models explore [EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2506.02896

Rajagopal M K and Murugan M S B. 2023. Artificial Intelligence based drone for early disease detection and precision pesticide management in cashew farming. [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2303.08556.pdf

Sanyal S and Roy K. 2025. ASMA: An Adaptive Safety Margin Algorithm for Vision-Language Drone Navigation via Scene-Aware Control Barrier Functions[EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2409.10283

Sarker G C, Azad A K M, Rahman S, Hasan M M. 2026. VLM-Nav: Mapless UAV navigation using monocular vision driven by vision-language models. PLOS ONE, 21(5): e0345778. DOI: 10.1371/journal.pone.0345778

Sautenkov O, Akhmetkazy A, Yaqoot Y, Mustafa M A, Tadevosyan G, Lykov A, et al. 2025. UAV-VLPA: Vision-Language Guided Global-Local UAV Mission Planning from Satellite Imagery//Proceedings of the 2025 IEEE International Conference on Robotics and Biomimetics (ROBIO 2025). Zhangjiajie, China: IEEE: 2354-2359 [DOI: 10.1109/ROBIO66223.2025.11377338]

Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B, et al. 2019. Habitat: A platform for embodied AI research//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 9339-9347 [DOI: 10.1109/ICCV.2019.00939]

Saxena P, Raghuvanshi N and Goveas N. 2025. UAV-VLN: End-to-End Vision Language guided Navigation for UAVs [EB/OL]. [2026-

04-10].
https://arxiv.org/pdf/2504.21432.pdf

Shah D, Osinki B, Ichter B and Levine S. 2022. LM-Nav: robotic navigation with large pre-trained models of language, vision, and action [EB/OL]. [2026-04-09].
https://arxiv.org/pdf/2207.04429.pdf

Shah S, Dey D, Lovett C, Kapoor A, Scherer S and Kumar V. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. Field and Service Robotics, 2018: 621-635 [DOI: 10.1007/978-3-319-67361-5_40]

Shao D, Xu Z Z, Wang P Y, Liu L, Wang Y, Shi J Q, et al. 2026. FineCog-Nav: Integrating Fine-grained Cognitive Modules for Zero-shot Multimodal UAV Navigation[EB/OL].[2026-05-06].
https://arxiv.org/pdf/2604.16298.

Shendryk Y, Sofonia J, Garrard R, Rist Y, Skocaj D and Thorburn P. 2020. Fine-scale prediction of biomass and leaf nitrogen content in sugarcane using UAV LiDAR and multispectral imaging. International Journal of Applied Earth Observation and Geoinformation, 91: 102177[DOI: 10.1016/j.jag.2020.102177]

Singla A, Padakandla S and Bhatnagar S. 2020. Memory-based Deep Reinforcement Learning for Obstacle Avoidance in UAV with Limited Environment Knowledge. IEEE Transactions on Intelligent Transportation Systems, 21(10): 4293 - 4304 [DOI: 10.1109/TITS.2019.2954952]

Sun X L, Ni W H, Li Y T, Wu D M, Xie F, Guan R W, et al. 2026. AutoFly: Vision-Language-Action Model for UAV Autonomous Navigation in the Wild[EB/OL].[2026-04-10].
https://arxiv.org/pdf/2602.09657.pdf

Tang Y, Ma J W, Zhang J R, Wang A J, Zhang D Y. 2026. Mitigating Error Accumulation in Continuous Navigation via Memory-Augmented Kalman Filtering[EB/OL]. [2026-04-11].
https://arxiv.org/pdf/2602.11183.pdf

Tian Y L, Lin F, Zhang X Y, Ge J W, Wang Y T and Dai X Y. 2024. LogisticsVISTA: 3D Terminal Delivery Services With UAVs, UGVs and USVs Based on Foundation Models and Scenarios Engineering[C]//2024 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). IEEE, Macau, China, 2024. [DOI:10.1109/SOLI63266.2024.10956119]

Tong H Y, Dong X Y, Ma X G, Zhao H R, Zhou Y M and Lin C H. 2026. ViSA-Enhanced Aerial VLN: A Visual-Spatial Reasoning Enhanced Framework for Aerial Vision-Language Navigation [EB/OL].[2026-04-10].
https://arxiv.org/pdf/2603.08007.pdf

Vemprala S, Bonatti R, Bucker A and Kapoor A. 2023. ChatGPT for Robotics: Design Principles and Model Abilities[EB/OL].[2026-04-10].
https://arxiv.org/pdf/2306.17582.pdf

Vyas V, Schuck M, Dahanaggamaarachchi D O, Zhou S, Zhang L and Schoellig A P. 2024. SwarmGPT-Primitive: A Language-Driven Choreographer for Drone Swarms Using Safe Motion Primitive Composition[EB/OL].[2026-05-06].
https://arxiv.org/pdf/2412.08428

Wang H. 2024. Grutopia: Dream general robots in a city at scale[EB/OL].[2026-04-10].
https://arxiv.org/pdf/2407.10943.pdf

Wang X Y, Yang D L, Liao Y, Zheng W H, Dai B, Wu W J, et al. 2025. UAV-Flow Colosseo: A Real-World Benchmark for Flying-on-a-Word UAV Imitation Learning [EB/OL].[2026-04-10].
https://arxiv.org/pdf/2505.15725.pdf

Wang X Y, Yang D L, Wang Z Q, et al. 2024. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology[EB/OL].[2026-04-10].
https://arxiv.org/pdf/2410.07087.pdf

Wang X, Huang Q, Celikyilmaz A, Gao J, Shen D, Wang Y F, et al. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE/CVF: 6639-6648 [DOI: 10.1109/CVPR.2019.00679]

Wang Z Y, Li R P, Li S Z, Xiang Y M, Wang H P and Zhao Z F. 2025. RALLY: Role-Adaptive LLM-Driven Yoked Navigation for Agentic UAV Swarms. IEEE Open Journal of Vehicular Technology, 2025, 6: 1-12. [DOI:10.1109/OJVT.2025.3610852]

Wolfe V, Frobe W, Shrinivasan V and Hsieh T Y. 2015. Detecting and locating cell phone signals from avalanche victims using unmanned aerial vehicles. In: 2015 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE: 1-8 [DOI: 10.1109/ICUAS.2015.7152353]

Wu J H, Yao F L, Liu Y Z, Zhang W Y, Zhu Z Q, Li C L, et al. 2025. AeroDuo: Aerial Duo for UAV-based Vision and Language Navigation//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin: Association for Computing Machinery: 2576-2585 [DOI: 10.1145/3746027.3754498]

Wu W, He H, He J, Wang Y, Duan C, Liu Z, Li Q and Zhou B. 2024. MetaUrban: An Embodied AI Simulation Platform for Urban Micromobility[EB/OL].[2026-04-10].
https://arxiv.org/pdf/2407.08725.pdf

Wu Y Z, Zhu M, Li X X, Du Y H, Fan Y X, Li W J, et al. 2025. VLA-AN: An Efficient and Onboard Vision-Language-Action Framework for Aerial Navigation in Complex Environments [EB/OL]. [2026-04-11].
https://arxiv.org/pdf/2512.15258.pdf

Xiao J Q, Sun Y X, Shao Y X, Gan B X, Liu R Q, Wu Y J, Guan W L and Deng X. 2025. UAV-ON: A Benchmark for Open-World Object Goal Navigation with Aerial Agents//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin: Association for Computing Machinery: 1-7 [DOI: 10.1145/3746027.3758251]

Xu H T, Hu Y, Gao C, Zhu Z Q, Zhao Y and Yin Q J. 2026. GeoNav:

Empowering MLLMs with Dual-Scale Geospatial Reasoning for Language-Goal Aerial Navigation. Pattern Recognition, 177: 113366 [DOI: 10.1016/j.patcog.2026.113365]

Xu P, Deng Z N, Deng J Y, Gu Z H and Wan S H. 2026. AerialVLA: A Vision-Language-Action Model for UAV Navigation via Minimalist End-to-End Control [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2603.14363.pdf

Xu Z F, Han X M, Shen H Y, Jin H Y and Shimada K. 2025. NavRL: Learning Safe Flight in Dynamic Environments. IEEE Robotics and Automation Letters, 10(4): 3668-3675 [DOI: 10.1109/LRA.2025.3546069]

Yao F L, Yue Y C, Liu Y Z, Sun X, Fu K. 2024. AeroVerse: UAV-Agent Benchmark Suite for Simulating, Pre-training, Finetuning, and Evaluating Aerospace Embodied World Models [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2408.15511.pdf

Yasheer A, Mustafa A, Sautenkov O, Lykov A, Serpiva V, Tsetserukou D. 2025. UAV-VLRR: Vision-Language Informed NMPC for Rapid Response in UAV Search and Rescue [EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2503.02465

Ye J L, Papaioannou S and Kolios P. 2025. VLM-RRT: Vision Language Model Guided RRT Search for Autonomous UAV Navigation [EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2505.23267

Yong S P and Yeong Y C. 2018. Human Object Detection in Forest with Deep Learning based on Drone's Vision. In: 2018 4th International Conference on Computer and Information Sciences (ICCOINS). IEEE: 1-5 [DOI: 10.1109/ICCOINS.2018.8510564]

Zha J M, Yang T and Mueller M W. 2024. Agri-fly: simulator for uncrewed aerial vehicle flight in agricultural environments [J]. IEEE Access, 2024, 12: 140900-140907. [DOI: 10.1109/ACCESS.2024.3467335]

Zhang C H, Huang G J, Liu L, Huang S, Yang Y N, Wan X, et al. 2023. WebUAV-3M: A Benchmark for Unveiling the Power of Million-Scale Deep UAV Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 9186-9205 [DOI: 10.1109/TPAMI.2022.3232854]

Zhang D X, Chen P, Xia X B, Su X, Zhen R C, Xiao J Q, et al. 2026. APEX: A Decoupled Memory-based Explorer for Asynchronous Aerial Object Goal Navigation [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2602.00551.pdf

Zhang D X, Chen P, Xia X B, Su X, Zhen R C, Xiao J Q, et al. 2026. APEX: A Decoupled Memory-based Explorer for Asynchronous Aerial Object Goal Navigation [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2602.00551.pdf

Zhang J Z, Wang K Y, Xu R T, Zhou G Z, Hong Y C, Fang X M, et al. 2024. NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2402.15852.pdf

Zhang L F, Zhang Y C, Li H S, Fu H X, Tang Y B, Ye H J, et al. 2025. Is your VLM Sky-Ready? A Comprehensive Spatial Intelligence Benchmark for UAV Navigation [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2511.13269.pdf

Zhang Q Y, Zheng S H, Sun J L, Li C X, Wu X K, Song Z H, et al. 2026. UAV-Track VLA: Embodied Aerial Tracking via Vision-Language-Action Models [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2604.02241.pdf

Zhang W C, Gao C, Yu S Q, Peng R Y, Zhao B N, Zhang Q, et al. 2025. CityNavAgent: Aerial Vision-and-Language Navigation with Hierarchical Semantic Planning and Global Memory. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31292 - 31309. [DOI: 10.18653/v1/2025.acl-long.1511]

Zhang X Y, Tian Y L, Lin F, Liu Y, Ma J, Szatmary K S and Wang F Y. 2025. LogisticsVLN: Vision-Language Navigation For Low-Altitude Terminal Delivery Based on Agentic UAVs [EB/OL]. [2026-05-01]. https://arxiv.org/pdf/2505.03460

Zhang Y H, Yu H S, Xiao J P and Feroskhan M. 2025. Grounded Vision-Language Navigation for UAVs with Open-Vocabulary Goal Understanding [EB/OL]. [2026-05-06]. https://arxiv.org/pdf/2506.10756.pdf

Zhang Y N, Wang H Y, Yan Q S, Yang J Q, Liu T, Fu M Q, et al. 2025. Research progress of unmanned mobile vision technology for complex dynamic scenes. Journal of Image and Graphics, 30(6): 1828-1871 (张艳宁,王昊宇,闫庆森,杨佳琪,刘婷,符梦芹,等. 2025. 面向复杂动态场景的无人移动视觉技术研究进展。中国图象图形学报,30(6): 1828-1871) [DOI:10.11834/jig.240458]

Zhao B N, Fang J J, Dai Z C, Wang Z Y, Zha J R, Zhang W C, et al. 2025. UrbanVideo-Bench: Benchmarking Vision-Language Models on Embodied Intelligence with Video Data in Urban Spaces//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics: 32400-32423 [DOI: 10.18653/v1/2025.acl-long.1558]

Zhao B N, Tang R Z, Jia M Y, Wang Z Y, Man F H, Zhang X, et al. 2025. AirScape: An Aerial Generative World Model with Motion Controllability [EB/OL]. [2026-04-10]. https://arxiv.org/pdf/2507.08885.pdf

Zhao G L, Li G B, Pan J and Yu Y Z. 2025. Aerial Vision-and-Language Navigation with Grid-based View Selection and Map Construction [EB/OL]. [2026-04-11]. https://arxiv.org/pdf/2503.11091.pdf

Zheng G Y, Ban Y T, Zhang M J, Zheng J P and Zhou B Y. 2026. OnFly: Onboard Zero-Shot Aerial Vision-Language Navigation toward Safety and Efficiency [EB/OL]. [2026-05-06]. https://arxiv.org/pdf/2603.10682.pdf

Zhong F. 2024. Unrealzoo: Enriching photo-realistic virtual worlds for

embodied ai [EB/OL].[2026-04-10]. https://arxiv.org/pdf/2412.20977.pdf

Zhou J X, Wang S B, Yang Z Y, Yu Z J and Li T. 2026. FreeFly-Thinking: Aligning Chain-of-Thought Reasoning with Continuous UAV Navigation[EB/OL].[2026-04-10]. https://arxiv.org/pdf/2603.07181

Zhou Y C, Lao C C, Yang Y L, Zhang Z T, Chen H Y, Chen Y W, et al. 2021. Diagnosis of winter-wheat water stress based on UAV-borne multispectral image texture and vegetation indices. Agricultural Water Management, 256: 107076 (雷印杰,徐凯,郭裕兰,杨鑫,武玉伟,胡玮,杨佳琪,汪汉云. 2024. “三维视觉 — 语言”推理技术的前沿研究与最新趋势。中国图象图形学报,29(6): 1747-1764)[DOI: 10.1016/j.agwat.2021.107076]

## 作者简介

郑周一,男,博士研究生,研究方向为计算机视觉。E-mail: zhengzhouyi@mail.nwpu.edu.cn

郭宸瑞,男,工程师,研究方向为目标检测、多模态大语言模型、特征反演、雷达和光学系统设计。E-mail: g-ch-r@163.com

单淳,男,本科生,研究方向为视觉语言导航。E-mail: shanchun@mail.nwpu.edu.cn

张磊,通信作者,男,教授,主要研究方向为计算机视觉。E-mail: nwpuzhanglei@nwpu.edu.cn

魏巍,男,教授,主要研究方向为计算机视觉。E-mail: weiweinwpu@nwpu.edu.cn