中图法分类号:TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3169-29

论文引用格式:Yu Y T, Cao C Q, Wang Z Y and Zhang Y N. 2026. Recent advances in large multimodal models for UAV visual understanding. Journal of Image and Graphics, 31(9):3169-3197(余雅婷, 曹聪琦, 王昭颖, 张艳宁. 2026. 面向无人机的多模态视觉理解大模型研究进展. 中国图象图形学报, 31(9):3169-3197)[DOI:10.11834/jig.260215]

面向无人机的多模态视觉理解大模型研究进展

余雅婷,曹聪琦*,王昭颖,张艳宁

西北工业大学计算机学院,西安　710129

摘　要:随着低空经济和智能无人系统的发展,无人机逐渐从传统的遥控飞行平台演化为集环境感知、语义理解与自主决策于一体的空中智能体。以视觉基础模型、视觉语言模型和多模态大模型为代表的视觉理解大模型,为无人机在复杂开放环境中的感知、理解与决策提供了新的技术范式。围绕无人机视觉理解能力的演进逻辑,本文构建了基础感知—语义推理—决策规划的3层能力分析框架,并以此为主线系统梳理无人机场景视觉理解大模型的研究进展。在任务层面,依据该能力框架构建无人机视觉理解任务体系,归纳了基础目标感知、事件语义分析、空间环境理解与飞行决策规划等典型任务,并分析航拍视觉在尺度变化、远距离观测与复杂动态环境中的关键挑战;在技术层面,沿着同一能力演进逻辑,系统回顾视觉理解方法从传统深度学习算法与视觉基础模型的感知建模,发展到大语言模型与多模态大模型的语义推理与跨模态交互,再到具身视觉—语言—行动模型的智能决策与任务规划的技术演进路径。在此基础上,重点综述视觉理解大模型在无人机视觉感知增强、视觉语义推理以及视觉决策规划3个核心能力维度的研究进展,并分析其在开放词汇感知、跨模态推理、复杂空间关系理解与具身智能决策等方面带来的关键能力提升。同时,对当前无人机视觉理解领域的主流数据集与评测基准进行了系统总结,并分析当前评测体系正由传统任务导向逐步向能力导向评估范式演进。最后,针对无人机平台资源受限、实时推理需求以及系统安全可靠性等问题,对视觉理解大模型在无人机领域未来的发展方向进行了展望。

关键词:多模态大模型(MLLM);视觉基础模型(VFM);无人机(UAV);视觉理解;智能决策;综述

Recent advances in large multimodal models for UAV visual understanding

Yu Yating, Cao Congqi*, Wang Zhaoying, Zhang Yanning

School of Computer Science, Northwestern Polytechnical University, Xi'an　710129, China

Abstract: With the rapid expansion of the low-altitude economy, accompanied by the increasing maturity of intelligent unmanned systems, unmanned aerial vehicles (UAVs) are gradually transforming from traditional remotely controlled flying platforms into autonomous aerial agents that are fully capable of perception, reasoning, and decision-making. Among the various sensing modalities available to UAVs, visual perception plays a central role in acquiring environmental information and enabling high-level situational awareness. Hence, visual understanding capability can directly determine the intelligence level and operational autonomy of UAV systems. In recent years, the emergence of visual foundation models, vision-language models, and multimodal large language models (LLMs) has significantly influenced the technical paradigm of UAV visual understanding. As a result, this paradigm shift has provided new opportunities for enabling UAV systems to

收稿日期:2026-04-16;**修回日期**:2026-05-20;**预印本日期**:2026-05-27

*通信作者:曹聪琦　congqi.cao@nwpu.edu.cn

基金项目:国家自然科学基金项目(62376217,62576279,62301434);中国科协青年人才托举工程项目(2023QNRC001)

Supported by: National Natural Science Foundation of China (62376217, 62576279, 62301434); Young Elite Scientists Sponsorship Program by CAST(2023QNRC001)

operate effectively in complex and open environments in which perception, reasoning, and decision making must be tightly coupled. To clarify this emerging research landscape, this paper introduces a capability-oriented analytical framework that organizes UAV visual understanding into three hierarchical levels: basic perception, semantic reasoning, and decision planning. This framework serves as the conceptual backbone of the survey and allows recent studies to be examined from a systematic capability-evolution standpoint. Specifically, from the task perspective, we construct a comprehensive taxonomy of UAV visual understanding tasks that includes four major categories: 1) basic object perception, 2) event semantic analysis, 3) spatial environment understanding, and 4) flight decision-making. Within this taxonomy, representative tasks (i. e., object detection, target tracking, human action recognition, visual question answering, spatial reasoning, navigation, and autonomous flight control) are analyzed in a unified manner. At the same time, this paper summarizes several fundamental challenges that arise in aerial visual perception. These challenges include significant scale variations caused by high-altitude viewpoints, long-range observations that lead to small object representations, complex backgrounds with strong visual clutter, and dynamic environmental changes that require robust temporal reasoning. From the technical perspective, this paper reviews the methodological evolution that underlies the development of UAV visual understanding models. Early approaches were largely based on conventional deep learning architectures that focused on supervised visual perception tasks. Then, subsequent advances in visual foundation models significantly improved representation learning by leveraging large-scale pretraining and open-vocabulary multimodal alignment. More recently, LLMs and multimodal LLMs have introduced powerful reasoning capabilities and cross-modal interaction mechanisms that allow UAV systems to interpret visual observations along with natural language instructions and contextual knowledge. Building upon these developments, embodied vision-language-action models have begun to connect perceptions with action generation, thus enabling UAVs to perform complex task planning and interactive decision making in real-world environments. From the capability perspective, this paper further examines how large visual understanding models enhance UAV intelligence across three dimensions: 1) visual perception enhancement, where large models improve robustness in open-vocabulary recognition, small-object detection, and fine-grained visual understanding; 2) vision-language reasoning, where multimodal models facilitate complex reasoning processes such as spatial relation reasoning, event interpretation, and cross-modal knowledge integration; and 3) visual decision planning, where embodied multimodal models enable UAVs to translate perception and reasoning outcomes into actionable flight strategies, mission planning procedures, and adaptive control policies. Additionally, representative UAV visual datasets and benchmarks that support the development and assessment of large multimodal models for visual understanding are summarized in this work. Particular attention is given to the evolution of evaluation protocols. Traditional benchmarks often measure performance in narrowly defined tasks, such as detection or classification. However, recent research has increasingly emphasized capability-oriented evaluation frameworks that assess broader competencies, namely reasoning ability, cross-task generalization, and decision support. This transition reflects a broader shift in the field toward evaluating integrated visual intelligence rather than isolated perception performance. Finally, this work discusses several promising research directions that may shape the future development of UAV visual understanding systems. These include the construction of general-purpose visual foundation models tailored for aerial scenarios, the advancement of embodied UAV intelligence through vision-language-action integration, the development of real-time reasoning techniques and lightweight deployment strategies suitable for resource-constrained aerial platforms, and the establishment of safety-aware, trustworthy, and privacy-preserving UAV perception systems.

Key words: multimodal large language model (MLLM); visual foundation model (VFM); unmanned aerial vehicle (UAV); visual understanding; intelligent decision-making; review

0 引 言

随着低空经济的蓬勃发展，无人机(unmanned aerial vehicle, UAV)技术在过去10年中经历了爆发式发展。在此背景与技术驱动之下，无人机凭借其卓越的灵活空间机动性与全方位视野，已由单纯的遥控飞行器演变为集感知、决策与执行于一体的空中智能体，已深度渗透至智慧交通、应急救援、精准农业及电力巡检等多元领域，成为现代智能城市管

理的核心底座(晏磊 等,2019)。在非结构化环境的作业任务中,视觉感知系统作为无人机系统的眼睛,是其获取外界信息最直观、最丰富的手段,更是实现自主化的首要前提。随着应用场景向复杂动态环境深度延伸,视觉理解能力已成为无人机实现高度智能化的动力源泉。然而,尽管深度学习在过去10年推动了视觉任务的飞速进展,无人机视觉系统在面对复杂真实场景以及多样化任务需求时仍面临严峻挑战,这促使无人机视觉理解范式发生深刻演变(Joshi等,2025;Phadke等,2024)。

无人机场景视觉理解的独特性对算法提出了极其苛刻的要求。高空俯瞰视角下的航拍图像普遍面临目标尺度剧烈变化、空间分布密集以及背景干扰严重等问题(冷佳旭 等,2023)。在长距离感知中,目标往往仅占据数个像素,传统检测器难以捕捉其关键语义特征。这种成像过程不仅容易受到大气湍流、雾霾或雨雪等恶劣天气的影响,还伴随着复杂的背景干扰和动态场景变化。为解决这些挑战,早期研究多依赖于基于规则的传统计算机视觉算法,通过人工设计的特征算子完成特定目标的识别,但在非结构化环境下的泛化性能极差。随着深度学习架构的崛起,无人机在目标检测、语义分割等基础视觉任务上取得了跨越式的感知能力。然而,此类研究范式集中于通过全监督学习训练无人机特定任务专用的视觉模型,不仅依赖大规模任务标注数据,而且通常只能处理预定义的闭集类别,一旦面对分布外目标或复杂的跨域场景,极易因缺乏语义先验而失效(Limberg等,2024;Zhang等,2024a)。

无人机场景视觉理解的多样化任务需求进一步放大了任务专用视觉模型的局限性。随着人机协作需求的日益增长,传统的边界框式感知已无法满足复杂的交互需求,系统迫切需要理解人类自然语言指令背后的深层语义(Li等,2024)。此外,无人机在执行搜救或监测任务时,往往需要在非结构化环境中处理具有高度不确定性的地理空间关系,这对视觉系统的三维空间推理能力提出了严峻考验(Gao等,2025b)。

大语言模型(large language model, LLM)与多模态大模型(multimodal large language model, MLLM)的涌现,为无人机视觉理解能力的多维度升级提供了全新的破局思路。这种范式变化的第一个核心体现在于从闭集识别向开放词汇感知的跨越。传统的任务专用视觉模型受限于训练集的标签空间,而视觉基础模型(vision foundation model, VFM),通过海量图文对的预训练,实现对任意自然语言描述目标的零样本检测(Kim等,2024a)与分割(Ma等,2024;Sezgin等,2025),赋予了无人机开放世界视觉泛化的能力。这意味着无人机无需针对每个新任务进行微调,即可根据人类口头描述识别并追踪陌生目标,极大降低了智能系统的部署门槛。

范式变迁进一步体现在从无人机基础视觉感知向语义综合推理能力的升级。大语言模型与多模态基座的融合,赋予了无人机对事件级的深层次理解能力与任务级的人机交互能力。De Zarzà 等人(2023)研究表明,通过将视频流转化为语义日志或结构化描述,无人机能够像人类一样对飞行行为进行解释、对异常事件进行逻辑推理,甚至在缺乏预制地图的情况下,通过常识知识推断房间的搜索序列。这种感知到认知能力的升级,不仅提升了系统在复杂任务中的自主性,还极大增强了人机交互的直观性。

更为深远的变革在于从被动观测向主动决策的进化。在具身无人机智能系统中,视觉信息不再仅仅是预测结果,而是影响飞行决策的关键指令。传统无人机系统在面对复杂指令时,往往需要多模块烦琐重组,通过视觉—语言—行动(vision language action, VLA)的一体化架构,实现了从高层语言指令到低层飞行控制的直接映射,有效缓解了传统控制链路中语义流失的问题。例如,在避障与路径规划任务中,大模型能够凭借其丰富的物理先验知识,在动态环境中推断安全缓冲距离,并实时调整飞行轨迹以规避潜在风险,展现出类似人类专家的综合决策水平(Cai等,2025b)。面向大脑的审慎规划与小脑的快速反应相结合的架构,正在重构无人机的作业模式(Zhao等,2023;Koubaa和Gabr,2025)。

视觉理解能力的这种代际升级,对于无人机系统而言具有决定性意义。在资源受限的边缘计算平台上,如何高效部署具有强大感知能力的轻量化模型,已成为当前工业界与学术界共同关注的焦点(Wang等,2025)。与此同时,随着大规模航拍视觉语言数据集及物理真实仿真平台的不断涌现(Yao等,2025a;Ferrag等,2025),视觉大模型在无人机领域的评测体系也从单一的算法精度演进为涵盖空间智能、伦理推理及鲁棒性的综合评估(Sautenkov等,

2025a)。

尽管大模型赋能无人机系统展现出巨大潜力，但如何将这些体量巨大、计算昂贵的模型高效集成到资源受限的无人机平台，仍是当前学术界与工业界亟待攻克的难题(Ahmmad 等，2025；Chen 等，2026)。此外，大模型固有的幻觉问题在安全敏感的航空任务中具有潜在风险，也为该领域的研究带来了新的不确定性。

与已有无人机特定应用领域或无人机大语言模型技术应用的综述不同，如图1所示，本文立足于无人机通用视觉能力，旨在系统梳理无人机场景下视觉理解大模型的研究进展，深入探讨视觉大模型如何从基础视觉感知、视觉语义推理以及视觉决策规划等多个维度赋能无人机通用视觉理解。首先，从任务层角度构建无人机视觉理解任务图谱；其次，深入解析从传统深度学习算法到多模态大模型的技术演进脉络；随后，重点分析大模型赋能下的无人机感知增强、语义推理与具身决策等核心能力；然后，总结当前主流的无人机视觉数据集与评测基准；最后，展望无人机视觉理解大模型面临的实时性、安全性与协同理解等未来挑战，旨在为下一代高度自主、透明可信的空中智能系统提供前瞻性的学术参考。

1 无人机视觉理解任务与挑战

无人机视觉理解的核心目标，是使飞行平台从

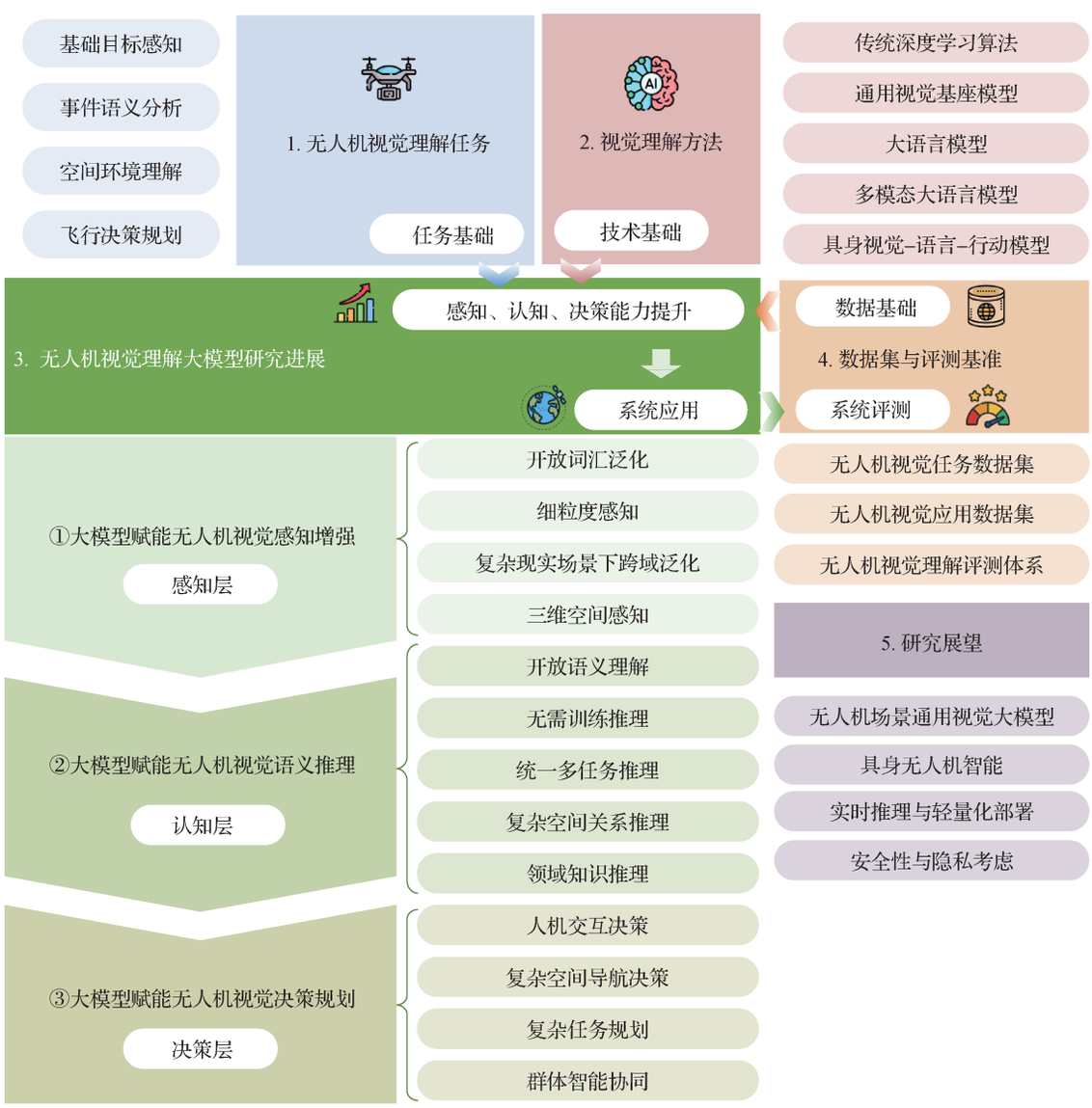


图1 本文整体框架

Fig. 1 Overall framework of the paper

机载图像或视频中获得对目标、场景、事件与空间关系的有效表征，并进一步支撑导航控制与任务执行。与地面视觉系统相比，无人机场景具有更强的视角变化、更大的尺度跨度、更远的观测距离以及更开放的环境分布，这使其视觉任务并非简单复用通用计算机视觉范式，而是在感知、理解和决策3个层面都呈现出鲜明的空域特征。

为了清晰呈现视觉理解大模型在这一领域的演进脉络，如表1所示，本研究构建了基于任务目标和处理层级的无人机视觉理解任务分类体系，并总结了不同任务中的挑战以及应用场景。具体可概括为4类：基础目标感知、语义理解与事件分析、空间理解与环境建模，以及飞行控制与决策规划。无人机视觉理解任务不仅要求从视觉中进行目标聚焦与空间推理，同时也进一步强调视觉中形成语义判断并驱动行动，共同构成从感知理解到智能决策的任务链条。

1.1 基础目标感知

基础目标感知是无人机视觉理解的起点，也是后续空间建模、语义推理与自主决策的前提。在本

表1 典型无人机视觉理解任务归纳总结

Table 1 Comprehensive overview of representative UAV visual understanding tasks

理解维度	任务类型	主要挑战	应用场景
基础目标感知	目标检测	小目标密集、尺度变化大、遮挡严重、远距离成像分辨率低以及恶劣天气干扰	交通监控、城市管理、工业巡检、灾难救援等
	目标分割、场景解析	复杂背景干扰、边界模糊、精确标注成本高、高分辨率影像处理复杂	地物识别、道路边界解析、农作物覆盖监测、设施巡检等
	目标跟踪	目标尺度变化大、相似目标易干扰、物体遮挡、消失重现、运动模糊	安防监控、交通流分析、目标轨迹追踪等
事件语义分析	行为识别	长时序依赖、动作细粒度差异、复杂交互关系	空地交互理解、人员动作分类、安全巡逻监控等
	视觉问答	视觉信息与语言语义对齐困难、复杂推理需求、领域知识受限	智能巡检问答、任务解释等
	意图识别	意图语义隐含性强，依赖上下文以及先验知识	低空安全防御、非合作无人机管控、态势感知等
	事故检测	事件稀缺、异常定义复杂、实时检测需求高	灾害应急响应、交通事故监测等
空间环境理解	三维空间推理	单目深度估计不确定、空间结构复杂、细粒度空间认知缺乏	三维建图、环境感知等
	地理定位	GPS信号拒止、现实场景复杂多变、视觉定位跨域泛化困难	野外搜救定位、无GPS环境导航、军事侦察等
飞行决策规划	视觉避障	动态障碍物轨迹预测不确定性，低延迟实时响应需求	自主飞行、复杂环境巡航等
	自主降落	动态非结构化环境中语义感知有限，着陆区域识别困难、复杂环境因素干扰	快递配送自动对接、充电平台对接、应急降落等
	视觉语言导航	语言指令的模糊性、仿真与现实场景的分布差距大、超长距离导航需求	智慧城市快递配送、复杂环境探索、物资自动投送
	目标搜索	搜索空间大、语义处理冗余、目标稀缺、相似物体歧义、路径优化复杂	野外搜救任务、巡检任务等
	任务规划	多任务协同复杂、资源调度困难	复杂任务执行调度等
	协同规划	多机间通信的语义不对称、任务分配不高效、路径优化的效率低	多无人机协作巡检等
	编队控制	编队稳定性与同步性与通信延迟高	无人机灯光秀表演、军事集群任务、空中协同侦察等

文的任务框架中，无人机基础目标感知主要涵盖目标检测、目标分割与场景解析、目标跟踪，以及与之密切相关的场景识别、变化检测和运动估计等任务。这一层级的核心目标，是从复杂航拍观测，即传感器捕获的原始多模态数据（如RGB、红外、遥感、深度图像）中提取具有语义信息的关键要素，例如目标在哪里、是什么、处于何种状态以及如何变化等信息，为更高层的环境理解和任务执行提供可靠视觉表征。就整体特征而言，高空视角导致目标在图像中占据的像素极少且观测尺度跨度大，导致环境背景噪声极易淹没目标特征。此外，在传统架构下，感知模块主要依赖预定义的固定类别识别，对模型的跨域、跨模态以及零样本泛化能力提出了严峻考验。在大模型范式下，感知任务正经历从封闭集识别向开放词汇感知与多模态语义理解的重大转变。

1.1.1　目标检测

目标检测是无人机视觉感知中最基础、最核心的任务之一，旨在确定航拍图像或视频中感兴趣目标的类别与精确位置。在实际应用，如交通监控、作物分析和灾难救援中，检测器需要处理各类关键设施目标，涵盖从大范围农作物、建筑物到微小行人、车辆等极宽目标范围。作为分割、跟踪、场景分析与事件理解的前置步骤，目标检测直接决定无人机系统对环境的初始认知质量。

1.1.2　目标分割与场景解析

目标分割（包括语义分割、实例分割及全景分割等）与场景解析旨在实现像素级的环境理解，进一步对目标边界的精确划分、场景的语义区域构成以及不同区域之间的结构关系等实现细粒度感知，是解析道路边界、农作物覆盖及建筑损毁的关键。然而，由于航空领域获取精确地面真值标注的成本极高，标注依赖成为制约分割任务发展的一大瓶颈。在无人机应用中，场景解析任务需要处理超高分辨率的航拍影像，分割任务不仅要求模型具备细粒度边界刻画能力，还要求其能够处理大幅面图像中的全局—局部语义耦合关系，这对特征提取的精细度提出了极高要求。而不同地域、气候导致的无人机航拍影像跨域差异大，使得模型在未知领域的泛化性能大幅下降。

1.1.3　目标跟踪

目标跟踪任务要求无人机在视频序列中对特定目标建立跨帧的关联，实现持续的动态监控，是连接静态感知与动态理解的重要桥梁。该任务不仅涵盖单一目标的单目标跟踪，还包括在拥挤场景下的多目标跟踪。无人机目标跟踪既服务于交通监控、安防巡逻和搜救任务，也直接支撑动态目标追踪、路径调整和多机协同。与地面跟踪相比，无人机视角下的跟踪面临目标遮挡、快速运动导致的模糊以及视角突变带来的特征漂移等严峻挑战，尤其在高空道路场景中，大量外观相近的车辆和行人会显著增加身份匹配难度，造成跟踪丢失。

1.1.4　其他目标感知任务

除检测、分割和跟踪外，无人机基础目标感知还包括场景识别、变化检测以及运动估计等任务。它们虽然常被单独讨论，但本质上都服务于无人机对环境状态的基础建模。场景识别关注对城市道路、农田、水域、园区和灾害区域等整体场景类型及空间语义的判断，是从目标级感知迈向环境级理解的过渡。变化检测强调对不同时相或不同巡检周期影像中的新增、缺失与异常区域进行识别，在基础设施巡检、灾后评估和国土监测中具有重要价值。运动估计则聚焦于目标、相机或场景运动状态的恢复，为视频稳定、目标跟踪、避障与后续时空推理提供动态先验。

1.2　事件语义分析

相较于目标检测、分割与跟踪等基础视觉任务，行为事件分析更关注发生了什么、为何发生、将导致什么后果这类高层语义问题，处于无人机视觉理解由感知走向推理的关键环节。无人机场景中的行为事件分析通常涵盖行为识别、视觉问答、意图识别与事故检测等任务，其共同目标是在连续观测中建模目标动作、场景关系与事件演化过程，为城市巡逻、交通监管、灾害处置与低空安防提供可解释的智能支持。

1.2.1　行为识别

无人机行为识别主要关注对人、车、群体或飞行体动作状态的判别，是无人机从目标级感知走向事件级理解的基础任务。在应用层面，该任务可服务于交通监管中的违规行为分析、安防巡逻中的异常活动识别，以及灾害现场中的人员动态监测。由于航拍视角带来的尺度压缩与姿态变形，无人机行为识别较少依赖精细人体骨架，而更强调场景上下文、运动趋势与区域语义的联合建模。

1.2.2　视觉问答

无人机视觉问答强调围绕图像或视频内容进行自然语言交互，是连接视觉理解与任务决策的重要接口。相比传统图像问答，无人机场景下的视觉问答更强调空间关系、区域定位、场景概括与任务相关推理，例如，事故发生在何处，哪一片区域存在异常，该场景是否适合继续飞行等。这类任务能够将复杂感知结果转化为自然语言反馈，显著提升无人机系统的人机协同能力。

1.2.3　意图识别

意图识别关注从观测到的运动状态、载荷信息、环境条件与先验知识中推断行为主体未来目标或潜在目的，在低空安全治理、非合作无人机监管和对抗博弈场景中具有重要价值。与行为识别不同，意图识别试图回答接下来想做什么与为何这样做等更强的知识推理问题。

1.2.4　事故检测

事故检测是面向安全关键场景的重要任务，通常指对交通事故、火灾以及碰撞等异常事件进行快速发现、定位与场景分析。无人机具备机动部署快、覆盖范围广和视角灵活的优势，因此在高速公路巡检、灾后应急和城市安全监管中具有明显应用潜力。与一般异常检测不同，事故检测不仅要求识别异常是否发生，还要求对事故类型、影响范围和处置优先级进行概括，从而服务后续决策。

1.3　空间环境理解

与无人机客观视角下的基础目标感知与行为时间分析任务不同，空间环境理解任务构成了无人机主观视角的环境建模，要求智能体在三维空间中建立稳健的几何与语义关联。相较于基础感知，空间环境理解不仅关注目标的类别与位置，更强调对机体自身位姿、环境拓扑结构及动态障碍物演化趋势的综合判断，从而支撑定位、导航、避障与自主飞行。更通俗来说，这类任务聚焦于对无人机本体身处何处、周围如何以及下一步如何运动的空间认知问题。

与地面机器人相比，无人机的空间环境理解具有更强的三维性、动态性和远距感知特征。由于机体具有俯仰、偏航和滚转等高机动自由度，飞行高度变化会显著改变目标尺度与视场范围，而俯仰滚转带来更复杂的视角扰动，单视角视觉往往难以维持全局一致的场景表征。此外，在全球定位系统（global positioning system，GPS）拒止或高层建筑密集的城市场景中，迫使无人机必须依赖纯视觉或多传感器融合技术进行自主定位与环境建模。开放环境中的建筑、电线、树木与临时飞行物又不断改变可通行空间，因此模型不仅要具备几何建模能力，还要具备语义理解、跨模态对齐与动态决策能力（许越越等，2025）。

1.3.1　三维空间推理

三维空间推理强调无人机基于视觉观测、历史轨迹与语言目标，对高度、距离、拓扑结构、遮挡关系与可达路径进行综合判断，其目标是使无人机不仅能识别地标，还能理解目标在何处、应从何处接近以及当前动作如何改变后续可达性。因为空中导航天然处于三维空间：同一目标在不同高度、方位和视角下呈现出显著差异，且无人机动作包含俯仰、滚转、偏航和升降等耦合控制，自然比地面机器人面临更复杂的空间决策约束。

1.3.2　地理定位

地理定位关注的是无人机在大范围环境中确定自身在哪里或目标在哪里，既包括基于视觉与地图的跨视角定位，也包括由语言描述、图像示例或多模态目标驱动的主动地理定位。与三维空间推理相比，地理定位更强调跨区域匹配、语义锚点提取与坐标级落地，其本质是将视觉内容映射到真实地理空间中的某个位置或区域。一方面，无人机视角下地物密集、场景杂乱、尺度跨度大，容易造成语言描述与视觉区域之间的粒度失配；另一方面，真实任务中的目标描述往往具有模糊性和不完备性，如“靠近红色屋顶旁边的停车区”这类表达同时包含相对位置、外观属性与场景先验，对模型的跨模态对齐提出更高要求。此外，天气变化、季节变化和视角变化也会削弱跨场景定位稳定性。

1.4　飞行决策规划

在无人机视觉理解任务体系中，飞行决策规划位于感知理解之后、任务执行之前，是连接视觉信息与自主行动的重要环节。与传统无人机依赖规则控制或预设路径不同，视觉驱动的飞行决策规划强调从视觉环境感知中提取语义信息，并据此完成路径规划、任务分解与行动控制。随着深度学习与多模态模型的发展，无人机逐渐具备在复杂环境中进行自主决策与动态规划的能力，使其从单纯的感知平台向具身智能体演进。相关研究指出，智能无人机系统通常由感知、规划、控制与通信等模块组成，其

中规划模块负责根据环境状态和任务目标生成飞行策略，是实现自主飞行的核心功能模块。

无人机场景下的飞行决策规划具有明显的任务特点。一方面，无人机具有三维空间机动能力，飞行路径需要在复杂空间结构中实时更新，这对环境建模与空间推理提出更高要求；另一方面，无人机作业环境具有高度动态性，如城市建筑、树木、电线以及移动目标等因素均可能影响飞行安全。此外，低空无人机往往需要在多任务场景中运行，例如巡检、搜索与救援或物流配送，这使得决策规划不仅涉及路径优化，还包括任务调度与协同控制问题。无人机运行环境在视角变化、高度变化以及任务场景多样性方面具有显著复杂性，这些特征使得传统规则驱动的飞行控制难以满足智能化需求。

1.4.1　视觉避障

视觉避障是无人机自主飞行中最基础且最关键的任务之一，其核心目标是在飞行过程中实时识别环境中的障碍物，并根据环境结构动态调整飞行轨迹，从而保证飞行安全。与依赖激光雷达或超声波传感器的传统避障方法相比，视觉避障利用摄像头获取的图像或视频信息进行环境理解，具有成本低、信息丰富和适应性强等优势。然而，无人机场景下的视觉避障面临多方面挑战。首先，航拍视角导致障碍物尺度变化显著，小尺度障碍物难以稳定识别；其次，复杂背景和光照变化会降低视觉检测的可靠性；此外，无人机高速飞行要求避障系统具备实时响应能力，这对算法效率提出了严格要求。Tian等人(2025)指出，无人机视觉任务常伴随视角变化和尺度变化，这些因素会显著增加环境感知与决策规划的难度。

1.4.2　自主降落

自主降落是无人机执行任务的重要环节，尤其在物流配送、应急救援和巡检等场景中具有关键作用。该任务通常要求无人机通过视觉信息识别降落区域，并在复杂环境中完成安全降落。典型方法包括降落标志检测、视觉定位与姿态估计等技术。无人机自主降落面临的主要挑战来自环境不确定性。例如，降落区域可能存在遮挡或背景干扰，视觉系统需要在复杂场景中准确识别降落目标；同时，风速变化与平台运动也可能影响降落稳定性。此外，在开放环境中，降落区域往往缺乏明确标志，这要求无人机具备更强的场景理解能力。

1.4.3　视觉语言导航

视觉语言导航(vision language navigation, VLN)是近年来无人机决策规划研究的重要方向，旨在使无人机根据自然语言指令和机载视觉观测在未见场景中完成定向飞行，本质上是综合视觉感知、语言理解与空间推理的导航任务。相较于地面VLN，无人机VLN面临更显著的高度变化、长距离视野稀疏和三维空间关系复杂等问题，因而语言中的左侧建筑后方、沿河道上空前进、绕开高压线后悬停等描述，往往需要更精细的视角转换与空间映射(王子豫等，2026)。尽管VLN展现出良好的发展前景，但仍面临诸多挑战。例如，跨模态语义对齐难度较大，自然语言指令往往存在歧义；同时，动态环境中的实时决策要求模型具备高效推理能力。此外，真实环境中的数据获取成本较高，限制了大规模训练数据的构建。

1.4.4　目标搜索

目标搜索是空间环境理解中最具综合性的任务之一，它要求无人机在开放环境中根据给定目标或语义线索主动探索、判断、定位并持续逼近目标区域。这一任务通常涉及目标检测、路径规划以及搜索策略优化等多个环节。与视觉语言导航相比，目标搜索不预设明确路径终点，也不保证目标始终可见，因此对环境建模、目标表征和主动决策的要求更高。该任务通常出现在灾害搜救、城市巡检、野外搜寻和安防侦察等场景中，需要无人机在不完整信息下进行连续推理和区域覆盖。然而，目标搜索任务通常具有环境复杂、目标稀疏以及任务范围广等特点，这使得搜索效率与路径规划成为研究重点。

1.4.5　任务规划

任务规划是无人机飞行决策中的高层智能模块，其核心目标是根据任务目标和环境信息生成完整的执行策略，并将复杂任务分解为可执行的飞行动作序列。与传统路径规划仅关注空间轨迹不同，任务规划强调任务语义理解、行为序列生成以及动态环境适应能力。在复杂应用场景中，无人机往往需要同时执行多种任务，例如巡检、监视、物流配送或灾害救援，这使得任务规划不仅涉及路径优化，还需要在任务优先级、资源分配以及环境约束之间进行综合决策。无人机任务规划面临多方面挑战。一方面，现实环境中的任务目标通常以自然语言或抽象指令形式出现，例如，巡查建筑屋顶或寻找受灾人

员，如何将高层语义指令转化为具体飞行策略仍具有较高难度。另一方面，复杂场景中的环境变化往往难以提前建模，例如突发障碍物或任务目标变化，需要规划系统具备实时调整能力。传统无人机系统通常依赖预设规则或离线规划方法，在动态环境中缺乏灵活性。

1.4.6　协同规划

在大规模无人机应用场景中，单个无人机往往难以完成复杂任务，多无人机协同系统逐渐成为重要研究方向。协同规划的目标是在多个无人机之间实现任务分配、路径协调以及信息共享，从而提升整体任务效率与系统鲁棒性。相比单机系统，多无人机系统在覆盖范围、任务效率和容错能力方面具有明显优势，因此广泛应用于灾害救援、农业监测以及城市巡检等场景。然而，多无人机协同规划具有更高的系统复杂性。首先，多无人机之间需要进行高效通信以共享环境信息和任务状态；其次，任务分配与路径规划问题往往具有组合优化特性，其计算复杂度随着无人机数量增加而迅速增长。研究指出，多无人机任务分配问题通常可建模为旅行商问题、车辆路径规划问题或混合整数规划问题等，其求解复杂度随系统规模呈指数增长。

1.4.7　编队控制

编队控制是多无人机协同系统中的重要研究方向，其目标是在保持无人机群体结构稳定的前提下，实现协同飞行与任务执行。典型应用包括大范围监视、环境监测以及通信中继等任务。在这些场景中，无人机群体通过形成稳定编队，可以在保证覆盖范围的同时提升系统效率与可靠性。无人机编队控制的关键挑战在于多智能体系统的协调与稳定性问题。由于无人机之间需要保持特定空间结构，系统必须在飞行过程中实时调整个体位置与速度。同时，通信延迟、环境干扰以及无人机动力学差异都会对编队稳定性产生影响。无人机群体系统通常采用分布式决策机制，使每个无人机根据邻居节点信息调整自身行为，从而实现整体协调。

1.5　小结

无人机视觉理解任务呈现出从基础视觉感知到语义理解，再到空间建模与飞行决策的层级化结构。与传统地面视觉任务相比，无人机场景下的视觉任务不仅需要完成目标识别，还需要具备跨模态语义理解与空间推理能力。当前研究虽然在目标检测、跟踪及行为分析等方面取得显著进展，但在复杂环境理解、跨场景泛化以及任务级决策能力方面仍存在明显不足。随着无人机应用逐渐走向开放环境和复杂任务场景，未来视觉理解任务将进一步向多模态语义理解、三维空间认知以及具身智能决策方向演进，从而推动无人机系统由单一感知平台向智能自主体转变。

2　视觉理解基座模型

视觉理解技术的发展正在经历从特定任务驱动的局部建模向通用具身智能驱动的全局闭环推理的深刻变革。早期视觉算法虽然在卷积神经网络（convolutional neural network，CNN）的支持下实现了从手工特征向数据驱动的跃迁，但在面对复杂场景的全局语义建模时，其感受野受限的问题逐渐凸显。ViT（vision Transformer）（Dosovitskiy 等，2021）的提出通过自注意力机制打破了卷积结构的局部归纳偏置，使得模型能够在统一架构下捕捉图像的长距离依赖关系。这种表征学习范式的演化，不仅奠定了高质量视觉特征的基础，更推动了视觉基础模型、大语言模型以及多模态生成系统的融合。当前的大模型的演进脉络清晰地展现出从静态语义对齐到生成式理解与推理，并最终走向具身决策执行的技术路径，如图 2 所示。

2.1　通用视觉基础模型

随着大规模数据训练和多模态学习的发展，视觉模型开始从单任务模型向通用视觉基础模型演进。这类模型通过在海量图像或图文数据上进行预训练，学习通用视觉语义表征，使模型具备了强大的跨任务迁移能力和开放语义理解能力。

视觉基础模型的重要代表之一，CLIP（contrastive language-image pre-training）（Radford 等，2021）通过对比学习范式联合训练图像编码器与文本编码器，使图像与文本在共享语义空间中对齐。训练目标鼓励匹配图像与文本描述在嵌入空间中接近，而非匹配样本保持距离。由于训练数据来自互联网规模的图文对，CLIP 能够学习到丰富的语义概念，从而实现开放词汇识别能力。例如，在不进行任务微调的情况下，为实现零样本迁移，CLIP 将分类任务重构为图文检索问题，即通过输入图像的嵌入与一组文本提示的嵌入进行相似度匹配，动态预测类别

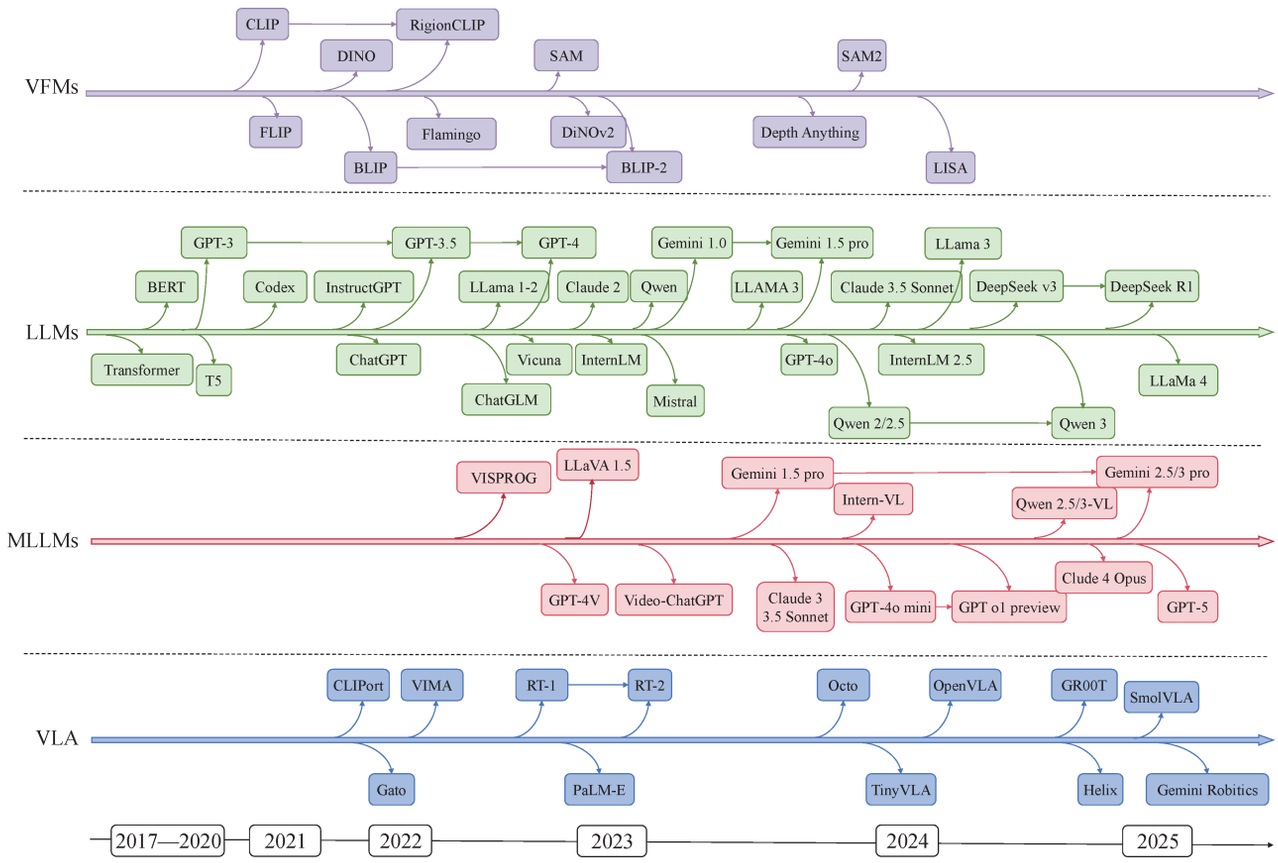


图2　视觉基础模型、大语言模型、多模态大语言模型、视觉—语言—行动模型的发展脉络

Fig. 2　Development timeline of VFMs, LLMs, MLLMs and VLA models

标签。该方法完全依赖自然语言作为监督信号，摆脱了传统模型对预定义类别体系的依赖，为开放域视觉任务提供了高度灵活的范式。然而CLIP主要侧重全局语义匹配，在细粒度时空理解或复杂视觉推理方面仍存在局限。

为进一步拓展视觉语言预训练模型的能力，BLIP（bootstrapping language-image pre-training）（Li等，2022）引入了统一的编码器—解码器架构，同时采用图文数据自举策略，通过过滤与重构噪声图文对提高预训练数据质量。与CLIP仅进行对比学习不同，BLIP在预训练阶段引入生成式学习目标，使模型能够胜任更加复杂的图像描述生成和视觉问答等任务。这种由对比学习向生成式建模的转变，增强了模型处理非结构化信息的能力。

在视觉空间理解领域，SAM（segment anything model）（Kirillov等，2023）展示了基础模型在通用的像素级感知中的巨大潜力。SAM通过大规模掩码数据训练，实现了提示驱动的分割机制。无论是点、框还是文本提示，该模型都能在未见领域生成精确的目标区域分割结果，证明了视觉基础模型不仅能够对齐视觉语义表示和实现生成式视觉语义理解，还能够在空间结构建模方面展现强泛化能力。

此外，以DINO（self-distillation with no labels）（Zhang等，2023b）为代表的基于自监督学习的视觉基础模型，通过自蒸馏机制在无标签图像数据上学习视觉表示。其核心思想是利用教师—学生网络的一致性约束，使模型能够学习到具有语义结构的特征表示。这类视觉特征在目标检测、分割以及检索任务中表现出较强的迁移能力，并且能够在无需人工标签的情况下形成稳定的视觉语义结构，为后续复杂的视觉推理奠定了低成本、高效率的表征基础。

2.2　大语言模型

面向自然语言的逻辑推理进一步扩大了传统视觉理解能力的边界，而大语言模型通过统一的语言接口，实现了通用的任务理解，使视觉系统具备大脑般的认知能力。

OpenAI（open artificial intelligence）提出的GPT（generative pre-trained transformer）系列（Brown等，

2020；Ouyang 等，2022；Achiam 等，2024）模型作为这一领域的开拓者，推动了生成式语言模型的快速发展。早期 GPT 模型主要展示了语言生成能力，而随着模型规模扩展至数十亿甚至上千亿参数，模型逐渐具备更强的语义理解与推理能力。在此基础上，后续模型进一步通过指令微调和人类反馈强化学习增强模型的指令遵循能力，使其能够更准确地理解用户意图并生成符合任务需求的输出。

开源语言模型的发展进一步推动了大语言模型研究的加速，其中 Meta 提出的 LLaMA（large language model meta artificial intelligence）系列（Touvron 等，2023；Grattafiori 等，2024）模型通过优化训练策略和数据筛选，在相对可控的模型规模下实现较强性能。与早期闭源模型相比，LLaMA 的开源策略使研究社区能够在统一基础模型上，针对高效训练与模型压缩开展更深入的探索，证明了通过优化训练策略与数据质量，中等规模模型亦能获得可与千亿级模型媲美的推理能力。

国内研究机构在这一浪潮中亦有显著贡献，例如，阿里巴巴提出的通义千问 Qwen 系列（Bai 等，2023；Yang 等，2025a）模型通过多阶段预训练和指令微调策略实现较好的中文与多语言能力，并在代码生成、数学推理和多任务理解方面表现出均衡且卓越的性能。上海人工智能实验室提出的书生—浦语（intern large model，InternLM）模型（Cai 等，2024）则更加关注开源生态建设，通过构建开放的工具链与数据生态，极大地降低了开发者进行模型适配与二次开发的门槛。近年来，DeepSeek 系列（Liu 等，2025；Guo 等，2025）模型在推理深度和训练高效性方面的突破引起了广泛关注。该系列模型通过专门针对推理任务设计的强化学习与高效推理优化策略，在处理数学证明、代码生成和高难度逻辑问题时表现出显著的竞争优势。

大语言模型不仅提供了理解真实世界的常识先验，更通过其技术优势，即通过提示工程或轻量化参数高效微调（parameter efficient fine-tuning，PEFT）实现快速适配未知场景的能力，为后续多模态大模型及具身智能系统发展提供了推理基础。

2.3 多模态大语言模型

随着视觉基础模型与大语言模型能力的持续提升，视觉理解范式正在从判别式任务模型向统一视觉与语言生成式架构演进。多模态大模型通常以大语言模型为核心，通过视觉编码器或跨模态对齐模块接入视觉多模态信息，使模型能够在统一生成式框架下完成视觉问答、图像理解、视觉推理以及跨模态生成等任务。与传统视觉模型依赖固定任务头不同，MLLM 通过自然语言指令驱动任务执行，从而实现更加灵活的视觉理解方式。

OpenAI 推出的 GPT-4V 在 GPT-4 语言模型基础上接入视觉流，使模型能够对图像内容进行理解并生成自然语言描述，在视觉问答、图像描述和复杂场景分析任务中表现出强大泛化能力。随后推出的 GPT-4o（Hurst 等，2024）进一步提升了多模态实时交互能力，通过统一架构同时支持文本、图像和语音输入输出，使多模态交互更加自然。

Google 提出的 Gemini 模型（Anil 等，2025）从设计之初便采用多模态统一架构。与传统视觉编码器 + 语言模型组合结构相比，Gemini 在训练阶段采用统一多模态数据进行联合学习，使不同模态之间能够共享表示空间。这种统一建模策略有效缓解了跨模态对齐过程中的语义损失，提升了系统在处理多模态复杂逻辑时的鲁棒性。然而统一训练通常需要更大规模的数据和计算资源，因此模型训练成本较高。

在开源领域，多模态大模型的发展主要以视觉语言模型为核心。LLaVA（large language and vision assistant）（Liu 等，2023）是较早公开的视觉语言模型之一，其核心思想是将 CLIP 视觉编码器输出通过线性映射接入 LLaMA 语言模型，并通过视觉指令数据进行微调，使模型能够完成视觉问答和图像理解任务。LLaVA 的贡献在于展示了一种相对简单且可复现的多模态模型构建方式，使研究社区能够在开源框架上探索视觉语言模型训练方法。随后的一系列迭代版本（如 LLaVA-1.5 等）通过更高质量数据和训练策略提升了模型性能。

针对更高分辨率与细粒度感知的需求，中国科研团队推出的 Intern-VL（intern vision-language model）（Chen 等，2024b）与 Qwen-VL（Bai 等，2025）在技术细节上进行了针对性优化。Intern-VL 系列模型通过结合高性能视觉编码器与语言模型，强调高分辨率视觉输入与视觉细节理解，使模型能够在复杂视觉场景中保持较强识别能力。Qwen-VL 模型在 Qwen 大语言模型基础上接入视觉模块，通过视觉指令微调增强图像理解能力。Qwen-VL 不仅支持常规

视觉问答任务，还能够完成图像目标与视频时空定位等细粒度理解任务，在开放词汇与时空语义关联中展现出显著的技术优势。

2.4 具身视觉—语言—行动模型

随着视觉基础模型与多模态大模型能力的不断提升，研究者逐渐意识到，仅具备视觉理解和语言推理能力并不足以支撑真实环境中的智能决策任务。在机器人操作、自动驾驶以及无人系统等应用场景中，模型不仅需要理解环境，还需要根据视觉信息与语言指令生成具体行动策略。由此，视觉—语言—行动模型逐渐成为多模态智能研究的重要方向。VLA模型通常通过统一建模视觉感知、语言理解以及动作控制，使系统能够在复杂环境中根据指令完成任务决策。这类模型的核心目标在于构建从感知到行动的闭环推理能力，使智能系统能够在现实世界环境中执行任务。

随着视觉语言模型的发展，研究者开始探索多模态指令驱动的操作策略，尝试在统一框架下建模视觉观察与动作序列。VIMA（visuomotor attention agent）（Jiang等，2023）为交互式操作领域的代表性框架，利用Transformer结构联合建模视觉观察、语言指令以及动作序列，实现了在物体搬运、排序及组合等复杂任务中的泛化。

Google提出的RT-2（robotics Transformer 2）（Zitkovich等，2023）将大规模视觉语言模型知识迁移至机器人控制模型中，使机器人能够利用互联网规模视觉语义知识进行决策。RT-2通过将机器人动作表示为离散文本标记，从而将机器人控制问题转化为序列生成问题。该模型在训练过程中融合互联网图文数据与机器人操作数据，使模型能够学习到更丰富的视觉语义概念，有效克服了机器人领域标注数据匮乏的挑战。在RT-2之后，研究者开始探索更大规模的视觉—语言—行动模型，RT-X系列（O'Neill等，2024）工作通过跨平台数据集共享与统一训练框架，进一步验证了多源操作经验的共享能够显著提升单一具身策略的稳健性。

与此同时，一些研究工作开始探索将多模态大模型直接应用于机器人决策系统。PaLM-E（pathways language model-embodied）模型（Driess等，2023）将视觉输入、语言信息以及机器人状态统一输入到语言模型中，使模型能够根据多模态信息生成机器人控制策略。该模型展示了大型语言模型在机器人规划任务中的潜力，使机器人能够通过语言推理进行任务分解和决策。然而由于语言模型并非专门为控制任务设计，其在低层控制稳定性方面仍然需要额外模块进行补充。

当前的OpenVLA（Kim等，2024b）等模型尝试通过大规模机器人数据训练统一视觉语言行动模型，通常结合视觉编码器、语言模型以及动作生成模块，通过端到端训练方式学习从视觉感知到动作执行的映射关系，使机器人能够在不同任务之间共享策略知识。这种大规模多任务训练显著提升了机器人策略的泛化能力，使模型能够在未见场景中完成类似任务。

2.5 小结

视觉理解模型的发展经历了从卷积神经网络到Transformer，再到视觉基础模型、语言模型及多模态大模型的持续演进过程。视觉基础模型通过大规模预训练提供通用视觉表征，大语言模型则赋予系统强大的语义理解与推理能力，而多模态大模型进一步实现了视觉与语言信息的统一建模，为复杂场景中的生成式理解与推理提供了基础。同时，视觉—语言—行动模型的出现标志着视觉理解技术正从理解世界走向与世界交互。然而，大模型在计算资源消耗、实时推理效率以及跨模态对齐稳定性等方面仍存在一定挑战。未来视觉理解基座模型的发展趋势将集中在统一多模态建模、模型轻量化以及感知—推理—行动闭环能力构建等方向。

3 无人机视觉理解大模型的研究进展

无人机视觉理解涉及从基础视觉感知到高层语义推理再到飞行决策规划的多层能力体系，不同技术范式在这一能力链条中的作用并不相同。表2对不同技术范式在无人机视觉任务中的能力边界与适用场景进行了系统对比，从核心能力、适用任务、适配优势以及主要局限等多个维度分析其在无人机场景中的技术特点。其中，视觉基础模型主要提升开放词汇识别与视觉感知能力，多模态大模型进一步强化视觉语义理解与跨模态推理能力，而视觉—语言—行动模型则尝试将视觉理解与飞行决策规划进行一体化建模。基于这一技术划分，本文接下来进一步从能力提升视角对大模型在无人机场景中的关键应用进行系统梳理，分别讨论其在视觉感知增强、视觉语义推理以及视觉决策规划等方面的研究

进展。

3.1 基于大模型的无人机视觉感知增强

无人机场景中的视觉感知长期受制于高空视角、远距离成像、小目标密集分布、背景纹理复杂以及天气与地域变化剧烈等因素，传统架构下的感知模块过度依赖预定义的固定类别识别，难以应对航拍视角下极端的尺度变化、光照波动以及复杂多变的运行环境。视觉基础模型与多模态大语言模型的引入，使无人机视觉感知范式和泛化能力进一步升级。模型不再仅依赖标注类别学习目标外观，而是借助语言先验、跨模态对齐和大规模预训练知识，逐步获得开放世界细粒度感知、跨域泛化和场景级理解能力。本节将从开放词汇泛化、细粒度感知、复杂场景稳健性及三维空间推理 4 个能力提升维度，深入探讨视觉理解大模型在无人机感知增强中的研究进展，如图 3 所示。

3.1.1 开放词汇泛化

传统深度学习算法在处理高分辨率航拍影像时，面临极高的人工标注成本与陌生新颖目标的挑战。无人机视觉大模型通过引入大规模预训练的先验知识，有效缓解了对手工标注的依赖，显著提升了其在零样本场景下的适配能力。Limberg 等人（2024）将 YOLO-World（you only look once-world）与 GPT-4V 结合，用于无人机图像中的零样本行人检测与动作识别。结果表明，大模型不仅能够在缺乏专门标注数据时完成候选目标筛选，还能够提供更具整体性的场景解释，体现出多模态模型在开放环境中的先验优势。面向更大尺度的遥感与航拍场景，

表 2 无人机视觉理解多模态大模型主要技术路线对比

Table 2 Comparison of large multimodal models for UAV visual understanding

技术范式	核心能力	适用无人机任务	适配优势	主要局限
视觉—语言对齐基础模型（CLIP、DINO 等）	图像—文本对齐、开放词汇识别	基础感知：航拍场景分类、开放词汇目标识别、检索	具有较强开放类别迁移能力，可减少人工标注依赖，适合快速构建无人机场景语义检索与粗粒度识别系统	可提供类别语义先验，但不能单独完成精确检测；对小目标、密集目标和细粒度局部差异不敏感；缺乏精确空间定位能力；在无人机小目标和细粒度类别中需领域适配或提示学习
图文生成基础模型（BLIP、BLIP-2 等）	图像描述、视觉问答、跨模态生成	基础感知与语义理解：航拍图像描述、场景问答、事件语义概括	能将无人机图像转化为自然语言描述，适合人机交互和任务描述生成	对复杂空间关系、异常事件因果关系和专业领域语义理解有限；容易生成泛化描述，细节可靠性不足
通用分割基础模型（SAM 等）	提示式分割、区域掩码生成、交互式标注	基础感知：目标分割、地物提取、巡检区域标注、数据集半自动标注	可显著降低无人机分割数据标注成本，对高分辨率航拍图像中的显著区域具有较好泛化能力	缺乏类别语义理解；对小目标、低对比度目标、复杂纹理背景和边界模糊区域表现不稳定；通常需要提示、后处理或领域微调
多模态大语言模型（GPT-4V、Gemini 系列、LLaVA 系列等）	图像视频理解、视觉问答、复杂语义推理	语义理解与飞行决策规划：航拍视觉问答、异常解释、任务指令理解、场景语义推理、初步决策辅助	能以统一生成式接口处理多种无人机视觉任务，适合从感知结果向语义解释和任务规划扩展	对细粒度空间定位、小目标识别和时序动态理解仍有限；存在幻觉风险；需要无人机领域指令数据和评测基准增强；对时序关系和因果链条仍不稳定，需结合视频模型、检索模块或结构化知识；全局场景语义理解力较强，但对精确几何、距离和方向判断不足，需融合深度估计、地图信息或多视角建模
视觉—语言—行动模型（RT-2、OpenVLA 等）	视觉—语言—行动映射、任务规划、具身决策	飞行决策规划：视觉导航、目标搜索、自主降落、避障、复杂任务协同	连通视觉理解与飞行控制，使无人机系统进一步具备感知—决策—控制一体化的闭环体系	当前多集中于机器人操作或通用具身任务，直接迁移到无人机需解决实时性、安全约束、物理控制和空域环境建模问题；适合高层任务规划和语言交互，但直接控制无人机仍需安全约束、实时感知和低层控制器协同

Zhang等人(2024c)构建了包含500万对图文数据的RS5M(remote sensing 5 million)数据集,并基于此微调GeoRSCLIP模型,在零样本分类任务中较传统模型性能提升了3%～20%,进一步验证了大规模领域数据预训练对于增强感知泛化性的重要价值。类似地,结合视觉语言模型与SAM,Ma等人(2024)面向高分辨率无人机道路场景解析提出一种新型无监督框架,通过结合视觉语言模型定位感兴趣区域并调用SAM生成分割掩码,在缺少人工像素级标注的条件下实现道路区域解析,通过自监督迭代训练实现了高达89.96%的交并比,证明了在无需人工定义类别的情况下自主学习新知识的可行性,更显示出视觉基础模型在高分辨率航拍场景中的高效适配性。Blei等人(2025)提出的CloudTrack进一步将开放词汇能力延伸到面向搜救场景的目标跟踪任务,该方法能够依据口头语义描述直接执行目标跟踪,

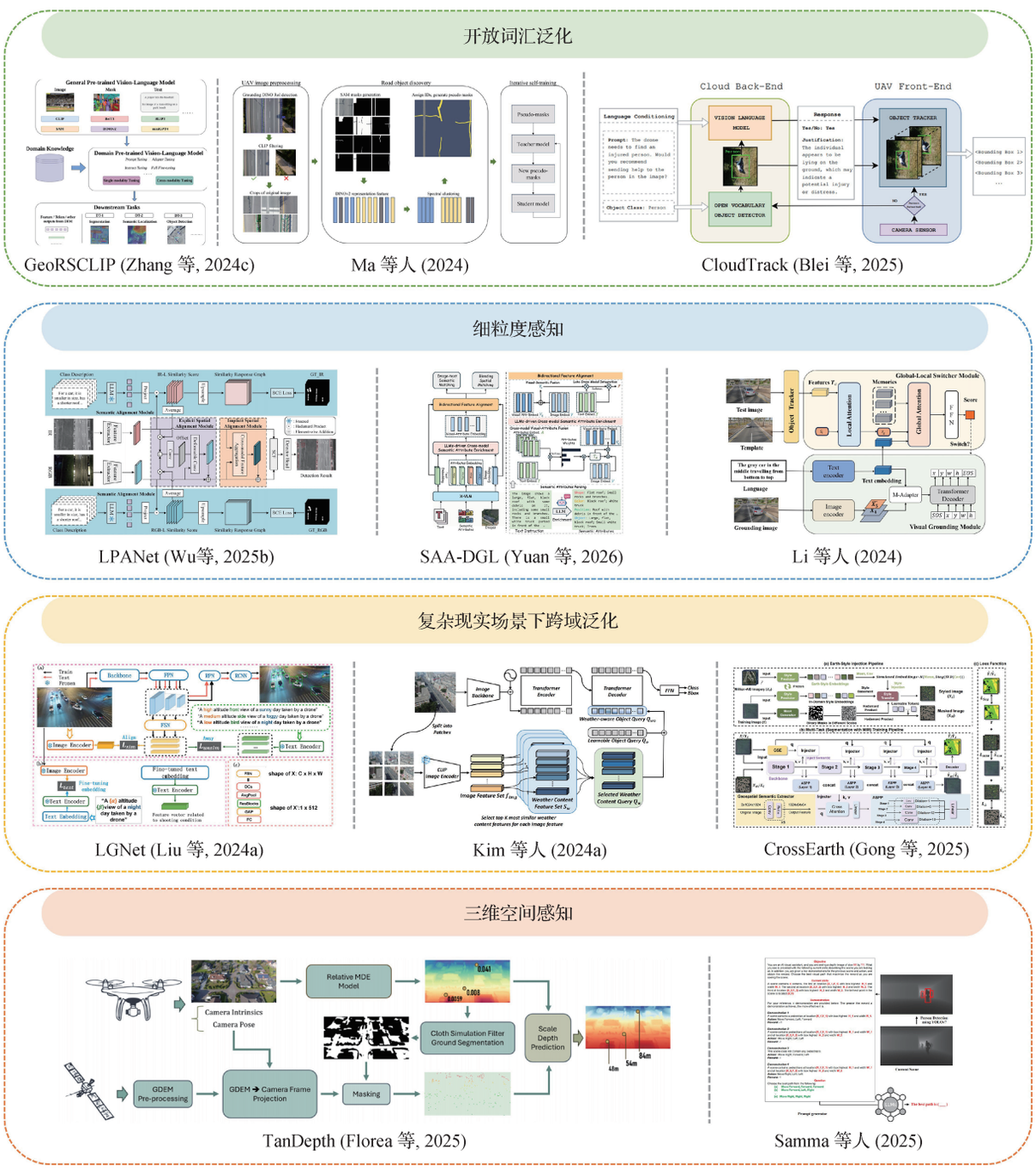


图3　基于大模型的无人机视觉感知增强的代表性研究进展

Fig. 3　Representative advancements on UAV visual perception enhancement with large models

而无需为每一类目标单独训练模型，提升了无人机执行搜救任务的灵活性。与传统先定义类别再训练模型的范式相比，视觉大模型使无人机系统在一系列下游任务中具备了依据语义信息感知关键目标的能力，更契合真实任务的动态性。

3.1.2　细粒度感知

无人机拍摄目标往往尺寸小、遮挡强，与背景混杂严重，目标特征极易被环境噪声淹没，而在一些拥挤场景下，外观相似的目标容易导致特征漂移，带来歧义干扰。针对微小目标难感知以及相似目标难辨别的问题，研究者致力于引入大模型的语言知识弥补视觉信息的不足，增强特征判别精度。

针对小目标识别难题，Wu 等人（2025）提出的 LPANet（large language model guided progressive feature alignment network）网络利用大语言模型生成目标的细粒度文本描述（如形状、颜色、位置属性等），并将其作为显式语义锚点引导多模态特征的渐进式对齐，直接参与跨模态感知特征优化，显著提升了复杂背景下的小目标检测性能。相比之下，Yuan 等人（2026）更关注复杂空间关系的细粒度属性建模不足的问题，在无人机地理定位中引入基于大语言模型的属性对齐机制，使其开发的 SAA-DGL（semantic attribute alignment-drone geo- localization）框架能更稳健地将颜色、结构和方位等细节语义嵌入视觉特征，强化跨模态信息关联。这种思路增强了模型在杂乱场景中识别模糊目标的能力，也同样适用于地标识别和精细目标辨别。为处理动态目标跟踪时，传统的边界框引导方式在处理相似目标时歧义大的问题，Li 等人（2024）指出，引入自然语言描述能够补充边界框难以表达的属性信息，并利用 CLIP 的多模态特征对齐能力，有效缓解了相似目标条件下的跟踪歧义，显著降低了车辆跟踪中的身份跳变率。

3.1.3　复杂现实场景下跨域泛化

无人机常需在雾天、低照度等恶劣气候或动态非受控环境下作业，这样复杂的现实场景下的视觉感知表现出的域偏移特征对模型的泛化性和鲁棒性提出了严峻考验。Liu 等人（2024a）在研究拍摄条件不敏感的无人机目标检测时发现，语言引导机制能够弱化对特定视觉域的依赖，使检测器在尺度、视角和非受控成像条件变化下保持更稳定的表现。他们提出的 LGNet（language guided UAV detection network）通过微调拍摄条件的文本提示嵌入，从视觉—语言特征空间中剔除领域特定的视觉干扰，使检测器能够适应高度剧变、视角旋转等非受控拍摄条件。Kim 等人（2024a）进一步将环境上下文显式引入检测框架，通过多模态大语言模型自适应提取天气信息并参与感知推断，显著增强了无人机在雨雾等极端气候下的目标检测能力。针对全球尺度的遥感解译难题，Gong 等人（2026）将视觉基础模型用于跨地域遥感语义分割，证明基础模型集成地球级风格注入和多任务学习，能够有效缓解由于地理位置、波长和传感器类型等差异带来的分布偏移与泛化退化问题，所推出的 CrossEarth 模型在覆盖 32 种跨域场景的基准测试中展现出卓越的领域泛化能力，显著优于现有最先进的领域自适应方法。由此可见，大模型凭借其强大的语境理解能力，能够有效提取环境上下文信息并将其融入感知框架。其在无人机场景中的关键价值，不仅体现为单域感知精度的提升，更在于实现了由环境知识驱动的跨域泛化能力增强。

3.1.4　三维空间感知

受限于机载载荷，无人机往往缺乏高精度激光雷达，而单纯依赖单目视觉的深度估计通常面临尺度不确定的困境。近期研究正致力于将大模型的逻辑推理能力与地理信息先验结合，以构建精准的三维空间意识。Florea 和 Nedevschi（2025）提出的 TanDepth 框架提供了一种创新的尺度恢复方案，通过将全球数字高程模型的测量值投影至单目视觉流，成功为无激光雷达的微型无人机提供了具有公制尺寸的高精度深度感知。而在无 GPS 的室内环境下，Samma 和 El-Ferik（2025）验证了利用 LLM 分析人员位置与深度信息生成路径的可行性，其碰撞率显著低于传统的深度强化学习模型。

3.2　基于大模型的无人机视觉语义推理

传统的视觉算法难以处理非受控环境下的复杂逻辑与因果推断，多模态大语言模型为无人机赋予了强大的语义推理能力，使其不仅能感知视觉信息，还能通过常识推理与逻辑分析完成场景描述、行为解释及异常识别等任务。本节将从开放语义理解、无需训练推理、统一多任务推理、复杂空间关系推理以及领域知识推理 5 个层面，探讨大模型在赋能无人机视觉语义推理的关键研究进展，如图 4 所示。

3.2.1　开放语义理解

早期无人机视觉系统主要输出类别标签、位置框和简单事件判别，其本质仍是任务专用的判别式

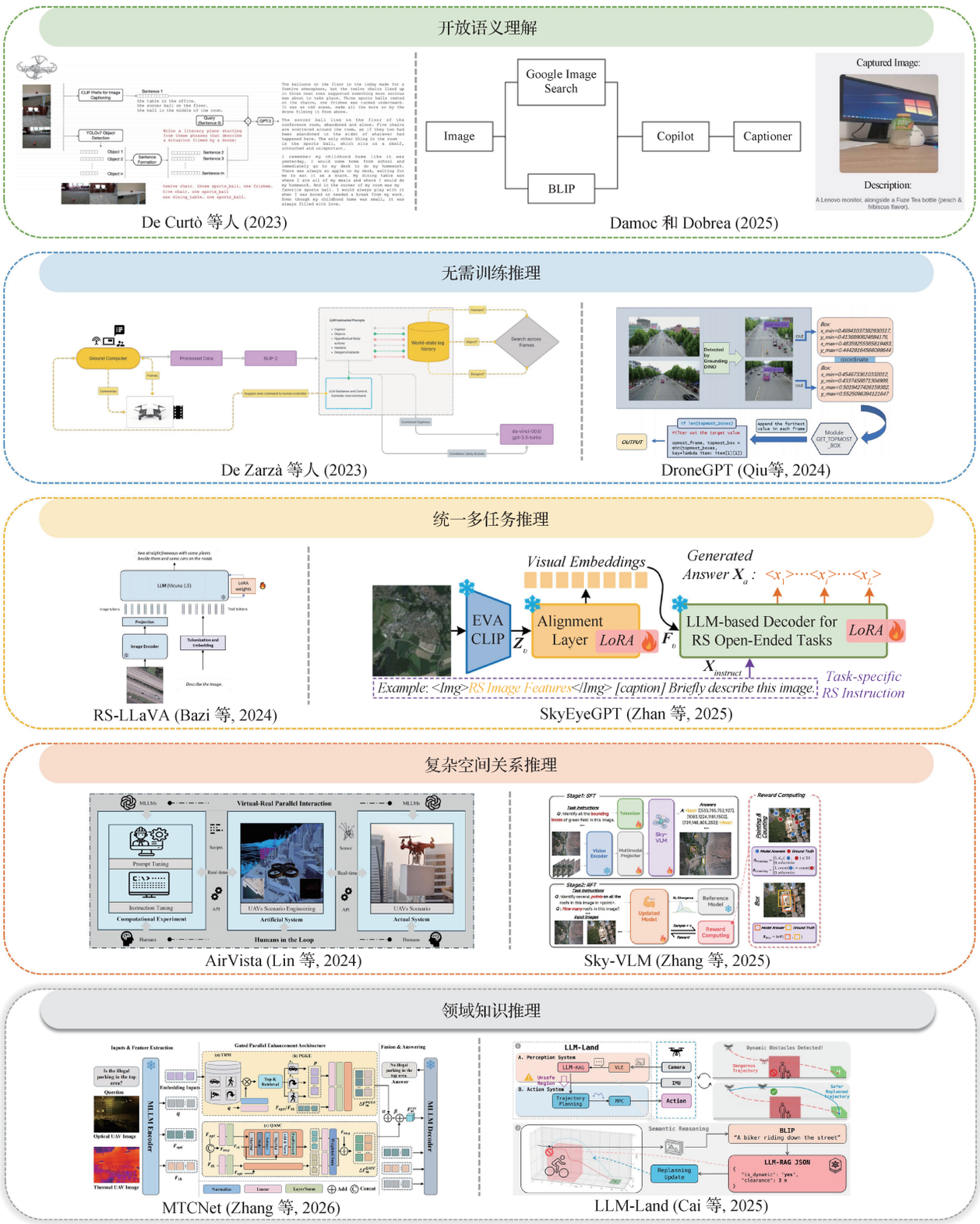


图4　基于大模型的无人机视觉语义推理的代表性研究进展

Fig. 4　Representative advancements on UAV vision semantic reasoning with large models

感知。随着视觉语言模型和多模态大模型的发展，研究开始强调将图像、视频转化为可解释的开放语义表示。De Curtò 等人(2023)指出，低成本微型无人机已能够借助大语言模型生成高质量场景描述，说明无人机视觉理解正在由黑箱化的结构化检测结果向自然语言语义解释转变。类似地，Damoc 和

Dobrea(2025)进一步指出,这种视觉描述生成实现了从简单检测到近人类语义解释的范式转变。与传统目标检测相比,这类方法的价值不在于替代检测器,而在于为后续问答、推理和决策提供统一语义接口。

3.2.2 无需训练推理

无人机执行任务的环境通常具有高度的不可预测性,多模态大语言模型在无需针对特定任务进行微调的情况下,也能展现出卓越的视觉文本推理潜力。De Zarzà等人(2023)通过BLIP-2等视觉语言模型将无人机采集的原始视频流转化为语义日志,记录下物体、人类及其潜在危险状态信息,之后在苏格拉底式推理框架内,利用LLM强大的语义推理能力,使无人机可以在极少人工干预下进行可解释的事件预测并为操作员提供决策辅助。在更为复杂的无人机场景视频问答任务中,Qiu等人(2024)提出的DroneGPT框架采用了神经符号方法,将GPT-3.5、Grounding DINO与程序化视觉推理结合起来,利用LLM的上下文学习能力将自然语言指令解析为模块化程序。通过调用现成的检测模型,以逻辑连接的方式实现对无人机视频的零样本推理,极大地拓展了无人机在语义层面的任务边界,说明复杂组合任务并不一定依赖专门训练的数据闭环。

3.2.3 统一多任务推理

无人机场景覆盖遥感、交通、巡检和搜救等多个领域,任务形式横跨图像描述、问答、定位、检索和事件理解。若为每一类任务都单独构建模型,系统代价极高,也不利于能力迁移。面向无人机的视觉描述与视觉问答系统正逐渐摆脱任务独立微调范式,而采用统一视觉语言建模,以形成可迁移的语义接口和跨任务共享表示。Bazi等人(2024)提出RS-LLaVA,将遥感图像字幕生成与视觉问答联合建模,利用指令微调将不同语义任务对齐到统一接口中。更进一步,Zhan等人(2025)提出SkyEyeGPT,通过高质量指令微调统一遥感视觉语言任务,在图像描述、问答与视觉定位等任务上表现出较强适配能力。

3.2.4 复杂空间关系推理

无人机在三维空间中运行,其视觉语义推理必须超越二维像素层级,建立对地理拓扑与三维几何的深刻理解。然而,现有大模型在处理高空视角的细粒度空间感知时仍存在不足。针对这一局限,研究者开始探索如何将空间知识显式地融入推理链条。Lin等人(2024)提出的AirVista框架,通过融合三维空间知识的指令微调策略,显著提升了模型在复杂城市环境中的推理效率。此外,为了系统性地评估大模型在无人机视角下的空间智能,Zhang等人(2025)推出了SpatialSky-Bench基准测试,涵盖距离测算、高度估计及降落安全分析等多个子类。通过构建百万级样本的SpatialSky-Dataset并训练Sky-VLM模型,研究者证明了多粒度空间推理在提升无人机场景解析有效性方面的核心作用。

3.2.5 领域知识推理

无人机在真实开放环境中执行特定任务(如交通巡检、灾害响应)时,常常面临视觉证据不足但决策风险很高的挑战,例如细粒度交通违规、低空飞行意图识别和复杂搜救任务中的异常判断。这类问题仅靠图像表面模式匹配往往不够,还需要规则知识、情境先验和任务常识的支持。

在交通场景理解中,传统的感知模型难以识别细粒度的违规行为。为此,Zhang等人(2026)提出了MTCNet(multi-modal traffic cognition network)架构,通过引入外部交通规则记忆库提取高层语义原型,将领域知识与视觉表征联合建模,赋予模型解析复杂交通违规逻辑的能力,支撑了无人机在复杂应用落地的可靠性。而在安全性要求极高的自主降落任务中,Cai等人(2025a)提出LLM-Land,利用轻量化模型结合检索增强生成技术,实时推断情境感知的安全缓冲距离(如针对行人与车辆设置不同的避障阈值),使无人机在动态非结构化环境中展现出优于传统视觉控制器的规避性能与飞行安全性。在更为极端的搜救场景下,Cai等人(2025b)提出的NEUSIS(neuro-symbolic system)框架通过集成神经符号视觉感知与推理模块,在信息不确定的条件下对特定目标进行定位。该系统利用概率世界模型进行环境表征,确保了推理过程的可解释性。

3.3 基于大模型的无人机视觉决策规划

随着无人机视觉理解从单纯的视觉任务模型转向语言驱动的视觉推理与行动系统,其研究范式进一步向具身决策的深刻演进。在这种新兴范式下,大模型不再仅仅充当静态的特征提取器,而是作为无人机的大脑,将高层语义指令、复杂的视觉场景信息与底层的飞控策略相结合,实现从感知到行动的闭环推理。本节从人机交互决策、复杂空间导航决策、复杂任务规划、群体智能协同4个方面,分析大

模型在无人机视觉决策规划中的研究进展，如图5和图6所示。

3.3.1 人机交互决策

传统无人机控制依赖于预定义的指令集或复杂的编程语言，控制接口门槛高而且任务描述方式僵硬，极大限制了非专家用户的交互效率，而大模型的引入显著降低了这一门槛，实现了基于自然语言的直观灵活交互。Joshi等人（2025）提出的Neuro-LIFT（neuromorphic and large language model-based interactive framework for autonomous drone flight）框架，通过结合LLM的语义理解能力与神经形态视觉系统，成功将人类语音转化为高级规划指令，在边缘端实现了低功耗、低延迟的自主避障与导航。然而，LLM在词元序列生成过程中固有的延迟问题往往制约了实时交互的性能。针对这一局限，Chen等人（2024a）开发了TypeFly系统，将GPT-4用于解析用户自然语言，并设计轻量级任务规划语言MiniSpec，以减少冗余生成、提高任务脚本生成效率，将无人机对自然语言指令的响应时间缩短了62%，显著提升了控制计划生成的实时性。TypeFly通过将语言理解前移为规划入口，使无人机能够从高层语义目标自动过渡到可执行任务序列。为了进一步确保自然语言控制的安全性，研究者开始关注指令验证机制。Tazir等人（2023）提出了首个针对大模型生成指令的验证系统，确保了日常语言转化为飞控动作过程中的合法性，不仅提升了人机协作的直观性，也为无人机在复杂任务中的动态重规划奠定了基础。

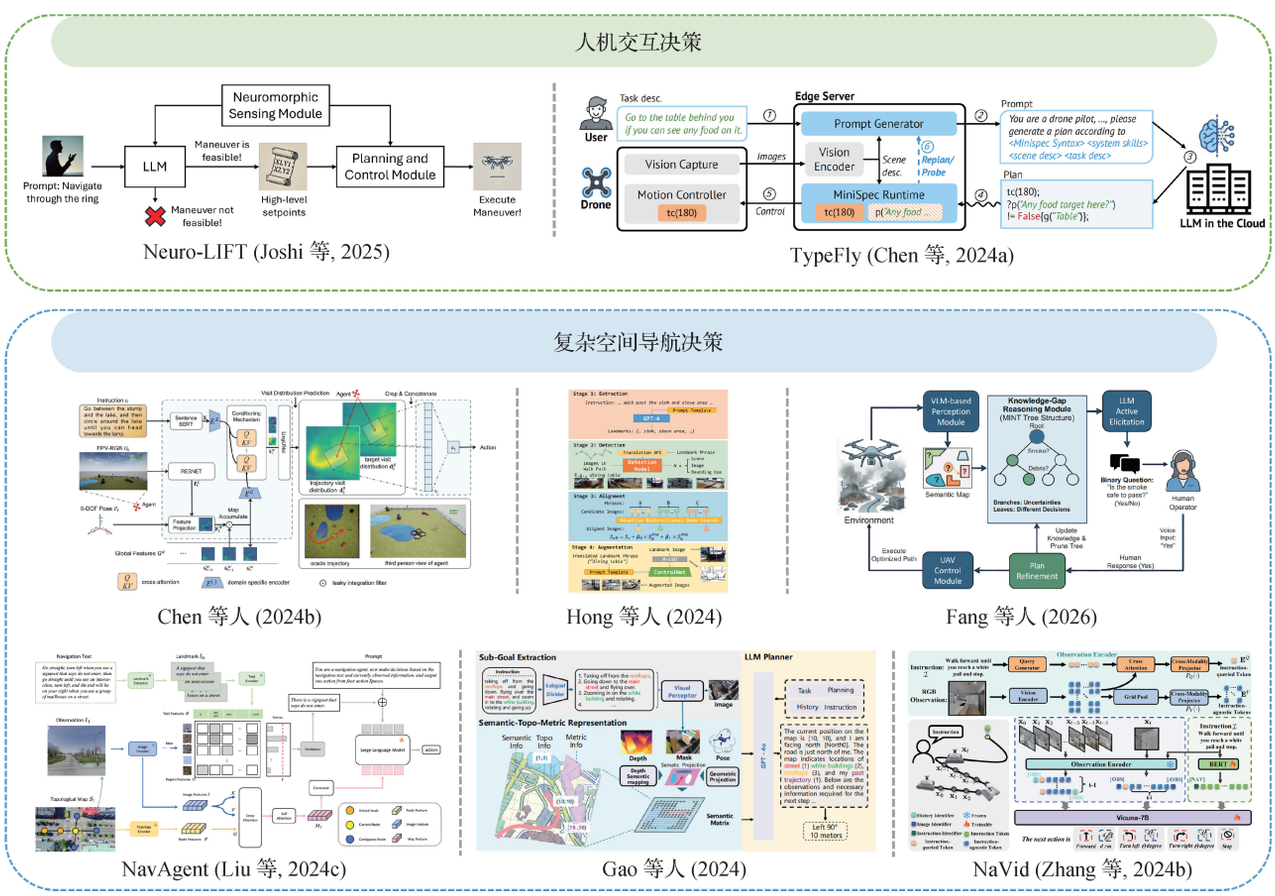


图5 基于大模型的无人机视觉决策规划的代表性研究进展

Fig. 5 Representative advancements on UAV vision decision-making and planning with large models

3.3.2 复杂空间导航决策

在真实的空中导航场景中，由于自然语言指令存在语义模糊性，且无人机任务常常牵涉空间关系、参考目标和隐含约束，仅依赖文本往往难以确保导航精度。为应对这一挑战，Chen等人（2023）在四旋翼视觉语言导航中引入基于提示的文本解析器，对用户指令进行语义重构；而Hong等人（2024）指出，单一的文本指令难以涵盖复杂的视觉先验，通过引

入多模态提示，允许用户结合图像与文本进行引导，可以有效消除语义歧义，使智能体在预探索环境中展现出超越纯文本模型的导航性能。针对指令模糊导致的决策挑战，Fang 等人（2026）提出的 MINT（minimal information neuro-symbolic tree）机制则将知识缺口结构化为可查询格式，通过主动向人类操作员发起最小化交互有效消除歧义。空间推理能力的缺失是制约无人机大模型应用的另一大瓶颈。为此，Liu 等人（2024c）的 NavAgent 则通过多尺度城市街景融合，将全局拓扑、全景视图和局部地标联合编码，以提升对复杂城市环境中细粒度目标的匹配能力。Gao 等人（2024）开发了一种语义—拓扑—度量表征，将与指令相关的语义掩码投影到俯视地图上，动态呈现周围地标的拓扑信息。这种将视觉语义与度量矩阵相结合的方法，增强了 LLM 在动作预测中的空间推理能力，在复杂导航任务中的成功率大幅超越了传统方法。与之相似，Zhang 等人（2024b）提出的 NaVid 模型进一步证明，仅通过机载单目摄像头获取的视频流，无需地图或里程计支持，即可实现从仿真到现实的高泛化导航，这种模拟人类导航逻辑的方法天然规避了传感器噪声带来的累积误差。由此可见，视觉信息进入决策规划之后，其作用已从辅助感知提升为驱动空间推理，而空间表征质量也会显著影响模型决策规划能力。

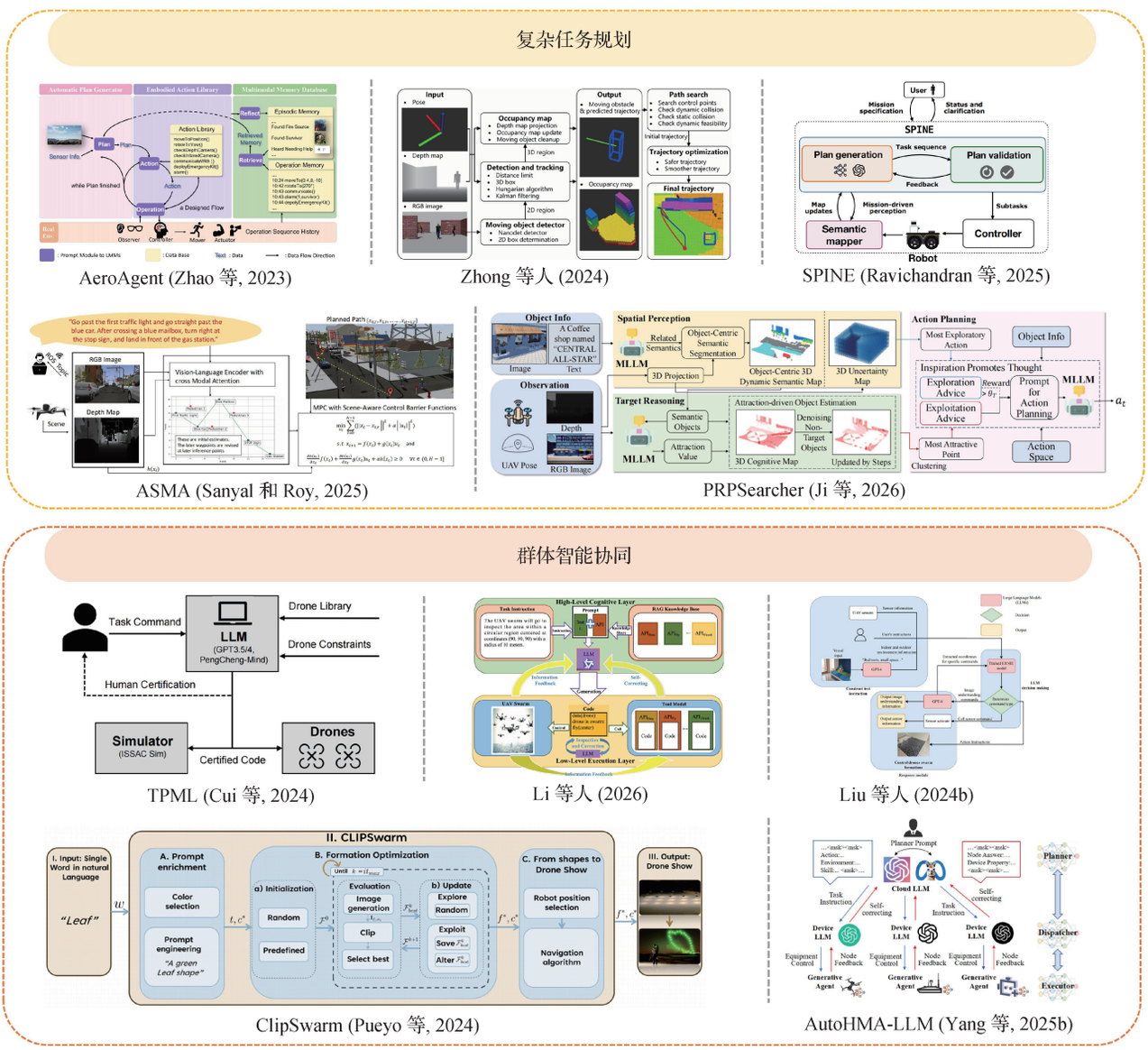


图 6　基于大模型的无人机视觉决策规划的代表性研究进展（续）

Fig. 6　Representative advancements on UAV vision decision-making and planning with large models (continued)

3.3.3　复杂任务规划

面对搜救、工业巡检等涉及多阶段推理的复杂任务，研究者倾向于采用层次化的决策框架。Zhao等人（2023）形象地提出了大脑代理—小脑控制器架构AeroAgent，其中MLLM作为大脑负责高层任务分解与环境意图推理，而底层的ROS控制器则作为小脑执行具体的飞控指令。这种架构有效缓解了大模型在底层实时控制稳定性上的不足。对于长周期、不确定性高的搜索任务，系统的可解释性与鲁棒性至关重要。Döschl和Kiam（2024）的Say-REAPEx则利用LLM对在线计划空间进行筛选，提升搜救任务中的规划效率。与之类似，Cai等人（2025a）的LLM-Land将检索增强与场景解析相结合，为自主降落推断情境感知的安全缓冲距离，再将结果输入模型预测控制（model predictive control，MPC）模块完成重规划，体现出语言模型与传统优化控制协同的典型路线。这一类研究说明，无人机决策规划并不意味着抛弃经典控制理论。相反，大模型更适合承担任务理解、风险解释、先验注入和候选策略生成，而安全控制、动力学约束和实时轨迹优化仍需要传统控制模块保障。Zhong等人（2024）将动态障碍物轨迹预测与视觉自主规划结合，并探索LLM参与安全规划；Sanyal和Roy（2025）提出ASMA（adaptive safety margin algorithm），通过场景感知控制屏障函数增强视觉语言导航中的安全边界。这表明当前较为可行的技术路径，并不是让大模型独占决策链条，而是形成大模型负责高层语义决策，小模型或控制器负责低层安全执行的协同架构。Ravichandran等人（2025）的SPINE则面向不完整自然语言任务描述，结合Grounding DINO、LLaVA与语义拓扑图进行在线语义规划，将复杂任务拆解为可执行子路径，体现出大模型在不确定任务规范下的任务补全能力。此外，针对资源受限的硬件环境，Wang等人（2025）提出的LMUCS（lightweight LLM-driven UAV control system）系统通过LoRA（low-rank adaptation）技术微调紧凑型LLM，结合YOLO检测与深度估计，证明了在低功耗平台上实现复杂物资投送任务的可行性。对于更复杂的城市搜救任务，Ji等人（2026）提出的PRPSearcher（perception-reasoning-planning searcher）智能体通过三维认知地图与去噪机制，模拟人类根据视觉线索进行思考推理的过程，有效解决了城市环境中相似物体的干扰问题，其搜索成功率较基线方法提升了37.69%。

3.3.4　群体智能协同

视觉大模型同样在多无人机集群协同领域展现出巨大潜力，改变了以往依赖刚性预定义策略的现状。无人机编队、协同搜索和集群任务不仅需要空间规划，还涉及意图共享、任务分配与角色协调，因此比单机导航更依赖高层语义推理。例如，Cui等人（2024）的TPML（task planning for multi-UAV system）验证了以大语言模型作为多机任务规划接口的可行性。Li等人（2026）提出的意图驱动协同控制框架，能够解析高层人类意图并自动生成Python集群控制代码，极大提升了编程效率与系统自适应性。在具体的编队任务中，Liu等人（2024b）创新性地利用多模态大模型处理领航无人机的实时图像，实现了基于视觉特征理解的集群编队规划，其成功率达到83.8%。为了简化多无人机表演等创意性任务的流程，Pueyo等人（2024）利用CLIP模型计算文本描述与队形视觉呈现的相似度，实现了从自然语言描述到无碰撞群体运动指令的直接映射。针对异构集群系统在资源受限环境下的任务调度难题，Yang等人（2025b）提出的AutoHMA-LLM（autonomous heterogeneous multi-agent systems with large language model）框架采用云端中央规划器与设备专用微型模型的混合架构，在确保任务完成准确率的同时，将通信开销减少了46%。

3.4　小结

大模型的引入显著提升了无人机系统在视觉感知增强、语义推理能力以及任务决策规划等方面的综合能力。通过融合大规模视觉—语言知识，模型不仅能够实现开放词汇感知和跨域泛化，还能够在复杂场景中进行多步语义推理与空间关系分析，为无人机执行高层任务提供更具解释性的决策支持。然而，现有研究仍主要集中在单一能力维度的提升，真正面向复杂任务的统一理解与推理框架仍有待进一步探索。此外，大模型在无人机平台上的部署仍面临实时性与资源受限等实际问题。未来研究需要在统一视觉理解框架、轻量化推理机制以及具身智能决策模型等方面进一步突破，以推动无人机系统向更高水平的自主智能发展。

4　无人机视觉理解数据集与评测基准

无人机视觉系统正经历从传统视觉感知向具身

智能体的范式演进。在这一过程中，高质量数据集与系统化评测基准发挥着关键支撑作用。随着视觉模型能力从单一任务识别逐渐扩展到多模态理解与推理，无人机视觉数据集的构建重点也从早期的单一视觉标注，逐渐转向视觉—语言—动作等多模态信息的联合建模与对齐。因此，构建覆盖多任务、多场景与多模态信息的数据体系，并设计能够反映模型综合能力的评测协议，已成为推动无人机视觉理解研究的重要基础。

4.1 无人机视觉任务数据集

近年来，无人机视觉任务数据集正从封闭类别、单模态标注逐渐演进为开放词汇、多模态语义描述的数据形态。这一演进不仅体现在数据规模的持续扩大，更重要的是通过引入自然语言描述与语义标注，使模型能够对复杂场景进行更深层次的理解。

在基础视觉感知任务方面，大规模标注数据为目标检测与分割任务提供了重要支撑。VisDrone（Zhu 等，2022）和 UAVDT（UAV detection and tracking）（Du 等，2018）等经典数据集为无人机视觉算法的发展奠定了重要基础，但这类数据集大多侧重视觉检测任务，缺乏丰富的语义信息支持。SynDrone（Rizzoli 等，2023）用合成环境生成约 7.2 万个像素级标注样本，为无人机场景下的检测与分割研究提供了高质量训练数据。然而，由于合成数据与真实航拍场景之间存在明显的分布偏移，基于真实环境采集的数据集往往更具研究价值。AirFisheye（Jaisawal 等，2024）面向城市复杂环境构建数据集，通过融合鱼眼相机、热成像以及 LiDAR（light detection and ranging）等多源传感器数据，显著提升了模型在极端视角与复杂光照条件下的感知能力。

在更大尺度的遥感场景理解任务中，数据规模与类别分布不均衡问题尤为突出。经典遥感数据集如 xView（Lam 等，2018）和 DOTA（dataset for object detection in aerial images）（Xia 等，2018）奠定了大规模检测数据的基础，虽然提供了数百万级标注实例，但其开放词汇与跨场景泛化能力仍然有限。而 RS5M（Zhang 等，2024c）等数据集的出现，标志遥感视觉数据正逐渐进入大规模图文对齐阶段。其中，RS5M 数据集通过利用预训练模型为纯标签数据自动生成文本描述，构建了百万级图文配对数据，从而显著增强了视觉基础模型在遥感领域的零样本适配能力。Gong 等人（2026）则更进一步，构建了包含 32 种跨域场景的 RSDG 基准，通过地球风格注入策略构建跨域数据分布，为遥感视觉基础模型的泛化能力评估提供了系统框架。

随着无人机视觉理解研究逐渐从静态感知扩展到动态理解，目标跟踪与行为事件识别等时序任务也开始融入多模态语义信息。在人类行为理解方面，UAV-Human（Li 等，2021）提供了跨 119 个受试者的 6.7 万组多模态序列，涵盖了姿态估计与属性识别，为低空视角下的人机交互研究提供了核心资源。TNL2K（tracking by natural language 2k）（Wang 等，2021）通过引入自然语言描述来指导目标跟踪，使模型需要理解目标的形状、属性以及空间关系，从而提升了跟踪任务的语义理解能力。WebUAV-3M（Zhang 等，2023a）构建了百万级规模的视频数据，通过整合视频、文本和音频信息，为多模态跟踪与视频理解提供了重要数据基础。在复杂动态背景下，模型对目标属性（如动作、姿态剧变）描述的准确性直接影响跟踪稳定性。为此，Li 等人（2024）提出的 UAVNLT（UAV-view natural language-guided tracking）基准通过引入细粒度的自然语言标注，有效缓解了相似目标干扰导致的身份跳变难题，使跟踪任务具备了人机交互能力。针对事件级的高层语义理解，CapERA（Bashmal 等，2023）在 ERA（event recognition in aerial videos）（Mou 等，2020）灾害与事故视频的基础上，引入了细粒度的文本描述，使事件识别从简单的类别标签扩展到语义层面的事件解释。Zhang 等人（2026）提出的 Traffic-VQA（UAV traffic visual question answering）数据集包含约 130 万组问答对，要求模型不仅具备基本视觉感知能力，还能够结合交通规则等领域知识进行推理。

针对视觉导航相关任务，虽然早期仿真环境数据使得相关研究取得了重大进展，但合成环境与真实航拍视角天然存在巨大的分布偏移。为此，Lee 等人（2025）推出的 CityNav 数据集覆盖了大规模真实城市区域，并要求智能体理解地标与目标之间的复杂空间关系。Gao 等人（2025a）开发的 OpenFly 平台，利用 3D 高斯泼溅技术实现真实场景到虚拟环境的高质量重建，为大规模航拍视觉语言导航任务提供了超过 10 万条高质量飞行轨迹。为进一步提升模型的推理能力，Cai 等人（2026）构建了基于真实城市航拍数据构建的大规模无人机 VLN 基准数据集 AirNav，其指令自然多样且完全脱离合成环境，通过

结合监督微调与强化微调来提升模型性能与泛化能力。

4.2 无人机视觉应用数据集

应用层数据集的构建则直接反映无人机视觉技术在实际场景中的应用需求，体现了技术向实际生产力的转化，涵盖城市治理、工业巡检、农业监测、灾害应急以及军事侦察等多个领域。

在城市治理与智慧交通场景中，数据集构建重点在于复杂环境下的目标识别与行为分析。例如，TrafficNight（Zhang等，2024a）通过融合可见光与红外数据，有效弥补了现有数据集在夜间和极端光照条件下覆盖不足的问题。Eesaar等人（2025）则构建了基于视觉语言模型的交通事故数据集，通过利用大模型自动生成结构化事件摘要，大幅提升了复杂交通场景下的态势感知能力。

在工业与农业巡检任务中，数据集更加关注细粒度目标识别。例如，InsPLAD（power line asset inspection dataset and benchmark）（Silva等，2023）针对电力巡检任务构建了包含约3万个实例的电力设备缺陷数据集，涵盖损坏、老化等多种异常状态，为自动化巡检系统的研究提供了重要数据基础。在精准农业领域，Avo-AirDB（avocado UAV database）（EL Amraoui等，2022）通过高分辨率航拍图像监测鳄梨作物的健康状态；WEED-2C（Tetila等，2024）则针对大豆田间杂草识别问题，展示了视觉模型在农业生产中的应用潜力。

在灾害应急与军事侦察等任务中，数据集通常强调场景真实性、情报准确性与响应实时性。SARsearchVL（Chen等，2025）面向搜救场景构建视觉定位数据集，通过结合语言描述增强大模型在山区、废墟等复杂环境下的语义对齐能力与目标识别能力。类似地，CityAVOS（city aerial visual object search）（Ji等，2026）主要模拟了人类根据视觉线索在城市环境中进行目标搜索与推理的过程，也为视觉推理研究提供了重要基准。而在灾害评估方面，FloodNet（Rahnemoonfar等，2021）主要提供灾后场景的高分辨率影像及问答数据，用于评估淹没程度与辅助救援规划。在军事侦察领域，MOCO（military objects in real combat）（Pan等，2024）提出了军事图像字幕生成任务，通过生成结构化文本情报，将无人机获取的视觉信息转化为可用于决策支持的语义信息，大幅提升了战态感知的效率。

4.3 无人机视觉理解能力评测体系及其需求

随着无人机系统逐渐向具身智能体范式发展，传统基于单一任务准确率的评测方式已难以全面反映视觉理解大模型的综合能力。研究表明，在复杂无人机场景中，大模型往往面临能力幻觉、物理常识缺失以及空间推理不足等问题。因此，有必要构建以核心能力为导向的综合评测体系。

首先是评测无人机智能视觉理解在开放世界下的零样本泛化能力。与封闭类别任务不同，无人机在真实环境中常常会遇到未见过的目标或突发事件。评测协议更应该强调模型在分布外场景中的空间推理与任务规划能力，而不仅仅关注分类准确率。

其次是长程空间推理与一致性评测指标。Zhang等人（2024b）指出，模型必须具备将自然语言指令转化为逻辑连贯的飞行轨迹的能力，这要求评测体系引入感知—推理—行动的闭环评估。而在SpatialSky-Bench（Zhang等，2025）中的评测结果表明，即便目前最先进的多模态大模型，在无人机视角的距离测算、高度估计及降落安全性分析中仍表现欠佳，反映出其在航拍几何理解上的明显缺陷。这种空间认知的局限性进一步制约了无人机在三维环境下的避障与规划效率。推理一致性还体现在视觉锚定可靠性上，UAV-CodeAgents（Sautenkov等，2025b）展示了模型将语义目标映射到像素坐标的能力对物理行动执行的重要性。

此外，指令遵循的一致性与逻辑性也是体现模型智能程度的重要指标。Yao等人（2025a）提出的评测框架首次定义了感知—推理—导航—规划的具身闭环评估协议，强调通过具身思维链机制衡量模型的决策逻辑是否符合物理现实。Ferrag等人（2025）提出的UAVBench数据集强调了伦理安全与资源约束下的决策评估，通过数万个已验证的飞行场景，测试模型在面对模糊指令时是否会出现幻觉。

随着无人机系统逐渐部署到边缘设备上，实时推理能力与可靠性也成为重要评测指标。为此，近年来研究开始探索多种模型轻量化与高效部署技术，包括网络剪枝、低比特量化、知识蒸馏以及参数高效微调（Lee等，2025；Wang等，2025）等方法，以在保证视觉理解能力的同时降低模型计算复杂度。例如，SlimYOLOv3（Zhang等，2019）通过通道剪枝显著降低网络参数与计算量，实现接近原始模型精度的同时获得更高推理速度。后续研究进一步结合动态

滤波器剪枝与轻量检测结构，实现近百帧每秒的航拍目标检测性能（Bellec，2026）。Yao 等人（2025b）则引入知识蒸馏用于训练轻量级 UAV 目标检测器，通过将大模型的语义知识迁移至小模型，使其在复杂航拍场景中保持较高检测精度。Sabaghian 等人（2025）进一步将结构剪枝与通道蒸馏联合应用，显著降低了计算量，在边缘计算平台上实现实时部署。此外，一些研究通过构建端—边—云协同架构（Yang 等，2025b；Yuan 等，2024），将低延迟的目标检测、避障与跟踪任务部署在机载端侧，将语义推理与多机协同分析放置于边缘计算节点，而复杂的任务规划与大模型推理则由云端完成，从而在系统层面实现计算资源的动态调度。在此基础之上，Zhao 和 Lin（2025）的研究表明，模型在边缘端的推理延迟直接决定了避障的成败，因此，评测体系应纳入每秒推理生成词元数、控制频率与端侧延迟等指标。

未来无人机视觉理解评测体系将以能力评估为关键动力，从开放世界泛化、空间推理、指令理解以及实时推理等多个维度系统衡量模型性能。构建无人机视觉理解评测体系的目标，应当是为研究人员提供一个衡量模型在真实空域下观察、思考以及行动一致性的度量。而 Embodied City（Gao 等，2024）等高保真模拟平台的出现，为这种能力评测提供了安全且可重复的数字孪生环境，使研究者能够在理解、问答、导航与任务规划等多个维度上对无人机智能体进行综合评估。通过以能力为中心的评测体系，将更好地推动无人机视觉系统向自主认知与具身智能方向发展，从而为下一代兼具安全性与可解释性的无人机具身智能奠定评估基础。

5 结 语

随着无人机技术与人工智能的深度融合，视觉理解范式正经历从特定任务驱动的局部建模向通用具身智能驱动的全局闭环推理的深刻变革。本文系统梳理了无人机场景下视觉理解大模型的研究进展，探讨了其从基础感知、语义推理到决策规划的全链条赋能路径。视觉理解大模型为无人机打破传统视觉系统的局限性提供了全新思路。在感知层面，通过引入视觉基础模型与多模态先验，无人机实现了从闭集识别向开放词汇感知的跨越，极大地增强了在非结构化环境中的泛化能力与细粒度检测精度；在语义推理层面，大语言模型与多模态大模型的融合，赋予了无人机对复杂事件的常识推理、空间关系理解及任务级人机交互能力；在决策规划层面，视觉—语言—行动一体化架构正在重构无人机的作业模式，实现了从高层指令到低层飞行控制的直接映射，推动无人机从被动观测者向自主具身智能体的进化。尽管大模型赋能无人机展现出巨大潜力，但要实现高度自主、透明可信的空中智能系统，仍需在以下方面深入探索：

1）构建无人机场景通用视觉大模型。当前视觉基础模型多基于互联网图文数据训练，与航拍视觉在视角结构、小目标分布以及空间语义表达方面存在明显差异，这在第 3 节中表现为小目标识别困难与跨域泛化能力不足。未来需进一步构建覆盖多任务、多模态的大规模航拍专用数据集，研发面向空域场景的无人机通用视觉基座模型，使无人机能够在开放环境中进行跨任务迁移和零样本推理，为下游无人机应用奠定技术基石。同时，第 4 节分析表明当前无人机视觉评测仍偏向单任务指标，统一视觉感知、空间理解与事件推理的综合评测体系也将成为推动模型能力发展的关键基础设施。

2）实现具身无人机智能与闭环控制。现有多模态模型虽然具备较强的语义理解能力，但其推理结果与低层飞行控制之间仍存在语义—动作鸿沟。VLA 模型为弥合这一问题提供了新的研究方向，但如何在动态环境中实现稳定可靠的闭环控制仍是关键挑战。未来研究需要进一步探索语言驱动的飞行策略学习，将高层语义规划与无人机动力学模型、轨迹优化算法以及安全约束机制进行深度融合，从而实现更加稳健的避障、导航与任务执行能力。同时，在多无人机系统中，还需要解决语义信息共享与协同决策问题，通过多智能体学习与任务分解机制提升无人机群体协同能力，从而实现复杂环境中的协同搜索、协同巡检与群体决策。

3）实时推理与轻量化部署技术。无人机边缘计算平台的资源受限与大模型巨大的计算开销之间存在显著矛盾。尽管已有工作通过网络剪枝、低比特量化、知识蒸馏及参数高效微调等方法在一定程度上降低了模型复杂度，但这些方法多集中于单模型层面的优化，尚未从系统层面解决感知—推理—控制闭环中的时延累积与资源分配问题。未来需突破模型压缩、量化、蒸馏以及如 LoRA 等轻量化微调技

术，优化端侧推理延迟，研发软硬件协同优化方案，以实现大模型在嵌入式设备上的高效运行。同时，通过云边协同计算、动态推理调度以及多模型协同机制，从模型级压缩向系统级协同优化转变，从而实现计算负载的动态分解与调度，在保证实时性的同时维持高质量视觉理解能力，从而实现复杂任务的在线决策。

4）安全性、隐私保护与伦理准则。大模型固有的幻觉问题在安全敏感的航空任务中具有巨大风险。未来需要加强模型可解释性研究，构建可验证的决策推理机制，并结合规则约束与安全验证模块，提升系统的可靠性。此外，无人机在采集大量视觉数据时可能涉及隐私保护与数据安全问题，因此需要在数据采集、存储与模型训练过程中引入隐私保护机制与安全策略，以确保技术应用符合伦理与法规要求。

参考文献（References）

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. 2024. GPT-4 Technical Report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2303.08774.pdf

Ahmmad S, Aditto Z A, Hossain M M, Yeasmin N and Hossain S. 2025. Autonomous navigation of cloud-controlled quadcopters in confined spaces using multi-modal perception and LLM-driven high semantic reasoning [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2508.07885.pdf

Anil R, Borgeaud S, Alayrac J B, Yu J H, Soricut R, Schalkwyk J, et al. 2025. Gemini: a family of highly capable multimodal models [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2312.11805.pdf

Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. 2023. Qwen Technical Report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2309.16609.pdf

Bai S, Cai Y X, Chen R Z, Chen K Q, Chen X H, Cheng Z S, et al. 2025. Qwen3-VL technical report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2511.21631.pdf

Bashmal L, Bazi Y, Al Rahhal M M, Zuair M and Melgani F. 2023. CapERA: captioning events in aerial videos. Remote Sensing, 15(8): #2139 [DOI: 10.3390/rs15082139]

Bazi Y, Bashmal L, Al Rahhal M M, Ricci R and Melgani F. 2024. RS-LLaVA: a large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing, 16(9): #1477 [DOI: 10.3390/rs16091477]

Bellec Y V. 2026. DroneScan-YOLO: redundancy-aware lightweight detection for tiny objects in UAV imagery [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2604.13278.pdf

Blei Y, Krawez M, Nilavadi N, Kaiser T K and Burgard W. 2025. CloudTrack: scalable UAV tracking with cloud semantics//Proceedings of 2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE: 15893-15899 [DOI: 10.1109/ICRA55743.2025.11128514]

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, et al. 2020. Language models are few-shot learners//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: ACM: #159

Cai H X, Rao Y J, Huang L G, Zhong Z Y, Dong J H, Tan J J, et al. 2026. AirNav: a large-scale UAV vision-and-language navigation dataset with natural and diverse instructions [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2601.03707.pdf

Cai S W, Wu Y W and Zhou L F. 2025a. LLM-land: large language models for context-aware drone landing [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2505.06399.pdf

Cai Z, Cao M S, Chen H J, Chen K, Chen K Y, Chen X, et al. 2024. InternLM2 Technical Report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/17297.pdf

Cai Z X, Cardenas C R, Leo K, Zhang C Y, Backman K, Li H B, et al. 2025b. NEUSIS: a compositional neuro-symbolic framework for autonomous perception, reasoning, and planning in complex UAV search missions. IEEE Robotics and Automation Letters, 10(9): 9502-9509 [DOI: 10.1109/LRA.2025.3592098]

Chen G J, Yu X J, Ling N W and Zhong L. 2024a. TypeFly: flying drones with large language model [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2312.14950.pdf

Chen J Y, Li H Y, Tang Z H, Li X D, Wu W J and Liu S. 2026. AerialVLA: a vision-language-action model for aerial navigation with online dialogue//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI: 18161-18169 [DOI: 10.1609/aaai.v40i22.38878]

Chen Y, Li Z Y, Lan B L, Liu Y F and Zheng T. 2025. Design of UAV visual-language collaborative search and rescue system based on visual grounding//Proceedings of the 10th International Conference on Computer and Information Processing Technology. Fushun, China: IEEE: 618-622 [DOI: 10.1109/ISCIPT67144.2025.11265419]

Chen Z, Li J Y, Fukumoto F, Liu P and Suzuki Y. 2023. 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, China: ACM: #76 [DOI: 10.1145/3595916.3626450]

Chen Z, Wu J N, Wang W H, Su W J, Chen G, Xing S, et al. 2024b. Intern VL: scaling up vision foundation models and aligning for generic visual-linguistic tasks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24185-24198. [DOI: 10.1109/CVPR52733.2024.02283]

Cui J Q, Liu G C, Wang H, Yu Y and Yang J K. 2024. TPML: task planning for multi-UAV system with large language models//Proceedings of 2024 IEEE 18th International Conference on Control and Automation. Reykjavík, Iceland: IEEE: 886-891 [DOI: 10.1109/ICCA62789.2024.10591846]

Damoc R M and Dobrea D M. 2025. Visual description generation from UAV-captured images using a large language model//Proceedings of 2025 International Symposium on Signals, Circuits and Systems. Iasi, Romania: IEEE: 1-4 [DOI: 10.1109/ISSCS66034.2025.11105649]

De Curtò J, De Zarzà I and Calafate C T. 2023. Semantic scene understanding with large language models on unmanned aerial vehicles. Drones, 7(2): #114 [DOI: 10.3390/drones7020114]

De Zarzà I, De Curtò J and Calafate C T. 2023. Socratic video understanding on unmanned aerial vehicles. Procedia Computer Science, 225: 144-154 [DOI: 10.1016/j.procs.2023.09.101]

Döschl B and Kiam J J. 2024. Say-REAPEx: an LLM-modulo UAV online planning framework for search and rescue//Proceedings of the 2nd CoRL Workshop on Learning Effective Abstractions for Planning. Munich, Germany: [s.n.]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net: [s.n.]

Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et al. 2023. PALM-E: an embodied multimodal language model//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #340

Du D W, Qi Y K, Yu H Y, Yang Y F, Duan K W, Li G R, et al. 2018. The unmanned aerial vehicle benchmark: object detection and tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 375-391 [DOI: 10.1007/978-3-030-01249-6_23]

Eesaar H, Ahmed A, Farhan M, Chong K T, Lee D J and Tayara H. 2025. Multi-modal autonomous drone system for real-time highway incident detection and analysis. IEEE Access, 13: 183314-183329 [DOI: 10.1109/ACCESS.2025.3623653]

EL Amraoui K, Lghoul M, Ezzaki A, Masmoudi L, Hadri M, Elbelrhiti H, et al. 2022. Avo-AirDB: an avocado UAV Database for agricultural image segmentation and classification. Data in Brief, 45: #108738 [DOI: 10.1016/j.dib.2022.108738]

Fang Z Y, Yu B, Liu C, Yang Z Y, Chen R Q, Lin Y X, et al. 2026. Reasoning knowledge-gap in drone planning via LLM-based active elicitation [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2603.07824.pdf

Ferrag M A, Lakas A and Debbah M. 2025. UAVBench: an open benchmark dataset for autonomous and agentic AI UAV systems via LLM-generated flight scenarios [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2511.11252.pdf

Florea H and Nedevschi S. 2025. TanDepth: leveraging global DEMs for metric monocular depth estimation in UAVs. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 5445-5459 [DOI: 10.1109/JSTARS.2025.3531984]

Gao C, Zhao B N, Zhang W C, Mao J Z, Zhang J, Zheng Z H, et al. 2024. EmbodiedCity: a benchmark platform for embodied agent in real-world city environment [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2410.09604.pdf

Gao Y P, Li C H, You Z R, Liu J L, Li Z, Chen P A, et al. 2025a. Openfly: a comprehensive platform for aerial vision-language navigation [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2502.18041.pdf

Gao Y P, Wang Z G, Han P F, Jing L L, Wang D and Zhao B. 2025b. Exploring spatial representation to enhance LLM reasoning in aerial vision-language navigation [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2410.08500.pdf

Gong Z Y, Wei Z X, Wang D, Hu X X, Ma X Z, Chen H R X, et al. 2026. CrossEarth: geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(5): 5147-5164 [DOI: 10.1109/TPAMI.2025.3649001]

Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. 2024. The llama 3 herd of models [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2407.21783.pdf

Guo D Y, Yang D J, Zhang H W, Song J X, Wang P Y, Zhu Q H, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 645(8081): 633-638 [DOI: 10.1038/s41586-025-09422-z]

Hong H D, Wang S, Huang Z, Wu Q and Liu J J. 2024. Why only text: empowering vision-and-language navigation with multi-modal prompts//Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, Korea (South): IJCAI.org: 839-847 [DOI: 10.24963/ijcai.2024/93]

Hurst A, Lerer A, Goucher A P, Perelman A, Ramesh A, Clark A, et al. 2024. GPT-4o system card [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2410.21276.pdf

Jaisawal P K, Papakonstantinou S and Gollnick V. 2024. AirFisheye dataset: a multi-model fisheye dataset for UAV applications//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 11818-11824 [DOI: 10.1109/ICRA57147.2024.10611092]

Ji Y T, Zhu Z Q, Zhao Y, Liu B D, Gao C, Zhao Y H, et al. 2026. Towards autonomous UAV visual object search in city space: benchmark and agentic methodology//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore, Singapore: AAAI: 18342-18350 [DOI: 10.1609/aaai.v40i22.38898]

Jiang Y F, Gupta A, Zhang Z C, Wang G Z, Dou Y Q, Chen Y J, et al. 2023. VIMA: robot manipulation with multimodal prompts//Pro-

ceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #611

Joshi A, Sanyal S and Roy K. 2025. Neuro-LIFT: a neuromorphic, LLM-based interactive framework for autonomous drone FlighT at the edge//Proceedings of 2025 International Joint Conference on Neural Networks. Rome, Italy: IEEE: 1-9 [DOI: 10.1109/IJCNN64981.2025.11228467]

Kim H, Lee D, Park S and Ro Y M. 2024a. Weather-aware drone-view object detection via environmental context understanding//Proceedings of 2024 IEEE International Conference on Image Processing. Abu Dhabi, United Arab Emirates: IEEE: 549-555 [DOI: 10.1109/ICIP51287.2024.10647388]

Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. 2024b. OpenVLA: an open-source vision-language-action model//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 2679-2713

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Koubaa A and Gabr K. 2025. Agentic UAVs: LLM-driven autonomy with integrated tool-calling and cognitive reasoning [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2509.13352.pdf

Lam D, Kuzma R, McGee K, Dooley S, Laielli M, Klaric M, et al. 2018. xView: objects in context in overhead imagery [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/1802.07856.pdf

Lee J, Miyanishi T, Kurita S, Sakamoto K, Azuma D, Matsuo Y, et al. 2025. CityNav: a large-scale dataset for real-world aerial navigation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 5912-5922 [DOI: 10.1109/ICCV51701.2025.00559]

Leng J X, Mo M J C, Zhou Y H, Ye Y M, Gao C Q and Gao X B. 2023. Recent advances in drone-view object detection. Journal of Image and Graphics, 28(9): 2563-2586 (冷佳旭, 莫梦竟成, 周应华, 叶永明, 高陈强, 高新波. 2023. 无人机视角下的目标检测研究进展. 中国图象图形学报, 28(9): 2563-2586) [DOI: 10.11834/jig.220836]

Li H Y, Liu X Y and Li G R. 2024. A benchmark for UAV-view natural language-guided tracking. Electronics, 13(9): #1706 [DOI: 10.3390/electronics13091706]

Li J N, Li D X, Xiong C M and Hoi S C H. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li T J, Liu J, Zhang W, Ni Y, Wang W Q and Li Z H. 2021. UAV-human: a large benchmark for human behavior understanding with unmanned aerial vehicles//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16261-16270 [DOI: 10.1109/CVPR46437.2021.01600]

Li Z X, Qian R R, Qi Y, Wang C F and Su H. 2026. Intent-driven cooperative control of UAV swarms: an LLM-based approach. Applied Sciences, 16(7): #3297 [DOI: 10.3390/app16073297]

Limberg C, Gonçalves A, Rigault B and Prendinger H. 2024. Leveraging YOLO-world and GPT-4V LMMs for zero-shot person detection and action recognition in drone imagery [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2404.01571.pdf

Lin F, Tian Y L, Wang Y Z, Zhang T C, Zhang X Y and Wang F Y. 2024. AirVista: empowering UAVs with 3D spatial reasoning abilities through a multimodal large language model agent//Proceedings of the 27th IEEE International Conference on Intelligent Transportation Systems. Edmonton, Canada: IEEE: 476-481 [DOI: 10.1109/ITSC58415.2024.10919532]

Liu A X, Feng B, Xue B, Wang B X, Wu B C, Lu C D, et al. 2025. DeepSeek-V3 Technical Report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2412.19437.pdf

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1516

Liu J, Cui J Z, Ye M, Zhu X T and Tang S. 2024a. Shooting condition insensitive unmanned aerial vehicle object detection. Expert Systems with Applications, 246: #123221 [DOI: 10.1016/j.eswa.2024.123221]

Liu Y T, Zhou Z H, Liu J W, Chen L M and Wang J K. 2024b. Multi-agent formation control using large language models [EB/OL]. [2026-04-05]. https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.172954477.70259514/v1

Liu Y Z, Yao F L, Yue Y C, Xu G L, Sun X and Fu K. 2024c. NavAgent: multi-scale urban street view fusion for UAV embodied vision-and-language navigation [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2411.08579.pdf

Ma Z H, Li Y S, Ma R G and Liang C. 2024. Applying unsupervised semantic segmentation to high-resolution UAV imagery for enhanced road scene parsing [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2402.02985.pdf

Mou L C, Hua Y S, Jin P and Zhu X X. 2020. ERA: a data set and deep learning benchmark for event recognition in aerial videos [software and data sets]. IEEE Geoscience and Remote Sensing Magazine, 8(4): 125-133 [DOI: 10.1109/MGRS.2020.3005751]

O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, et al. 2024. Open X-embodiment: robotic learning datasets and RT-X models: open X-embodiment collaboration[0]//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 6892-6903 [DOI: 10.1109/ICRA57147.2024.10611477]

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P,

et al. 2022. Training language models to follow instructions with human feedback//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #2011

Pan L Z, Song C T, Gan X Z, Xu K Y and Xie Y. 2024. Military image captioning for low-altitude UAV or UGV perspectives. Drones, 8(9): #421 [DOI: 10.3390/drones8090421]

Phadke A, Hadimlioglu A, Chu T X and Sekharan C N. 2024. Integrating large language models for UAV control in simulated environments: a modular interaction approach [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2410.17602.pdf

Pueyo P, Montijano E, Murillo A C and Schwager M. 2024. CLIP-Swarm: generating drone shows from text prompts with vision-language models//Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi, United Arab Emirates: IEEE: 11917-11923 [DOI: 10.1109/IROS58592.2024.10801327]

Qiu H J, Li J Q, Gan J H, Zheng S W and Yan L Q. 2024. DroneGPT: zero-shot video question answering for drones//Proceedings of 2024 International Conference on Computer Vision and Deep Learning. Changsha, China: ACM: #49 [DOI: 10.1145/3653804.3654608]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Rahnemoonfar M, Chowdhury T, Sarkar A, Varshney D, Yari M and Murphy R R. 2021. FloodNet: a high resolution aerial imagery dataset for post flood scene understanding. IEEE Access, 9: 89644-89654 [DOI: 10.1109/ACCESS.2021.3090981]

Ravichandran Z, Murali V, Tzes M, Pappas G J and Kumar V. 2025. SPINE: online semantic planning for missions with incomplete natural language specifications in unstructured environments//Proceedings of 2025 IEEE International Conference on Robotics and Automation. Atlanta, USA: IEEE: 13714-13721 [DOI: 10.1109/ICRA55743.2025.11128238]

Rizzoli G, Barbato F, Caligiuri M and Zanuttigh P. 2023. SynDrone — multi-modal UAV dataset for urban scenarios//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 2202-2212 [DOI: 10.1109/ICCVW60793.2023.00235]

Sabaghian M, Keyvanrad M A and Moghadami S M. 2025. A novel compression framework for YOLOv8: achieving real-time aerial object detection on edge devices via structured pruning and channel-wise distillation [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2509.12918.pdf

Samma H and El-Ferik S. 2025. UAV visual path planning using large language models. Transportation Research Procedia, 84: 339-345 [DOI: 10.1016/j.trpro.2025.03.081]

Sanyal S and Roy K. 2025. ASMA: an adaptive safety margin algorithm for vision-language drone navigation via scene-aware control barrier functions. IEEE Robotics and Automation Letters, 10(9): 9232-9239 [DOI: 10.1109/LRA.2025.3592138]

Sautenkov O, Yaqoot Y, Lykov A, Mustafa M A, Tadevosyan G, Akhmetkazy A, et al. 2025a. UAV-VLA: vision-language-action system for large scale aerial mission generation//Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction. Melbourne, Australia: IEEE: 1588-1592 [DOI: 10.1109/HRI61500.2025.10974117]

Sautenkov O, Yaqoot Y, Mustafa M A, Batool F, Sam J, Lykov A, et al. 2025b. UAV-CodeAgents: scalable UAV mission planning via multi-agent ReAct and vision-language reasoning [EB/OL]. [2026-04-05]. https://arxiv.org/pdf/2505.07236.pdf

Sezgin A, Sezgin B D and Keskin R. 2025. Semantic object understanding in UAV operations using vision-language models//Proceedings of the 12th International Conference on Future Internet of Things and Cloud. Istanbul, Turkiye: IEEE: 374-380 [DOI: 10.1109/FiCloud66139.2025.00057]

Silva A L B V E, De Castro Felix H, Simões F P M, Teichrieb V, Dos Santos M, Santiago H, et al. 2023. InsPLAD: a dataset and benchmark for power line asset inspection in UAV images. International Journal of Remote Sensing, 44(23): 7294-7320 [DOI: 10.1080/01431161.2023.2283900]

Tazir M L, Mancas M and Dutoit T. 2023. From words to flight: integrating OpenAI ChatGPT with PX4/gazebo for natural language-based drone control//Proceedings of the 13th International Workshop on Computer Science and Engineering. Singapore, Singapore: WCSE: 215-222 [10.18178/wcse.2023.06.031]

Tetila E C, Moro B L, Astolfi G, Da Costa A B, Amorim W P, Belete N A De S, et al. 2024. Real-time detection of weeds by species in soybean using UAV images. Crop Protection, 184: #106846 [DOI: 10.1016/j.cropro.2024.106846]

Tian Y L, Lin F, Li Y D, Zhang T C, Zhang Q Y, Fu X, et al. 2025. UAVs meet LLMs: overviews and perspectives towards agentic low-altitude mobility. Information Fusion, 122: #103158 [DOI: 10.1016/j.inffus.2025.103158]

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. 2023. LLaMA: open and efficient foundation language models [EB/OL]. [2026-04-05]. https://arxiv.org/pdf 2302.13971.pdf

Wang P, Shuai Z H, Li Q C, Wang K, Liu L Y, Ye F F, et al. 2025. Lmucs: lightweight LLM-driven UAV control system with multimodal perception for autonomous material deliver. Rochester, NY, USA: Social Science Research Network: #5397714 [DOI: 10.2139/ssrn.5397714]

Wang X, Shu X J, Zhang Z P, Jiang B, Wang Y W, Tian Y H, et al. 2021. Towards more flexible and accurate object tracking with natural language: algorithms and benchmark//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13763-13773 [10.1109/CVPR46437.

2021.01355]

Wang Z Y, Du C X and Liu Y. 2026. A survey of vision-language navigation models for UAVs: from perceptual comprehension to intelligent decision-making. Journal of South China University of Technology (Natural Science Edition), 54(6): 173-182 (王子豫, 杜宸旭, 刘洋. 2026. 无人机视觉语言导航模型综述: 从感知理解到智能决策. 华南理工大学学报(自然科学版), 54(6): 173-182) [DOI: 10.12141/j.issn.1000-565X.260003]

Wu W T, Li C L, Wang X, Luo B and Liu Q. 2025. Large language model guided progressive feature alignment for multimodal UAV object detection [EB/OL]. [2026-04-05]. https://arxiv.org/pdf 2503.06948.pdf

Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, et al. 2018. DOTA: a large-scale dataset for object detection in aerial images// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3974-3983 [DOI: 10.1109/CVPR.2018.00418]

Xu Y Y, Du H J and Guo S W. 2025. Research progress on embodied navigation of low-altitude UAV. Aerospace Control, 43(4): 7-14 (许越越, 杜华军, 郭尚伟. 2025. 低空无人机具身导航研究进展. 航天控制, 43(4): 7-14) [DOI: 10.3969/j.issn.1006-3242.2025.04.002]

Yan L, Liao X H, Zhou C H, Fan B K, Gong J Y, Cui P, et al. 2019. The impact of UAV remote sensing technology on the industrial development of China: a review. Journal of Geo-information Science, 21(4): 475-495 (晏磊, 廖小罕, 周成虎, 樊邦奎, 龚健雅, 崔鹏, 等. 2019. 中国无人机遥感技术突破与产业发展综述. 地球信息科学学报, 21(4): 475-495) [DOI: 10.12082/dqxxkx.2019.180589]

Yang A, Li A F, Yang B S, Zhang B C, Hui B Y, Zheng B, et al. 2025a. Qwen3 Technical Report [EB/OL]. [2026-04-05]. https://arxiv.org/pdf 2505.09388.pdf

Yang T T, Feng P, Guo Q X, Zhang J D, Zhang X F, Ning J H, et al. 2025b. AutoHMA-LLM: efficient task coordination and execution in heterogeneous multi-agent systems using hybrid large language models. IEEE Transactions on Cognitive Communications and Networking, 11(2): 987-998 [DOI: 10.1109/TCCN.2025.3528892]

Yao F L, Yue Y C, Liu Y Z, Sun X and Fu K. 2025a. AeroVerse: UAV-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models [EB/OL]. [2026-04-05]. https://arxiv.org/pdf 2408.15511.pdf

Yao L, Liu F, Zhang C Y, Ou Z Q and Wu T. 2025b. Domain-invariant progressive knowledge distillation for UAV-based object detection. IEEE Geoscience and Remote Sensing Letters, 22: 1-5 [DOI: 10.1109/LGRS.2024.3492187]

Yuan C S, Zhou Y H, Guo C H, Han D J, Shi G and Wang W W. 2026. Seeing with words: interpretable language-guided drone geo-localization via LLM-enriched semantic attribute alignment. IEEE Transactions on Multimedia, 28: 2132-2144 [DOI: 10.1109/TMM.2025.3642913]

Yuan Z H, Xie F F and Ji T W. 2024. Patrol agent: an autonomous UAV framework for urban patrol using on board vision language model and on cloud large language model//Proceedings of the 6th International Conference on Robotics and Computer Vision. Wuxi, China: IEEE: 237-242 [DOI: 10.1109/ICRCV62709.2024.10758606]

Zhan Y, Xiong Z T and Yuan Y. 2025. SkyEyeGPT: unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS Journal of Photogrammetry and Remote Sensing, 221: 64-77 [DOI: 10.1016/j.isprsjprs.2025.01.020]

Zhang C H, Huang G J, Liu L, Huang S, Yang Y N, Wan X, et al. 2023a. 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 G X, Liu Y M, Yang X Y, Huang H L and Huang C. 2024a. TrafficNight: an aerial multimodal benchmark for nighttime vehicle surveillance//Proceeding of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 36-48 [DOI: 10.1007/978-3-031-73650-6_3]

Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. 2023b. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview. net: [s.n.]

Zhang J Z, Wang K K, Xu R T, Zhou G Z, Hong Y C, Fang X M, et al. 2024b. NaVid: video-based VLM plans the next step for vision-and-language navigation//Robotics: Science and Systems XX. Delft, the Netherlands: [s.n.]

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]. [2026-04-05]. https://arxiv.org/pdf 2511.13269.pdf

Zhang P Y, Zhong Y X and Li X Q. 2019. SlimYOLOv3: narrower, faster and better for real-time UAV applications//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 37-45 [DOI: 10.1109/ICCVW.2019.00011]

Zhang Y, Zhao Z C, Luo Z, Li C L and Tang J. 2026. UAV traffic scene understanding: a regulation embedded multi-modal network and a unified benchmark [EB/OL]. [2026-04-05]. https://arxiv.org/pdf 2603.10722.pdf

Zhang Z L, Zhao T C, Guo Y L and Yin J W. 2024c. RS5M and GeoRSCLIP: a large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-23 [DOI: 10.1109/TGRS.2024.3449154]

Zhao H R, Pan F X, Ping H Q Y and Zhou Y M. 2023. Agent as cerebrum, controller as cerebellum: implementing an embodied LMM-

based agent on drones［EB/OL］.［2026-04-05］. https://arxiv.org/pdf 2311.15033.pdf

Zhao J and Lin X. 2025. General-purpose aerial intelligent agents empowered by large language models［EB/OL］.［2026-04-05］. https://arxiv.org/pdf 2503.08302.pdf

Zhong J G, Li M, Chen Y L, Wei Z H, Yang F and Shen H R. 2024. A safer vision-based autonomous planning system for quadrotor UAVs with dynamic obstacle trajectory prediction and its application with LLMs//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa, USA: IEEE: 920-929［DOI: 10.1109/WACVW60836.2024.00131］

Zhu P F, Wen L Y, Du D W, Bian X, Fan H, Hu Q H, et al. 2022. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11): 7380-7399［DOI: 10.1109/TPAMI.2021.3119563］

Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. 2023. RT-2: vision-language-action models transfer web knowledge to robotic control//Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR: 2165-2183

作者简介

余雅婷，女，博士研究生，主要研究方向为多模态视频理解。
E-mail：yatingyu@mail.nwpu.edu.cn

曹聪琦，通信作者，女，副教授，主要研究方向为计算机视觉、智能视频理解、数据高效学习与时空预测。
E-mail：congqi.cao@nwpu.edu.cn

王昭颖，女，硕士研究生，主要研究方向为多模态视频理解。
E-mail：wangzhaoying@mail.nwpu.edu.cn

张艳宁，女，教授，主要研究方向为图像处理、计算机视觉、空间环境探测动态视觉计算理论与技术。
E-mail：ynzhang@nwpu.edu.cn