无人机遇见大模型

1 基础模型

1.1 VLMs

VLMs 是多模态模型，通过整合视觉和文本信息扩展了 LLMs 的能力[[234]]。这些模型旨在解决需要视觉和语言理解的一系列任务，比如视觉问答(VQA) 和图像标注 [[235], [236], [237], [238], [239]]。本节介绍了几种典型的 VLM 模型，重点介绍它们的技术特点和应用场景。

子类别	模型名称	机构 / 作者
通用	GPT-4V, GPT-4o, GPT-4o mini, GPT o1-preview	OpenAI
	Claude 3 Opus, Claude 3.5 Sonnet	Anthropic
	Step-2	Jieyue Xingchen
	LLaVA, LLaVA-1.5, LLaVA-NeXT	Liu et al.
	MoE-LLaVA	Lin et al.
	LLaVA-CoT	Xu et al.
	Flamingo	Alayrac et al.
	BLIP	Li et al.
	BLIP-2	Li et al.
	InstructBLIP	Dai et al.
视频理解	LLaMA-VID	Li et al.
	IG-VLM	Kim et al.
	Video-ChatGPT	Maaz et al.
	VideoTree	Wang et al.
视觉推理	X-VLM	Zeng et al.
	Chameleon	Lu et al.
	HYDRA	Ke et al.
	VISPROG	PRIOR @ Allen Institute for AI

OpenAI 的 GPT-4V [[167]] 是视觉语言模型 (VLMs) 中的一个重要代表，展现了强大的视觉感知能力 [[240]]。升级版的 GPT-4o 引入了更先进的优化算法，使其能够接受任意组合的文本、音频和图像输入，同时提供快速响应。轻量级版本 GPT-4o mini 专为移动设备和边缘计算场景设计，通过减少计算资源消耗，平衡高效性能与可部署性 [[241]]。GPT o1-preview 在推理方面表现出色，尤其是在编程和解决复杂问题上 [[242]]。Anthropic 的 Claude 3 Opus 展现了强大的多任务泛化和可控性，而 Claude 3.5 Sonnet 通过优化推理速度和成本效率提升了实际价值 [[168]]。Step-2 模型采用了创新的专家混合 (Mixture of Experts, MoE) 架构，支持在万亿参数规模上进行高效训练，并显著改善复杂任务的处理和模型的可扩展性。

刘等 [[169]] 提出了 LLaVA，这是一种代表性的视觉语言模型（VLM）。该模型利用 GPT-4 生成遵循指令的数据集，并集成了 CLIP 视觉编码器 ViT-L/14 [[186]] 与 Vicuna [[243]]，通过端到端的微调提升其在多模态任务中的表现。其最新版本 LLaVA-NeXT [[171]] 在 LLaVA-1.5 [[170]] 的基础上进行了显著改进，特别是增强了捕捉视觉细节的能力，并在复杂的视觉和逻辑推理任务中表现突出。MoE-LLaVA 用 MoE 架构替代 LLaVA 中的语言模型，显著提高了大型多任务场景下的推理效率和资源利用率 [[172]]。LLaVA-CoT 通过对大规模视觉问答样本的结构化推理注释与束搜索方法结合，提升了在推理密集任务中的准确性 [[173]]。 另一个重要的架构类别包括 Flamingo [[174]] 和 BLIP 系列 [[175], [176]]，它们通过将预训练的视觉特征编码器与预训练的 LLMs 相结合，使得 LLMs 能够从多模态输入生成相应的文本输出。Flamingo 引入了感知重采样器和门控跨注意力机制，有效地将视觉、多模态信息与语言模型集成，从而显著提高了多模态任务的性能。BLIP-2 [[176]] 采用了一种结合逐阶段冻结图像编码器与 LLMs 的预训练策略，并引入了查询变换器（Q-Former），有效解决了视觉和语言模态之间的对齐问题。InstructBLIP [[177]] 结合了大规模任务指导微调机制，进一步提高了模型对多模态任务的适应性。

此外，视觉语言模型（VLMs）在各种任务和场景中展示了广泛的应用潜力。在视频理解方面，代表性模型如 LLaMA-VID [[178]]、IG-VLM [[179]]、Video-ChatGPT [[180]] 和 VideoTree [[181]] 在视频内容分析和多模态任务中表现出色。在视觉推理方面，X-VLM [[182]]、Chameleon [[183]]、HYDRA [[184]]

和 VISPROG [[185]] 通过创新的架构设计和推理机制提高了复杂视觉推理任务的准确性和适应性。

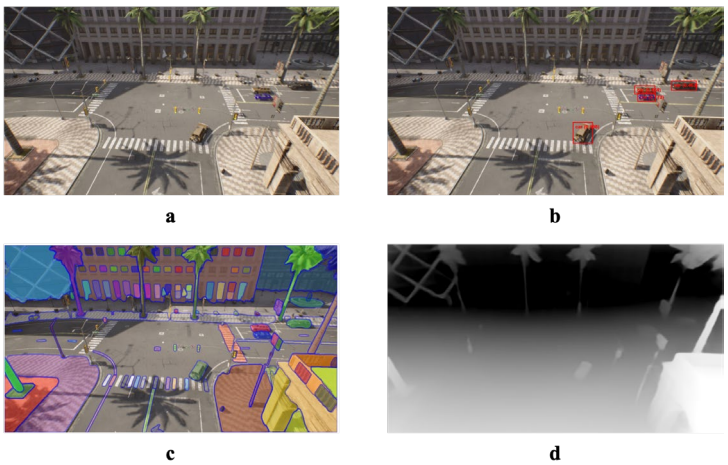
1.2 视觉特征模型（VFM）

近年来，视觉特征模型（VFM）作为计算机视觉中的核心技术概念应运而生。VFM 的主要目标是提取多样且高度表达性的图像特征，使其能够直接应用于各种下游任务。这些模型通常以大规模参数、显著的泛化能力和出色的跨任务迁移性能为特征，尽管其训练成本相对较高 [[194]]。CLIP 是 VFM 领域的一个开创性代表。通过对大规模图像-文本对进行弱监督训练，它高效地对齐了视觉和文本嵌入，为多模态学习打下了坚实的基础 [[186]]。后续的研究进一步提高了 CLIP 的训练效率和性能，包括 FILIP [[187]]、RegionCLIP [[188]] 和 EVA-CLIP [[189]] 等模型。

子类别	模型名称	机构 / 作者
通用	CLIP	OpenAI
	FILIP	Yao et al.
	RegionCLIP	Microsoft Research
	EVA-CLIP	Sun et al.
目标检测	GLIP	Microsoft Research
	DINO	Zhang et al.
	Grounding-DINO	Liu et al.
	DINOv2	Meta AI Research, FAIR}
	AM-RADIO	NVIDIA
	DINO-WM	Zhou et al.
	YOLO-World	Cheng et al.
图像分割	CLIPSeg	Lüdecke and Ecker
	SAM	Meta AI Research, FAIR
	Embodied-SAM	Xu et al.
	Point-SAM	Zhou et al.
	Open-Vocabulary SAM	Yuan et al.

子类别	模型名称	机构 / 作者
	TAP	Pan et al.
	EfficientSAM	Xiong et al.
	MobileSAM	Zhang et al.
	SAM 2	Meta AI Research, FAIR
	SAMURAI	University of Washington
	SegGPT	Wang et al.
	Osprey	Yuan et al.
	SEEM	Zou et al.
	Seal	Liu et al.
	LISA	Lai et al.
深度估计	ZoeDepth	Bhat et al.
	ScaleDepth	Zhu et al.
	Depth Anything	Yang et al.
	Depth Anything V2	Yang et al.
	Depth Pro	Apple

VFMs 展现了卓越的适应性，在多种计算机视觉任务中取得了显著成果，包括零-shot 目标检测、图像分割和深度估计。如图 [3] 所示，我们从 SynDrone 数据集中 Town10HD 场景中选择了一张样本图像 [[244]]，该数据集特定于无人机领域，以直观展示多个 VFMs 在零-shot 条件下的性能。这个例子为理解它们的实际应用潜力提供了有力支持。



a b c d

1.2.1 用于目标检测的 VFM

VFM 在目标检测中的核心优势在于其强大的 zero-shot 检测能力。GLIP [[190]] 统一了目标检测和短语定位任务，在各种对象级识别任务中展现出卓越的 zero-shot 和少-shot 迁移能力。Zhang 等. [[191]] 提出了 DINO，优化了 DETR 模型的架构 [[245]]，显著提升了检测性能和效率。后续工作 Grounding-DINO [[192]] 引入了文本监督以提高准确性。此外，DINOv2 [[193]] 采用了判别自监督学习的方法，使得能够提取强健的图像特征，并在下游任务中实现了优异的性能，而无需微调。 AM-RADIO [[194]] 通过多教师蒸馏方法整合了 CLIP [[186]]、DINOv2 [[193]] 和 SAM [[198]] 等 VFMs 的能力，从而产生了强大的表示能力，以支持复杂的视觉任务。DINO-WM [[195]] 将 DINOv2 集成到世界模型中，使零 shot 规划能力成为可能。此外，YOLO-World [[196]] 通过高效的预训练方案增强了 YOLO 检测器的泛化能力，在开放词汇和零 shot 检测任务中取得了卓越的性能。

1.2.2 图像分割的变分自由能

VFMs 在图像分割任务中相较于传统方法展现了显著的改进。Lüdecke 等 [[197]] 提出了基于 CLIP 模型的 CLIPSeg，该模型支持语义分割、实例分割和零样本分割。Kirillov 等[[198]] 开发了 Segment Anything Model (SAM)，通过在大规模和多样化数据集上进行预训练，实现了在多种场景下的零样本分割能力。后续研究进一步扩展了 SAM 的应用，如 Embodied-SAM [[199]] 和 Point-SAM [[200]]，扩展了 SAM 在三维场景中的功能。Open-Vocabulary SAM [[201]] 将 SAM 与 CLIP 的知识迁移策略相结合，有效地优化了分割和识别任务。Pan 等。[[202]] 提出了 TAP（Tokenize Anything），一个以视觉感知为中心的基础模型，通过引入视觉提示来改善 SAM 架构，从而能够同时完成任意区域的分割、识别和描述任务。EfficientSAM [[203]] 和 MobileSAM [[204]] 优化了 SAM 的表示，显著降低了模型复杂性，并在保持出色任务性能的同时实现了轻量化设计。最近，SAM 2 [[205]] 将内存模块引入到原始模型中，使得能够对任意长度的视频进行实时分割，同时解决了遮挡和多目标跟踪等复杂挑战。SAMURAI [[206]] 在 SAM 2 的基础上构建，通过集成卡尔曼滤波器，解决了 SAM 2 中内存管理的局限性，并实现了优越的视频分割性能，无需重训练或微调。

超越 SAM 系列，其他 VFM 架构也显著推进了图像分割。诸如 SegGPT [[207]]、Osprey [[208]] 和 SEEM [[209]] 等模型在任意分割任务和多模态场景中表现出了显著的适应性。此外，VFM 在其他分割任务中也显示出了重要的应用。例如，Liu 等 [[210]] 提出了 Seal 框架，用于对点云序列进行分割，而 LISA [[211]] 则采用了 Embedding-as-Mask 的方法，使多模态大模型具备基于推理的分割能力。LISA 能够处理复杂的自然语言指令并生成细粒度的分割结果，拓宽了分割模型应用的范围和复杂性。

1.2.3 单目深度估计的 VFM

在单目深度估计领域，VFM 也展示了显著的技术优势。ZoeDepth [[212]] 通过结合相对和绝对深度估计方法实现了零-shot 深度估计。ScaleDepth [[213]] 将深度估计分解为两个模块：场景规模预测和相对深度估计，在室内、室外、无约束和未见场景中实现了先进的性能。此外，Depth Anything [[214]] 利用许多未标记的单目图像训练出一种高效且鲁棒的深度估计方法，在零-shot 场景中展示了卓越的性能。Depth Anything V2 [[215]] 对原始模型进行了多项优化，进一步提升了复杂场景中的预测性能，并能够生成高质量的深度图像，细节丰富。基于多尺度 ViT 架构的 Depth Pro [[216]]，能够快速生成具有高分辨率和高频细节的度量准确深度图像，使其成为处理复杂深度估计任务的有效工具。

2 无人机的数据集和平台

本节回顾了与无人机研究相关的公开可用数据集和模拟平台，这些是推动基于基础模型（FM）无人机系统综合研究的重要资源。高质量的数据集构成了无人机视觉算法和自主行为学习的基础，通过提供多样化和全面的训练数据。与此同时，三维模拟平台为无人机系统的发展、测试和验证提供了安全和受控的虚拟环境。这些平台能够模拟复杂的场景和环境条件，使研究人员能够以无风险和成本效益高的方式进行实验。

我们收集了一系列开源数据集，这些数据集主要用于无人驾驶飞行器（UAV）系统的开发，所有数据集均已验证为可公开下载。

这些数据集涵盖多种格式，包括视频、RGB 图像（表格中的默认格式）、LiDAR 点云、红外图像、深度图像和文本数据（如描述或注释）。视频和 RGB

图像是主要的数据类型，而文本数据则较为少见。值得注意的是，一些数据集已更新以包含新功能。例如，EAR 数据集 [[246]] 增加了字幕和问答能力，演变为 CapEAR 数据集 [[247]]，现在适用于视觉问答（VQA）任务。表中列出的多数数据集均来自户外环境，并分为两类：一般领域数据集和特定领域数据集。

2.1 一般领域数据集

一般领域数据集旨在满足广泛场景的需求，进一步根据特定任务进行分类，包括环境感知、事件识别、目标跟踪、动作识别和导航。在环境感知类别中，我们专注于诸如目标检测、分割和深度估计等任务。虽然事件识别、目标跟踪和动作识别也可以视为环境感知的一部分，但为了更清晰地呈现数据集，我们将它们单独列出。

2.1.1 环境感知

这一部分展示了主要用于目标检测、分割和深度估计的数据集。例如，AirFisheye 数据集 [[248]] 专门设计用于 UAV 捕获的复杂城市环境中的目标检测、分割和深度估计等任务。它的多模态数据包括视觉、热成像和激光雷达，为分析这些复杂城市环境中的场景提供了全面的信息。SynDrone 数据集 [[244]] 是一个使用 Carla 生成的大规模合成数据集，旨在用于城市环境中的检测和分割任务。WildUAV 数据集 [[249]] 提供高分辨率 RGB 图像和深度真值数据，专注于单目视觉深度估计，同时支持在复杂环境中的精确无人机飞行控制。

名称	年份	类型	数量
AirFisheye	2024	鱼眼图像、深度图像、点云、IMU	总计超过 26,000 张鱼眼图像。数据以每秒 10 帧的速度采集。
SynDrone	2023	图像、深度图像、点云	包含 72,000 个标注样本，提供 28 种像素级和对象级标注。
WildUAV	2022	图像、视频、深度图像、元数据	测绘图像以 24 位 PNG 文件的形式提供，分辨率为 5280x3956。视频图像以 JPG 文件形式提供，分辨率为 3840x2160。详细说明了 16 种可能的类别标签。

2.1.2 事件识别

EAR 数据集 [[246]] 作为事件识别的视频基准，涵盖了 25 个事件类别，包括地震后、洪水、火灾、滑坡、泥石流、交通碰撞、交通拥堵、收割、耕作、建设、警方追捕、冲突、各种体育活动（如棒球、篮球、骑行）和社交活动（如聚会、音乐会、抗议、宗教活动）。该数据集包含 2864 个视频，每个视频时长为 5 秒，通过将“无人机”和“UAV”作为搜索关键词从 YouTube 收集而来。类似地，VIRAT 数据集 [[250]] 专注于监控视频中的事件识别，包括交通事故和人群聚集等事件。尽管并非由 UAV 捕获，VIRAT 数据集提供了类似的空中视角，使其与基于无人机的场景分析相关联。这些数据集为事件检测和场景理解的研究提供了宝贵的资源，特别是在将 UAV 应用与 LLMs 整合的背景下。

名称	年份	类型	数量
CapERA	2023	视频、文本	2864 个视频，每个有 5 个描述，总计 14,320 条文本。每个视频持续 5 秒，以 30 帧/秒的速度捕获，分辨率为 640 × 640 像素。
ERA	2020	视频	总共 2,864 个视频，包括灾难事件、交通事故、体育竞赛等 25 个类别。每个视频为 24 帧/秒，持续 5 秒。
VIRAT	2016	视频	25 小时的静态地面视频和 4 小时的动态空中视频。涉及 23 种事件类型。

2.1.3 对象跟踪

对象跟踪任务依赖于多样化的数据集，以推动各个领域的研究。WebUAV-3M 数据集 [[251]] 是一个大规模的无人机对象跟踪基准，包含 4,500 个视频和 233 个对象类别。它为一般场景中的无人机跟踪提供了坚实的基础，并包括多模态数据，例如音频和自然语言描述，使得多模态无人机跟踪方法的探索成为可能。TNL2K 数据集 [[254]] 专注于自然语言引导的对象跟踪，包含 2,000 个带有边界框标注和详细自然语言描述的视频序列，这些描述捕捉了目标对象的类别、形状、属性、特征和空间位置。为了应对挑战性的场景，TNL2K 包含对抗样本和外观变化显著的序列，提供 RGB 和红外模态以支持跨模态跟踪研究。VOT2020 数据集 [[256]] 提供了五个针对特定任务的专业数据集的全面集合：短期跟踪、实时跟踪、长期跟踪、热成像跟踪和深度跟踪。这些数据集共同应对了广泛的跟踪挑战，促进了不同跟踪范式的创新。

名称	年份	类型	数量
WebUAV-3M	2024	视频、文本、音频	4,500 个视频，总计超过 330 万帧，包含 223 个目标类别，提供自然语言和音频描述。
UAVDark135	2022	视频	135 个视频序列，包含超过 125,000 帧手动标注的图像。
DUT-VTUAV	2022	RGB-T 图像	近 170 万对对齐良好的可见光-热红外 (RGB-T) 图像对，包含 500 个序列，用于揭示 RGB-T 跟踪的潜力。涵盖 13 个子类别和跨 2 个城市的 15 个场景。
TNL2K	2022	视频、红外视频、文本	2,000 个视频序列，包含 1,244,340 帧和 663 个单词。
PRAI-1581	2020	图像	39,461 张图像，包含 1581 个行人身份。
VOT-ST2020/VOT-RT2020	2020	视频	1,000 个序列，每个长度不一，平均长度约为 100 帧。
VOT-LT2020	2020	视频	50 个序列，每个长度约为 40,000 帧。
VOT-RGBT2020	2020	视频、红外视频	50 个序列，每个长度约为 40,000 帧。
VOT-RGBD2020	2020	视频、深度图像	80 个序列，总计约 101,956 帧。
GOT-10K	2019	图像、视频	420 个视频片段，属于 84 个目标类别和 31 个运动类别。
DTB70	2017	视频	70 个视频序列，每个序列由多个视频帧组成，每帧包含分辨率为 1280x720 像素的 RGB 图像。
Stanford Drone	2016	视频	19,000+ 条目标轨迹，包含 6 种目标类型，约 20,000 次目标交互，40,000 次目标与环境交互，涵盖大学校园内的 100 多个场景。
COWC	2016	图像	标注了 32,716 辆独特的车辆和 58,247 个非车辆目标。涵盖 6 个不同的地理区域。

2.1.4 动作识别

使无人机理解人类动作并通过手势解释指令是一个关键的研究领域。Aeriform In-Action 数据集 [[261]] 针对空中视频中的人类动作识别，包含 32 个高分辨率视频，跨越 13 个动作类别。该数据集专门设计用于应对空中监控中与动作识别相关的独特挑战。MEVA 数据集 [[262]] 提供了一个大规模、多视角、多模态的数据集，包含 9300 小时由无人机和地面摄像机拍摄的连续视频。它涵盖 37 个活动类别并促进先进任务，例如多视角活动检测。此外，UAV-Human 数据集 [[79]] 提供了 67,428 个多模态视频序列，包括 119 个对象用于动作识别。除了动作识别外，它还支持姿态估计和人员重识别等任务。 该数据集涵盖了多样的背景、光照条件和环境，为基于无人机的人类行为分析提供了全面的基准。

名称	年份	类型	数量
Aeriform in-action	2023	视频	32 个视频，13 种动作类型，55,477 帧，40,000 个标注（callouts）。
MEVA	2021	视频、红外视频、GPS、点云	总计 9,300 小时视频，144 小时活动注释，37 种活动类型，超过 270 万个 GPS 轨迹点。
UAV-Human	2021	视频、夜视视频、鱼眼视频、深度视频、红外视频、骨架	67,428 个视频（155 种动作类型，119 个主体），22,476 帧标注关键点（17 个关键点），41,290 帧人员重识别（1,144 个身份），22,263 帧属性识别（如性别、帽子、背包等）。
MOD20	2020	视频	20 种动作类型，2,324 个视频，503,086 帧。
NEC-DRONE	2020	视频	5,250 个视频，包含 256 分钟的动作视频，涉及 19 位演员和 16 个动作类别。
Drone-Action	2019	视频	240 个高清视频，66,919 帧，13 种动作类型。
UAV-GESTURE	2019	视频	119 个视频，37,151 帧，13 种手势类型，10 位演员。

2.1.5 导航与定位

CityNav 数据集 [[267]] 是一个专为语言指导的空中导航任务而设计的数据集，旨在帮助无人机使用自然语言指令在城市规模的三维环境中导航。该数据集包含 32,000 个任务，提供广泛的地理信息和详细的城市环境模型。AerialVLN 数据集 [[268]] 关注通过视觉和语言线索的整合进行无人机导航，使无人机能够基于自然语言命令在复杂环境中执行飞行任务，从而增强其在动态环境中的适应能力。VIGOR 数据集 [[269]] 提供了一个跨视角图像定位数据集，促进了从不同视角对无人机的精确地理定位，提高了在复杂地理环境中图像匹配和位置校准的精度。 University-1652 数据集 [[270]] 作为跨视角地理定位的基准，填补了地面视角和卫星视角之间的视觉差距，采用了无人机视角图像。它包括来自 1,652 所大学的合成无人机、卫星和地面相机的配对图像，支持两个任务：无人机视角目标定位和无人机导航。

名称	年份	类型	数量
CityNav	2024	图像、文本	32,000 条自然语言描述和配套轨迹。
CNER-UAV	2024	文本	12,000 个标注样本，包含 5 种地址标签类型（例如：建筑、单元、楼层、房间等）。
AerialVLN	2023	模拟器路径、文本	25 个城市级场景，8,446 条路径，每条路径 3 条自然语言描述，总计 25,338 条指令。
DenseUAV	2023	图像	训练集：6,768 张无人机图像，13,536 张卫星图像。测试集：2,331 张无人机查询图像，4,662 张卫星图像。
map2seq	2022	图像、文本、地图路径	29,641 张全景图像，7,672 条导航指令文本。
VIGOR	2021	图像	90,618 张航空图像，238,696 张街道全景图像。
University-1652	2020	图像	1,652 栋大学建筑，72 所大学，50,218 张训练图像，37,855 张无人机查询图像，701 张卫星查询图像，以及 21,099 张普通图像和 5,580 张街景图像。

2.2 领域特定数据集

与通用领域数据集相比，领域特定数据集旨在针对特定应用进行定制，并根据其所涉及的具体领域进行分类，包括交通、遥感、农业、工业应用、应急响应、军事行动和野生动物。

2.2.1 交通运输

交通场景是无人机数据集中最常见的场景之一，本部分重点介绍专为交通监测以及车辆和行人检测任务而设计的数据集（见表 [7]），这些任务是无人机技术的关键应用。TrafficNight 数据集 [[274]] 是一种用于夜间车辆监测的航空多模态数据集，旨在解决现有航空数据集在光照条件和车辆类型代表性方面的局限性。该数据集结合了垂直 RGB 和热红外成像技术，覆盖多种场景，包括拥有大量半挂车的场景，并提供专业注释。它还包括相应的 HD-MAP 数据，以便进行多车辆追踪。VisDrone 数据集 [[275]] 是一个大规模基准，支持图像和视频中的目标检测，以及单一和多目标追踪。该数据集收集自中国 14 个城市，具有高度的多样性和挑战性场景，使其非常适合在复杂的城市和郊区环境中评估算法。CADP 数据集 [[276]] 强调交通事故分析，通过使用 CCTV 交通监控视频提高小物体检测准确性（例如行人）。它整合了上下文挖掘技术和基于 LSTM 的架构用于事故预测。CARPK 数据集 [[277]] 引入了一种新颖的停车场车辆计数方法，采用空间正则化区域建议网络（LPN）。该数据集包含超过 90,000 辆车辆的高分辨率无人机影像，增强了物体检测和计数性能。iSAID 数据集 [[278]] 为实例分割任务提供高质量注释，包括 15 个类别下的 655,451 个标注实例，从而支持无人机应用中的准确物体检测和场景分析。这些数据集共同推动了车辆检测、物体跟踪、交通监控和无人机自主导航的研究，为智能交通、无人机巡逻和配送系统的应用提供了强大的数据资源。

名称	年份	类型	数量
TrafficNight	2024	图像、红外图像、视频、红外视频、地图	数据集包含 2,200 对带标注的热红外和 sRGB 图像数据，以及来自 7 个交通场景的视频数据，总时长约 240 分钟。每个场景都包含一张高精度地图，提供详细的布局和拓扑信息。

名称	年份	类型	数量
VisDrone	2022	视频、图像	263 个视频，179,264 帧。10,209 张静止图像。超过 2,500,000 个目标实例标注。数据涵盖 14 个不同的城市，覆盖范围广泛的天气和光照条件。
ITCVD	2020	图像	总共收集了 173 张航空图像，其中训练集 135 张，包含 23,543 辆车辆，测试集 38 张，包含 5,545 辆车辆。图像之间有 60% 的区域重叠，但训练集和测试集之间没有重叠。
UAVid	2020	图像、视频	30 个视频，300 张图像，8 个语义类别标注。
AU-AIR	2020	视频、GPS、高度、IMU、速度	32,823 帧视频，分辨率 1920x1080，30 FPS，分为 30,000 个训练验证样本和 2,823 个测试样本。8 个视频总时长约 2 小时，共有 132,034 个实例，分布在 8 个类别中。
iSAID	2020	图像	总图像数：2,806 张。总实例数：655,451 个。测试集：935 张图像（不公开标注，用于服务器评估）。
CARPK	2018	图像	1448 张图像，约 89,777 辆车辆，提供框标注。
highD	2018	视频、轨迹	16.5 小时，110,000 辆车，5,600 次变道，45,000 公里，总计约 447 小时的车辆行驶数据；4 个预定义驾驶行为标签。
UAVDT	2018	视频、天气、高度、摄像机角度	100 个视频，约 80,000 帧，每秒 30 帧，包含 841,500 个目标框，涵盖 2,700 个目标。
CADP	2016	视频	总时长 5.24 小时，1,416 个交通事故片段，205 个全时空标注视频。
VEDAI	2016	图像	1,210 张图像（1024 × 1024 和 512 × 512 像素），9 种车辆类型，总共包含约 6,650 个目标。

2.2.2 遥感

在遥感领域，多个创新数据集，如表 [8] 所示，为目标检测、分类、定位和图像分析等任务提供了实质性的支持 [[14]]。xView 数据集 [[285]] 是一种大规模卫星图像数据集，包含超过一百万个注释，涵盖多个物体类别，使其特别适合于目标检测和图像分割任务，尤其是在复杂背景和具有挑战性的环境中。DOTA 数据集 [[286]] 专注于高分辨率航空图像中的目标检测，涵盖多个物体类别，如飞机、车辆和建筑物，适合于复杂场景中的多目标检测和分类任务。RSICD 数据集 [[287]] 主要用于遥感图像中的场景分类任务，并支持语言描述生成，提供了一个标准化的基准，促进了图像理解和自动注释技术的研究。RemoteCLIP [[288]] 引入了一种遥感视觉语言模型，通过自监督学习和掩蔽图像建模增强遥感图像的语义分析和图像检索，推动了无人机在遥感数据分析中的应用。

名称	年份	类型	数量
RET-3	2024	图像、文本	约 13,000 个样本。包括 RSICD、RSITMD 和 UCM。
DET-10	2024	图像	在目标检测数据集中，每张图像的目标数量范围为 1 到 70，总计约 80,000 个样本。
SEG-4	2024	图像	分割数据集涵盖不同区域和分辨率，总计约 72,000 个样本。
DIOR	2020	图像	23,463 张图像，包含 192,472 个目标实例，涵盖 20 个类别，包括飞机、车辆、船只、桥梁等，每个类别包含约 1,200 个实例。
TGRS-HRRSD	2019	图像	总图像数：21,761 张。13 个类别，包括飞机、车辆、桥梁等。目标总数约为 53,000 个。
xView	2018	图像	有超过 100 万个目标和 60 个类别，包括车辆、建筑物、设施、船只等，分为七个父类别和几个子类别。
DOTA	2018	图像	2806 张图像，188,282 个目标，15 个类别。
RSICD	2018	图像、文本	10,921 张图像，54,605 条描述性句子。

名称	年份	类型	数量
HRSC2016	2017	图像	3,433 个实例，总计 1,061 张图像，包括 70 张纯海洋图像和 991 张包含混合陆海区域的图像。2,876 个标记的船只目标。610 张未标记图像。
RSOD	2017	图像	包含 4 种目标类型（坦克、飞机、立交桥、操场），有 12,000 个正样本和 48,000 个负样本。
NWPU-RESISC45	2017	图像	总共 31,500 张图像，涵盖 45 个场景类别，每个类别 700 张图像，分辨率 256 × 256 像素，空间分辨率从 0.2m 到 30m。
NWPU VHR-10	2014	图像	800 张高分辨率图像，其中 650 张包含目标，150 张是背景图像，涵盖 10 个类别（如飞机、船只、桥梁等），总计超过 3,000 个目标。

2.2.3 农业

农业部分总结了过去两年中仅公开可用的数据集，如表 [9] 所示，因为几篇综述文章已覆盖 2023 年前的数据集。农业数据集通常用于目标检测，以识别杂草、入侵植物或植物疾病和害虫，而语义分割通常用于田地划分。Avo-AirDB 数据集 [[295]] 专门用于农业图像分割和分类，提供高分辨率的鳄梨作物图像，以支持在精准农业中的植物识别和健康监测。CoFly-WeedDB 数据集 [[296]] 由 201 幅航空图像组成，捕捉到三种干扰棉花作物的杂草，以及相应的注释图像。WEED-2C 数据集 [[297]] 专注于训练无人机图像，以在大豆田中识别杂草物种，自动识别两种杂草物种。

名称	年份	类型	数量
WEED-2C	2024	图像	包含 4,129 个标注样本，涵盖 2 种杂草。
CoFly-WeedDB	2023	图像、健康数据	包含 201 张航空图像、3 种受干扰行作物（棉花）的不同杂草类型及其对应的标注图像。
Avo-AirDB	2022	图像	984 张高分辨率 RGB 图像（5472 × 3648 像素），其中 93 张具有详细的多边形标注，分为 3 到 4 个类别（小、中、大和背景）。

2.2.4 行业

使用无人机影像进行工业检查，特别是在基础设施维护方面，变得越来越重要。表 [9] 列出了几个典型数据集。UAPD 数据集 [[298]] 专注于通过无人机影像和 YOLO 架构检测沥青路面裂缝，旨在通过自动化裂缝检测提高公路和高速公路的维护效率。InsPLAD 数据集 [[299]] 专门为电力线路资产检测而设计，包含集中于基础设施（如电力线路、塔和绝缘子的无人机影像）。通过提供在多种环境条件下的影像，该数据集支持开发自动检查系统，以识别电力设备中的损坏或老化，从而提高电力线路检查的效率和准确性。

名称	年份	类型	数量
UAPD	2021	图像	原始数据中有 2,401 张裂缝图像，数据增强后有 4,479 张裂缝图像。
InsPLAD	2023	图像	10,607 张无人机图像，包含 17 类电力资产，总共有 28,933 个标注实例，以及 5 种资产的缺陷标签，总共 402 个缺陷样本，分为 6 种缺陷类型。

2.2.5 紧急响应

这些数据集通常用于增强无人机在灾害救援场景中的视觉理解能力，特别是在灾后场景分析、灾区监测、环境评估和救援操作中。它们促进了快速图像识别、目标检测和场景理解任务。航空合成孔径雷达数据集 [[300]] 探讨了无人机在自然灾害监测和搜救操作中的应用，强调了在灾区快速部署和自主管理无人机的潜力。AFID 数据集 [[301]] 提供了用于水道监测和灾害预警的航空图像，支持深度语义分割模型的训练。FloodNet 数据集 [[302]] 提供了高分辨率航空图像，用于灾后场景理解，主要旨在协助灾后评估和紧急救援操作。 通过利用这些数据集，研究人员可以显著提高在灾难响应中的图像分析能力，并推动无人机技术在灾难救援工作中的实际应用。

名称	年份	类型	数量
AFID	2023	图像	总共 816 张图像，分辨率为 2720 × 1536 和 2560 × 1440。包含 8 个语义分割类别。
FloodNet	2021	图像、文本	整个数据集有 2,343 张图像，分为训练集（约 60%）、验证集（约 20%）和测试集（约 20\%）。语义分割标签包

名称	年份	类型	数量
			括：背景、淹没的建筑物、未淹没的建筑物、淹没的道路、未淹没的道路、水、树木、车辆、水池、草地。
Aerial SAR	2020	图像	2,000 张图像，包含 30,000 个动作实例，涵盖多种人类行为。

2.2.6 军事

MOCO 数据集 [[303]] 是为军事图像字幕生成（MilitIC）任务而设计的，该任务专注于从低空无人机（UAV）和无人地面车辆（UGV）在军事背景下捕获的图像中生成文本信息。该数据集包含一个训练集，包括 7192 张图像和 35960 个字幕，以及一个测试集，包含 257 张图像和 1285 个字幕。军事图像字幕生成作为一种视觉-语言学习任务，旨在自动生成军事图像的描述性字幕，从而增强态势感知并支持决策制定。通过将图像数据与文本描述结合，这种方法提高了军事领域的情报能力和作战效率。

名称	年份	类型	数量
MOCO	2024	图像、文本	7,449 张图像，37,245 条文字描述（captions）。

2.2.7 野生动物

WAID [[304]] 是一个大规模、多类别、高质量的数据集，专门设计用于支持无人机在野生动物监测中的应用。该数据集包括 14,375 幅在各种环境条件下拍摄的无人机图像，涵盖六种野生动物和多种栖息地。

名称	年份	类型	数量
WAID	2023	图像	14,375 张无人机图像，涵盖 6 种野生动物和多种环境类型。

2.3 3D 模拟平台

三维仿真平台在无人机的发展和应用中发挥着至关重要的作用，为智能无人机训练提供安全、可控和多样化的测试场景，这些场景位于高度仿真的虚拟环境中。这些环境涵盖复杂的条件，如变化的天气、光照、风速、地形和障碍物。这类平台能够生成大规模、准确标注的多模态数据集，用于训练和验证。此外，仿真平台支持多机协作任务的建模，评估无人机在共享空间内的协作能

力、通信和避障策略。这有效降低了与实际测试相关的风险和成本。硬件在环（HIL）仿真进一步将虚拟测试与真实硬件相结合，帮助识别潜在问题并验证系统可靠性。总之，三维仿真平台在智能训练、数据集生成、协作任务执行和硬件验证方面起到了关键作用，显著加速了无人机技术的发展和部署。

名称	出版物
AirSim	Airsim: 高保真视觉和物理模拟，用于自动驾驶车辆
Carla	CARLA: 一个开放的城市驾驶模拟器
NVIDIA Isaac Sim	基于物理的机器人模拟平台，建立在 NVIDIA Omniverse 平台之上，为机器人和自主系统的开发、测试和验证提供高精度的虚拟环境。
AerialVLN Simulator	Aerialvln: 无人机视觉与语言导航
Embodied City	EmbodiedCity: 用于现实世界城市环境中具身智能体的基准平台

2.3.1 AirSim

AirSim [[305]] 是由微软开发的开源跨平台模拟器，专为无人机、自动驾驶车辆及其他自主系统的研究与开发而设计。该平台基于虚幻引擎构建，提供高度逼真的物理模拟环境和视觉效果，使用户能够在虚拟场景中测试和验证算法的性能。AirSim 支持各种设备和传感器的模拟，包括摄像头、激光雷达、惯性测量单元（IMU）、全球定位系统（GPS）等，同时通过强大的 API 提供对环境和车辆的全面控制。开发者可以使用 Python 和 C++扩展该平台，实现机器学习、计算机视觉和机器人等领域前沿技术的集成。除了模拟无人机和地面车辆外，该平台还可以建模复杂的动态场景，包括天气变化、碰撞检测和物理交互，帮助用户在安全可控的虚拟环境中加速原型验证和算法优化。

2.3.2 Carla

CARLA [[306]] 是一个基于虚幻引擎的开源自动驾驶模拟平台，广泛用于智能系统算法的开发、训练和验证。其高度真实的模拟环境支持复杂的城市场景，包括道路网络、动态交通、行人行为以及多样的天气和光照条件，为感知、定位、规划和控制算法提供了一个虚拟测试场。CARLA 支持多种传感器的模拟，如摄像头、激光雷达、雷达、惯性测量单元和全球定位系统，并允许用户访问其 Python 或 C++ API，以及支持 ROS 的接口，使研究人员能够快速开发和测试

导航、避障、路径规划和环境感知的算法。此外，CARLA 提供数据录制和回放功能，支持多智能体任务，并集成强化学习应用，为低空物流、监控和无人机巡逻等场景中的算法开发提供安全、高效和可重复的测试平台。

2.3.3 NVIDIA Isaac Sim

NVIDIA Isaac Sim [[307]] 是一个基于物理的机器人模拟平台，建立在 NVIDIA Omniverse 平台之上，为机器人和自主系统的开发、测试和验证提供高精度的虚拟环境。该平台利用 NVIDIA 强大的 GPU 加速和物理引擎技术，包括 PhysX 和 RTX 实时渲染，呈现高度逼真的模拟场景，具有准确的物理交互、光照效果和多传感器数据生成。Isaac Sim 提供广泛的工具和插件，允许与各种机器人框架集成，并支持从感知、运动规划到控制算法的完整开发过程。除了在传统机器人领域的应用，Isaac Sim 还可以扩展到无人机领域，通过灵活的环境配置、传感器模拟（包括摄像头、激光雷达、惯性测量单元和 GPS）以及复杂的动力学建模，支持无人机导航、障碍避让、目标跟踪和多智能体协作任务。该平台结合了模拟强化学习能力、数据收集功能和数字双胞胎支持，以应对现实世界场景，从而加速无人机在物流、环境监测和灾难响应等领域的算法开发，同时为研究人员和开发者提供高效、安全和可扩展的测试环境。

2.3.4 空中虚拟导航模拟器

AerialVLN 仿真器 [[268]] 是一个高保真虚拟仿真平台，专门用于无人机代理的研究。它结合了虚幻引擎 4 和微软 AirSim 技术，真实地模拟典型的 3D 城市环境，包括上海、深圳等城市、校园和居民区，覆盖范围从 30 公顷到 3,700 公顷。该平台支持多样的环境设置，包括昼夜不同的光照条件、天气模式如晴天、阴天和小雪，以及季节变化，使无人机代理能够在与现实条件密切相关的环境中进行训练。平台配备前、后、左、右和顶部的多视角摄像头，能够生成 RGB 图像、深度图像和目标分割图等高分辨率数据，为场景理解和空间建模提供丰富的视觉输入。 此外，AerialVLN 模拟器支持无人机的动态飞行操作，提供对其三维位置、方向和速度的精确控制，同时允许执行复杂的机动，如转弯、爬升和避障，确保飞行动作的平滑和灵活。基于“真实到模拟再到真实”的设计理念，该平台显著缩小了虚拟环境与现实世界应用之间的差距，使其特别适

合于核心无人机任务的研究和优化，如场景感知、空间推理、路径规划和运动决策。

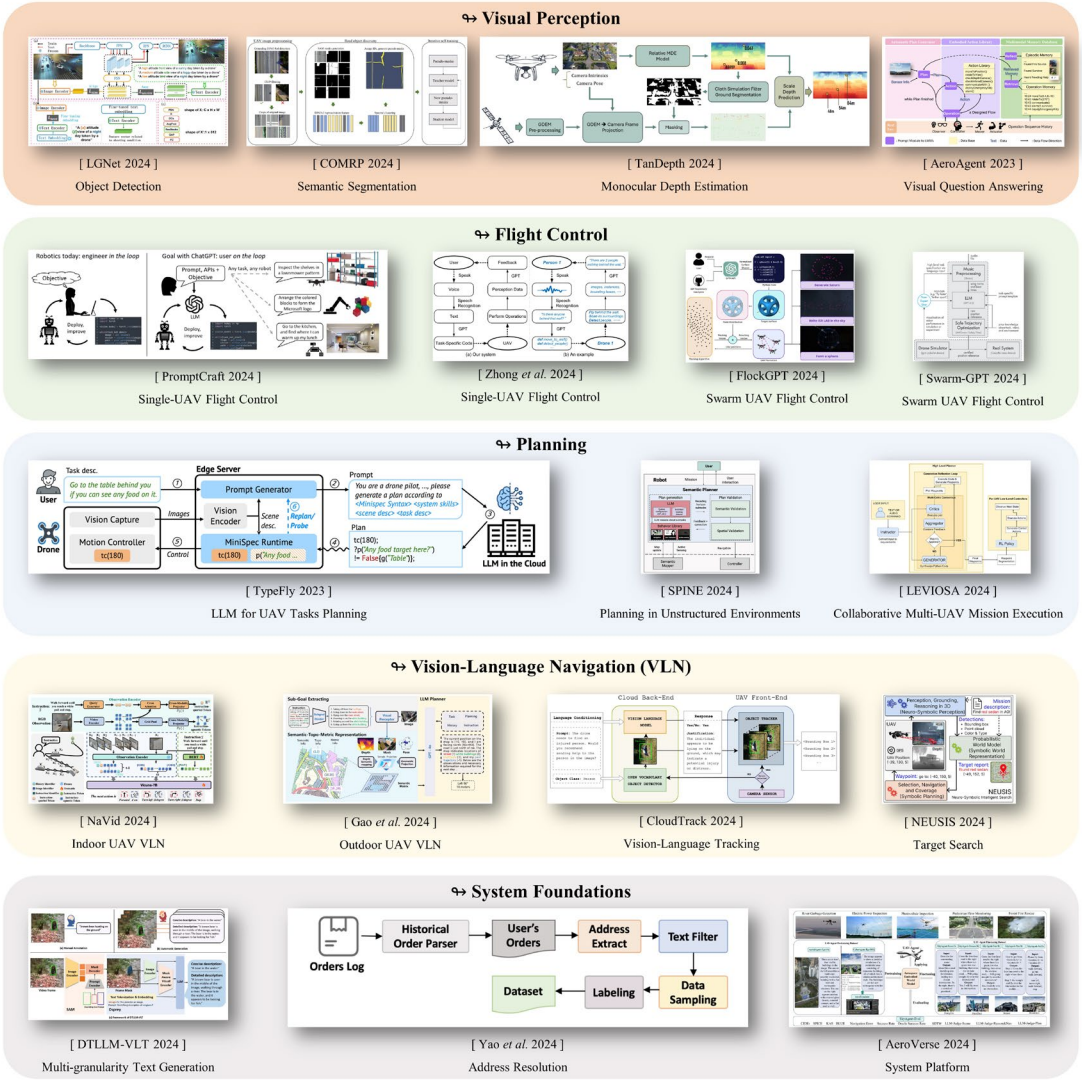
2.3.5 具身城市

具体现实城市 [[308]] 是一个先进的高保真 3D 城市仿真平台，专为具体现智能的评估和开发而设计。其核心特征是基于现实世界城市区域（如北京的商业区）构建的逼真虚拟环境，包括高度详尽的建筑模型、街道网络以及行人和车辆交通的动态仿真。该平台以虚幻引擎为技术基础，将历史数据与仿真算法相结合，为各种具体现代理（如无人机和地面车辆）提供持续的感知和交互能力。通过集成 AirSim 接口，它支持多模态输入和输出，包括 RGB 图像、深度图像、激光雷达、GPS 和 IMU 数据，便于在仿真中进行运动控制和环境探索。设计涵盖五个任务领域：场景理解、问答、对话、视觉语言导航和任务规划。通过一个易于使用的 Python SDK 和一个在线平台，用户可以方便地远程访问和测试多种代理行为，同时支持多达八个代理的实时操作。

3 基于基础模型的无人机系统的进展

将人工智能算法，如机器学习和深度学习，整合到无人机系统中已成为一种主流趋势。然而，在无人机任务中应用传统人工智能模型仍面临许多挑战。首先，这些模型通常依赖于特定任务的数据集进行训练，因此在实际场景与训练分布之间存在显著差异时，泛化能力不足和鲁棒性差。此外，传统人工智能模型通常是针对单一任务进行优化的，使它们在应对多任务协作的复杂需求时效果不佳。此外，这些模型在人机交互和任务协作方面存在明显的局限性 [[309], [310], [311]]。

LLMs、VFMs 和 VLMs 的引入通过自然语言理解、零-shot 自适应、多模态协作和直观的人机交互为无人机系统注入了新颖的智能能力。



本节探讨了将 LLMs、VFMs 和 VLMs 整合到无人机系统中的现有研究，并分析了这些技术为不同任务带来的优势。图 [4] 中展示了几个典型的工作。根据技术类型和任务特征，无人机相关任务被分类为以下几种类型：

1. **视觉感知：** 这些包括对象检测、语义分割、深度估计、视觉描述和视觉问答（VQA）。此类任务侧重于环境感知和语义信息提取，为无人机系统中的高层决策提供基础。
2. **视觉-语言导航（VLN）：** VLN 代表了计算机视觉和自然语言处理深度集成的典型应用。在 VLN 任务的基础上，开发了更复杂的多模态任务，例如视觉-语言跟踪（VLT）和目标搜索。这些任务整合了多个组件，包括感知、规划、决策、控制和人机交互，构成了无人机智能任务执行的核心框架。

3. **规划：** 这包括路径优化、任务分配和动态环境中的自适应任务优化。
4. **飞行控制：** 这些涉及低层控制任务，如姿态稳定、路径跟踪和避障。
5. **基础设施：** 这主要集中在为无人机系统提供全面的技术和数据支持，包括集成框架和平台的开发，以及高质量数据集的创建和处理。这些努力不仅提高了无人机在多模态任务中的应用效率，还为无人机领域的基础研究和技术创新提供了重要支持。

3.1 视觉感知

标题	类型	出版物	代码
Li et al. (A Benchmark for UAV-View Natural Language-Guided Tracking)（无人机视角自然语言引导跟踪基准）	VFM	*MDPI*	*GitHub*
Ma et al. (Applying Unsupervised Semantic Segmentation to High-Resolution UAV Imagery for Enhanced Road Scene Parsing)（将无监督语义分割应用于高分辨率无人机图像以增强道路场景解析）	VFM	*Arxiv*	-
Limberg et al. (Leveraging YOLO-World and GPT-4V LMMs for Zero-Shot Person Detection and Action Recognition in Drone Imagery)（利用 YOLO-World 和 GPT-4V LMM 进行无人机图像中的零样本人物检测和动作识别）	VFM+VLM	*Arxiv*	-
Kim et al. (Weather-Aware Drone-View Object Detection Via Environmental Context Understanding)（通过环境上下文理解实现天气感知无人机视角目标检测）	VLM+VFM	*ICIP 2024*	-
LGNet (Shooting condition insensitive unmanned aerial vehicle object detection)（对拍摄条件不敏感的无人机目标检测）	VFM	*Expert Systems with Applications*	-
Sakaino et al. (Dynamic Texts From UAV Perspective Natural Images)（来自无人机视角的自然图像中的动态文本）	VLM+VFM	*ICCV 2023*	-

标题	类型	出版物	代码
COMRP (Unsupervised semantic segmentation of high-resolution UAV imagery for road scene parsing)（高分辨率无人机图像的无监督语义分割以进行道路场景解析）	VFM	*Arxiv*	*GitHub*
CrossEarth (CrossEarth: Geospatial Vision Foundation Model for Domain Generalizable Remote Sensing Semantic Segmentation)（CrossEarth：用于域泛化遥感语义分割的地理空间视觉基础模型）	VFM	*Arxiv*	*GitHub*
TanDepth (TanDepth: Leveraging Global DEMs for Metric Monocular Depth Estimation in UAVs)（TanDepth：利用全球数字高程模型（DEM）进行无人机中的度量单目深度估计）	VFM	*Arxiv*	*GitHub*
DroneGPT (DroneGPT: Zero-shot Video Question Answering For Drones)（DroneGPT：无人机零样本视频问答）	VLM+LLM+VFM	*CVDL 2024*	-
de Zarzà et al. (Socratic video understanding on unmanned aerial vehicles)（无人机上的苏格拉底式视频理解）	LLM	*Procedia Computer Science*	-
AeroAgent (Agent as Cerebrum, Controller as Cerebellum: Implementing an Embodied LMM-based Agent on Drones)（Agent 作大脑，控制器作小脑：在无人机上实现基于具身 LMM 的智能体）	VLM	*Arxiv*	-
RS-LLaVA (Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery)（RS-LLaVA：用于遥感图像联合字幕生成和问答的大型视觉语言模型）	VLM	*MDPI*	-
GeoRSCLIP (RS5M and GeoRSCLIP: A large scale vision-language dataset and a large vision-language model for remote sensing)（RS5M 和 GeoRSCLIP：用于遥感的超大	VFM	*IEEE Transactions on*	*GitHub*

标题	类型	出版物	代码
规模视觉语言数据集和大型视觉语言模型）		*Geoscience and Remote Sensing*	
SkyEyeGPT (Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model)（SkyEyeGPT：通过大型语言模型指令微调统一遥感视觉语言任务）	VFM+LLM	*Arxiv*	*GitHub*
AirVista (AirVista: Empowering UAVs with 3D Spatial Reasoning Abilities Through a Multimodal Large Language Model Agent)（AirVista：通过多模态大语言模型智能体赋予无人机 3D 空间推理能力）	VFM+VLM	*ITSC2024*	-

3.1.1 目标检测

传统的目标检测算法在无人机应用中面临重大挑战。无人机的飞行高度和视角的变化导致视野的变化，使得多尺度目标检测成为关键研究焦点 [[357], [358], [359]]。然而，动态环境条件、多样化拍摄场景和不可预测性进一步增加了检测任务的复杂性 [[360], [361]]。此外，不同场景的特定领域特征使模型在多样化环境中实现稳健泛化变得困难。为了解决这些挑战，一些研究尝试通过改进训练策略来增强模型的鲁棒性，例如为特定无人机场景训练专用模型或引入多任务学习框架 [[362], [363], [364]]。然而，这些方法通常会产生高昂的训练成本，并且在泛化能力上仍然存在局限性。 此外，传统的监督学习方法严重依赖大量手动标记的数据，进一步增加了与数据集构建相关的时间和资源成本。

从更高的角度来看，传统的基于边界框的目标检测方法主要关注对象的几何特征，缺乏建模上下文信息的能力。这种离散的低级特征表示难以捕捉复杂的语义信息，限制了它们在高层次任务中的潜力。自然语言的引入为无人机目标检测任务开辟了新的途径。自然语言与视觉的结合利用了它们的互补优势。凭借视觉语言模型（VLMs）和视觉特征模型（VFMs）的灵活性、零样本学习能力、上下文理解以及强泛化性能，这些模型能够有效应对复杂任务，并通过多模态整合显著提高无人机目标检测的准确性和适应性。

在特定应用中，Li 等[[312]] 将 CLIP [[186]] 与传统目标跟踪模块结合，以实现无人机的自然语言跟踪（TNL）任务。Ma 等[[313]] 通过整合 Grounding DINO [[192]] 和 CLIP，增强了无人机图像中道路场景检测的准确性。Limberg 等[[314]] 利用 YOLO-World [[196]] 和 GPT-4V [[167]] 的组合，实现了无人机图像中的零样本人类检测和动作识别。Kim 等[[315]] 使用 LLaVA-1.5 [[170]] 通过将视觉特征与天气和光照等语言提示相结合，为无人机图像生成天气描述。通过使用 CLIP 编码器，他们将图像特征与天气相关信息融合。 基于这一框架，实施了天气感知物体查询，有效利用天气信息进行物体检测任务，从而显著提高了检测精度和鲁棒性。

值得注意的是，CLIP 的多模态表示能力能生成高质量的领域不变特征，为传统物体检测模型的训练提供强有力的支持。例如，LGNet [[3]] 引入了 CLIP 的多模态特征，显著增强了在多样拍摄条件下无人机物体检测的鲁棒性和性能。此外，LLMs、VLMs 和 VFMs 在一般物体检测任务中积累了丰富的研究经验，为无人机物体检测任务提供了重要的见解。相关例子包括 LLM-AR [[365]]、Han 等[[366]]、Lin 等[[367]]、ContextDET [[368]] 和 LLMI3D [[369]]。

然而，仅依靠 VFMs 或 VLMs 进行物体检测可能会在某些场景中由于模型幻觉或任务特定适应性不足而导致性能限制 [[370], [371], [372]]。虽然传统深度学习模型在特定任务中表现可靠，但它们缺乏跨任务泛化能力。更好的解决方案是采用“大型模型 + 小型模型”的协同架构，利用大型模型的强泛化能力和小型模型的领域专业化。例如，Hidetomo Sakaino 视觉识别组 [[316]] 提出了一种将深度学习模型与 VLMs 相结合的方法，用于可见度和天气条件估计。该方法有效地解决了图像处理中的诸如尺度变化、视角变化和环境干扰（包括天空背景干扰和远距离小物体检测）等挑战。它在各种环境和天气条件下展现了卓越的鲁棒性和稳定性。

3.1.2 语义分割

作为一种计算机视觉任务，UAV 语义分割面临许多与目标检测相似的挑战，例如，由对抗性视觉条件引起的不足的泛化能力以及对手动标注数据的高度依赖。此外，UAV 遥感图像在语义分割任务中遇到复杂的前景-背景交互，涉及多尺度场景，这在该领域提出了独特的挑战 [[373], [374]]。尽管领域适应（DA）

技术已广泛应用于跨域语义分割，但这些方法主要迫使模型从源域调整到预定义的目标域。然而，这些方法对未见域的泛化能力有限，突显了在多样化场景中提升模型性能的迫切需要更灵活和鲁棒的策略 [[375], [376]]。

VLMs 和 VFMs 的引入为无人机语义分割任务注入了新的技术动力。这些模型能够高效地执行零-shot 语义分割，同时通过自然语言交互灵活地定义和指导分割任务，展现出满足多样化场景需求的卓越潜力。例如，COMRP [[313]] 专注于解析高分辨率无人机图像中的道路场景。其方法首先利用 Grounding DINO [[192]] 和 CLIP [[186]] 提取与道路相关的区域，并使用 SAM [[198]] 自动生成分割掩膜。然后，使用 ResNet [[377]] 和 DINOv2 [[193]] 提取特征，并通过谱聚类方法对掩膜特征向量进行聚类以生成伪标签。这些标签用于训练教师模型，迭代优化学生模型的性能。COMRP 消除了对人工标注的依赖，为无人机道路场景解析提供了一种高效和自动化的解决方案。

此外，CrossEarth [[319]] 是一种针对遥感领域设计的跨领域泛化语义分割 VFM。它结合了两种互补策略：地球风格注入和多任务训练，显著增强了跨领域泛化能力。地球风格注入将来自地球领域的多种风格融入源领域数据，扩展了训练数据的分布范围。多任务训练利用共享的 DINOv2 主干网络，同时优化语义分割和掩膜图像建模任务，从而能够学习到稳健的语义特征。

3.1.3 深度估计

无人机感知系统的核心功能之一是对地形和自然环境进行三维建模，从而生成飞行区域的一致且准确的三维几何表示。近年来，基于神经辐射场（NeRF）和三维高斯喷溅（3DGS）的方法在这个任务上取得了显著进展，例子包括 UAV-NeRF [[378]] 和 AGS [[379]]。然而，这些方法在大规模场景中仍面临诸多挑战。在此背景下，单目深度估计（MDE）逐渐成为一种更有优势的解决方案 [[380], [381], [382]]。

Florea 等[[320]] 提出了 TanDepth 框架，该框架结合了 Depth Anything[[215]] 模型的相对深度估计与全球数字高程模型（GDEM）数据，采用尺度恢复方法生成具有真实世界尺寸的高精度深度图像。对多个无人机数据集的实验结果表明，TanDepth 在复杂地形和动态飞行环境中展现出卓越的准确性和鲁棒性。这

一方法为无人机深度估计任务开辟了新的技术方向，特别展示了其在缺乏高精度深度传感器的场景中的效率和适应性。

3.1.4 视觉标题和视觉问答

视觉标题和视觉问答属于计算机视觉与自然语言处理的跨模态融合领域，侧重于图像和视频内容的语义理解及自然语言表示。传统方法通常基于深度学习框架，其中视觉特征提取和语言生成被设计为独立模块。然而，这种分离设计在复杂场景、开放领域问题和细粒度描述生成中存在显著限制，主要受到模型表达能力和多模态特征不对齐的约束 [[383], [384], [385], [386]]。随着视觉语言模型（VLMs）和视觉特征模型（VFMs）的快速发展，这些模型利用视觉和语言的联合表示学习来显著增强对复杂跨模态信息的理解。 预训练于大规模多模态数据集的 VLM 和 VFM 展示了卓越的任务泛化能力，并能够在复杂场景中生成细致的语义描述，展现出对开放领域任务的高度适应性 [[169], [174], [175], [176], [180], [185]]。

在无人机视觉标注和视觉问答任务中，研究主要集中在两个方向：第一是根据特定的无人机需求，选择或结合现有的视觉语言模型（VLMs）和视觉功能模型（VFMs），使其适应无人机任务场景；第二是对 VLMs 或 VFMs 进行领域特定数据的训练或微调，以构建适用于无人机垂直应用的专业模型，解决独特的挑战。这些研究方向旨在进一步增强无人机在复杂环境中的视觉感知、语义推理和任务执行能力，为智能且用户友好的人机交互提供强有力的支持。

对于第一个研究方向，几项研究探讨了将现有的视觉语言模型（VLMs）和视觉功能模型（VFMs）结合，以适应无人机（UAV）场景。例如，邱等人 *et al.* [[4]] 提出了基于视觉推理模型 VISPROG [[185]] 的 DroneGPT 框架，其中 GPT-3.5 [[141]] 将用户的自然语言查询转换为任务逻辑代码。这些代码调用 Grounding DINO [[192]] 解析视觉信息并执行语义推理，最终输出清晰准确的视觉问答结果。De Zarzà *et al.* 设计了一个将 BLIP-2 [[176]] 与 GPT-3.5 结合的框架，以实现高效的无人机视频场景理解和语义推理，其中 BLIP-2 从每个视频帧中提取初步的语义信息，而 GPT-3.5 生成高级场景描述。AeroAgent 架构 [[322]] 从代理的角度优化无人机视觉问答模块。 基于 GPT-4V [[167]]，它构建了一个

可检索的多模态记忆数据库（类似于 RAG 框架），显著提高了复杂场景下的理解和回答准确性，同时减轻了生成模型中的幻觉问题。

对于第二个研究方向，现有工作主要集中在无人机遥感领域，旨在通过开发针对该领域的视觉语言模型（VLM）和视觉功能模型（VFM），增强对遥感图像的语义理解。传统的遥感视觉分析方法严重依赖领域专业知识，产生高昂的标注成本，并且在处理复杂场景和语义交互方面表现有限。Bazi 等人 提出了 RS-LLaVA [[323]]，该模型对 LLaVA-1.5 [[170]] 进行预训练和微调，以适应特定领域的任务，使得能够为遥感图像生成字幕和进行视觉问答（VQA）。Zhang 等人 [[324]] 通过构建大规模遥感图像-文本配对数据集 RS5M，并对 CLIP [[186]] 模型进行全微调或参数高效微调（PEFT），开发了 GeoRSCLIP 模型。该模型在零-shot 分类（ZSC）、跨模态检索（RSCTIR）和语义定位（SeLo）方面表现出色，展示了强大的领域适应性和任务泛化能力。SkyEyeGPT [[325]] 是遥感领域视觉-语言任务的统一框架，采用 EVA-CLIP [[189]] 作为视觉编码器提取图像特征，并使用 LLaMA2 [[154]] 作为语言解码器生成任务输出。经过指令调优的数据集优化了模型，以支持图像描述、视觉问答（VQA）和视觉定位等任务。

3.2 视觉-语言导航

标题	类型	出版物	代码
Neuro-LIFT (Neuro-LIFT: A Neuromorphic, LLM-based Interactive Framework for Autonomous Drone FlighT at the Edge)（Neuro-LIFT：一个基于神经形态和大模型的交互式框架，用于边缘自主无人机飞行）	LLM	*Arxiv*	-
NaVid (Navid: Video-based vlm plans the next step for vision-and-language navigation)（NaVid：基于视频的视觉语言模型为视觉与语言导航规划下一步）	VFM+LLM	*Arxiv*	-
VLN-MP (Why Only Text: Empowering Vision-and-Language Navigation with Multi-modal Prompts)（为何只有文本：使用多模态提示增强视觉与语言导航）	VFM	*Arxiv*	*GitHub*

标题	类型	出版物	代码
Gao et al. (Aerial Vision-and-Language Navigation via Semantic-Topo-Metric Representation Guided LLM Reasoning)（通过语义-拓扑-度量表示引导的大模型推理实现空中视觉与语言导航）	VFM+LLM	*Arxiv*	-
MGP (CityNav: Language-Goal Aerial Navigation Dataset with Geographic Information)（CityNav：具有地理信息的语言目标空中导航数据集）	LLM+VFM	*Arxiv*	*GitHub*
UAV Navigation LLM (Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology)（迈向真实的无人机视觉语言导航：平台、基准和方法）	LLM+VFM	*Arxiv*	*GitHub*
GOMAA-Geo (GOMAA-Geo: GOal Modality Agnostic Active Geo-localization)（GOMAA-Geo：目标模态无关主动地理定位）	LLM+VFM	*Arxiv*	*GitHub*
NavAgent (NavAgent: Multi-scale Urban Street View Fusion For UAV Embodied Vision-and-Language Navigation)（NavAgent：用于无人机具身视觉与语言导航的多尺度城市街景融合）	LLM+VFM+VLM	*Arxiv*	-
ASMA (ASMA: An Adaptive Safety Margin Algorithm for Vision-Language Drone Navigation via Scene-Aware Control Barrier Functions)（ASMA：通过场景感知控制障碍函数实现视觉语言无人机导航的自适应安全裕度算法）	LLM+VFM	*Arxiv*	-
Zhang et al. (Demo Abstract: Embodied Aerial Agent for City-level Visual Language Navigation Using Large Language Model)（演示摘要：使用大型语言模型进行城市级视觉语言导航的具身空中智能体）	VFM+LLM	*IPSN 2024*	-
Chen et al. (Vision-Language Navigation for Quadcopters with Conditional Transformer and Prompt-based Text Rephraser)（带条件 Transformer 和基于提示文本复述器的四旋翼无人机视觉语言导航）	LLM	*MMAsia 2023*	-

标题	类型	出版物	代码
CloudTrack (CloudTrack: Scalable UAV Tracking with Cloud Semantics)（CloudTrack：基于云语义的可扩展无人机跟踪）	VFM+VLM	*Arxiv*	-
NEUSIS (NEUSIS: A Compositional Neuro-Symbolic Framework for Autonomous Perception, Reasoning, and Planning in Complex UAV Search Missions)（NEUSIS：用于复杂无人机搜索任务中自主感知、推理和规划的组合神经-符号框架）	VFM+VLM	*Arxiv*	-
Say-REAPEx (Say-REAPEx: An LLM-Modulo UAV Online Planning Framework for Search and Rescue)（Say-REAPEx：用于搜救的 LLM-Modulo 无人机在线规划框架）	LLM	*Openreview*	-
OpenFLY (OpenFly: A Versatile Toolchain and Large-scale Benchmark for Aerial Vision-Language Navigation)（OpenFLY：用于空中视觉语言导航的多功能工具链和大规模基准）	LLM+VLM	*Arxiv*	*GitHub*

近年来，基于深度学习的方法在视觉语言导航（VLN）方面取得了显著进展。例如，基于 Transformer 架构的视觉与语言融合技术已广泛应用于 VLN 及其衍生任务 [[268], [387], [388]]。然而，VLN 任务仍面临许多挑战。一方面，多模态特征的对齐和融合仍然是一个核心难题，特别是在动态和复杂场景中，不一致的特征可能导致任务决策的不稳定。另一方面，过度依赖现有注释数据的方法限制了模型在未注释环境中的可迁移性。此外，在开放领域任务中，模型的泛化能力和鲁棒性仍需进一步提高。

随着 VLMs 和 VFMs 的引入，VLN 及其衍生任务迎来了新的发展轨迹。通过大规模的预训练，这些模型能够有效学习跨模态特征的对齐表示，显著增强任务理解和执行能力。在复杂和动态的场景中，它们表现出卓越的泛化性能，为智能无人机导航、目标跟踪和目标搜索提供了更强的技术支持。

无人机视觉-语言导航（UAV VLN）涉及通过将视觉输入与自然语言指令相结合在三维空间中进行路径规划。与传统的地面导航相比， aerial 导航需要考虑飞行高度以及三维空间感知和推理的复杂性。此外，无人机视觉-语言导航任

务在不同场景中差异显著：室内环境具有更明确的几何约束，从而简化了任务规划，而室外环境由于开放空间的尺度和动态环境变化则引入了更大的复杂性。

3.2.1 室内

在室内无人机视觉语言导航（UAV VLN）中，NaVid [[326]] 利用 EVA-CLIP [[189]] 提取视觉特征，并结合 Q-Former [[175], [176]] 生成视觉标记和几何标记。跨模态投影对齐视觉和语言特征，而 Vicuna-7B [[243]] 解析自然语言指令并生成特定的导航动作。该系统仅依赖单目视频流，而不需要地图、里程计或深度信息。通过将历史观测编码为时空上下文，它实现了低级导航动作的实时推理，展现了在室内环境中卓越的路径规划和动态调整能力。此外，多模态提示在无人机视觉语言导航任务中显示出显著潜力。Hong 等 [[327]] 提出了 VLN-MP 框架，通过多模态提示增强任务理解，减少自然语言指令中的模糊性，并支持多样化和高质量的提示设置。 该系统利用数据生成管道生成与地标相关的图像提示，结合了 Grounding DINO [[192]] 或 GLIP [[190]]，同时 ControlNet [[389]] 增强了数据多样性。最后，该系统通过视觉编码器和多层 Transformer 模块融合图像和文本特征，以生成精确的导航动作。

3.2.2 户外

针对户外无人机视觉语言导航（UAV VLN），Liu *et al.* [[268]] 提出了 AerialVLN，解决了空中导航研究中的空白。这项任务要求无人机根据自然语言指令和第一人称视觉感知导航到目标位置，将所有未被占用的点视为可导航区域，而无需预构建导航地图。在此任务基础上，Liu *et al.* 开发了一种扩展的基线模型，该模型建立在传统的跨模态对齐（CMA）导航方法之上，为空中导航提供了初步解决方案。后续研究结合了 LLMs 以增强任务性能。例如，Gao *et al.* [[328]] 设计了一种基于 LLM 的端到端无人机视觉语言导航框架。该系统使用 GPT-4o 将自然语言指令分解为多个子目标，并结合 Grounding DINO [[192]] 和 Tokenize Anything (TAP)[[202]] 来提取语义掩码和视觉信息。RGB 图像和深度图像被转化为语义-拓扑-度量表示（STMR）。 通过设计的多模态提示，包括任务描述、历史轨迹和语义矩阵，GPT-4o 执行思维链推理以生成导航动作（方向、旋转角度和移动距离），显著提高了 AerialVLN 数据集上的导航成功率。

其他显著的研究包括由李等人提出的 CityNav 数据集及其伴随模型 MGP[[267]]。MGP 使用 GPT-3.5[[141]] 来解释地标名称、空间关系和任务目标，结合 Grounding DINO[[192]] 和 MobileSAM[[204]] 生成高精度的目标区域，以用于导航地图构建和路径规划。王等人 [[329]] 开发了一个 UAV VLN 系统框架，引入了新颖的基准任务 UAV-Need-Help，并通过 OpenUAV 模拟平台构建了相关数据集。他们的 UAV 导航 LLM 模型基于 Vicuna-7B[[243]] 和 EVA-CLIP[[189]]，提取视觉特征并采用分层轨迹生成机制以实现高效的自然语言导航。GOMAA-Geo[[2]] 框架通过将各种 LLM 与 CLIP[[186]] 整合，专注于多模态主动地理定位任务。 它充分利用多模态目标描述（如自然语言、地面图像和空中图像）及视觉线索，以实现高效且准确的目标定位，展现出出色的零样本泛化能力。NavAgent[[1]] 框架结合了先进的模型，如 LLaMA2[[154]]、BLIP-2[[176]]、GPT-4[[143]] 和 GLIP[[190]]。解析自然语言导航指令以提取地标描述，并利用经过微调的地标识别模块在全景图像中实现精确的地标定位。该框架在城市户外场景中的路径规划和导航任务中表现出色，为复杂环境中的无人机导航提供了强大的技术支持。相关研究，如 ASMA [[330]]、Zhang 等 [[331]] 和 Chen 等[[332]] 也探讨了户外环境下无人机 VLN 解决方案，值得进一步关注。

值得注意的是，刘等人*提出了体积环境表示（VER）* [[390]]，为无人机视觉导航（VLN）任务提供了一种创新视角。这种方法将环境划分为 3D 体素网格，将多视角的 2D 视觉特征聚合到 3D 空间中，以生成统一的环境表示。通过使用多任务学习框架，该系统可以预测 3D 占用状态、房间布局和物体边界框。基于 VER 的系统通过多层 Transformer 模块估计状态，并借助存储历史观测的记忆模块进行全局规划。局部和全局行动决策模块执行导航任务。这种 3D 空间表示方法不仅适用于室内 VLN 任务，还有潜力扩展到户外开放环境。通过对户外场景进行分割，并在各个分段之间应用相同的体素建模方法结合时间连接机制，该方法可能进一步支持复杂动态环境中的导航任务。

3.2.3 VLT

VLT 任务旨在基于多模态输入实现连续目标跟踪，同时动态调整飞行路径以应对目标遮挡和环境干扰等挑战。目前，VLT 任务通常利用多模态注意机制

有效整合视觉和语言信号。然而，与 VLN 和对象检测任务类似，它们仍面临跨模态特征对齐、泛化能力不足以及适应动态环境的挑战 [[391], [392]]。

Li 等 [[312]] 引入了 UAVNLT 数据集，并基于此开发了一种无人机自然语言跟踪（TNL）的基线方法。该方法中的视觉定位模块采用 CLIP [[186]]，利用其多模态特征在第一帧中精确定位目标。与 VLN 任务类似，VLT 任务将自然语言描述与目标边界框结合起来，使用自然语言作为辅助信息，以减少边界框引入的模糊性。TNL 系统中的自然语言描述清晰地指定了目标属性，帮助系统在复杂场景中准确识别和跟踪目标，从而有效解决动态环境下的跟踪挑战。Blei 等 [[333]] 提出了 CloudTrack，这是一种开放词汇的目标检测与跟踪系统，旨在无人机救援任务中使用。该系统采用云边协作架构，将 Grounding DINO [[192]] 与 VLM 结合，解析语义描述，使复杂目标的检测和过滤成为可能。 CloudTrack 为资源受限环境中的智能无人机感知和动态任务执行提供可靠的技术支持，展示了多模态技术在无人机智能任务中的潜力。

3.2.4 目标搜索

目标搜索任务整合了多模态目标感知与智能任务规划，代表了一种复杂的高级自主无人机任务。它可以被视为“视觉-语言导航 + 物体检测 + 高效路径规划”的结合。与传统的视觉-语言导航任务相比，目标搜索要求无人机在导航的同时有效地感知和定位目标 [[393], [394]]。

Cai 等人 [[334]] 提出了 NEUSIS 框架，这是一种神经-符号方法，用于复杂环境中的目标搜索任务，使得无人机能够在不确定的情况下执行自主感知、推理和规划。该框架包括三个主要模块：首先，感知、定位和 3D 推理模块（GRiD）整合了 VFMs 和神经-符号方法，例如用于动态视觉推理的 HYDRA[[184]]、用于目标属性分类的 CLIP[[186]]、用于开放集目标定位的 Grounding DINO[[192]] 以及用于高效实例分割的 EfficientSAM[[203]]，以完成目标检测、属性识别和 3D 投影等任务。其次，概率世界模型模块采用贝叶斯滤波和分布排名机制，通过融合噪声数据来维持概率目标地图和 3D 环境表示，从而支持动态目标定位和可靠报告生成。 最后，选择、导航和覆盖模块（SNaC）利用高级区域选择、中级路径导航和低级区域覆盖。通过 A $*$ 算法和基于信念图的优化方法，它生成有效的路径规划方案，确保无人机在有限的时间约束内

最大化目标搜索任务。Döschl 等人 [[335]] 提出了 Say-REAPEx 框架，用于无人机搜索与救援任务的在线任务规划与执行。该框架以 GPT-4o-mini 作为主要语言模型，并测试 Llama3 [[155]]、Claude3 [[168]] 和 Gemini [[152]] 以解析自然语言任务指令。它使用观测数据动态更新任务状态，并生成相应的行动计划。该框架还采用在线启发式搜索优化无人机任务路径，显著增强动态环境中的实时响应能力和自主决策能力。 Say-REAPEx 为复杂任务提供高效可靠的技术解决方案。

3.3 规划

标题	类型	出版物	代码
TypeFly (Typefly: Flying drones with large language model)（Typefly：使用大型语言模型控制无人机飞行）	LLM	*Arxiv*	-
SPINE (SPINE: Online Semantic Planning for Missions with Incomplete Natural Language Specifications in Unstructured Environments)（SPINE：在非结构化环境中针对不完整自然语言规范任务的在线语义规划）	LLM+VFM+VLM	*Arxiv*	-
LEVIOSA (LEVIOSA: Natural Language-Based Uncrewed Aerial Vehicle Trajectory Generation)（LEVIOSA：基于自然语言的无人机轨迹生成）	LLM	*MDPI*	*GitHub*
TPML (TPML: Task Planning for Multi-UAV System with Large Language Models)（TPML：使用大型语言模型进行多无人机系统任务规划）	LLM	*ICCA 2023*	-
REAL (Real: Resilience and adaptation using large language models on autonomous aerial robots)（REAL：在自主飞行机器人上使用大型语言模型实现弹性和适应性）	LLM	*Arxiv*	-
Liu et al. (Multi-Agent Formation Control Using Large Language Models)（使用大型语言模型进行多智能体编队控制）	LLM	*Techrxiv*	-
AutoHMA-LLM (AutoHMA-LLM: Efficient Task Coordination and Execution in Heterogeneous Multi-	LLM	*IEEE*	-

标题	类型	出版物	代码
Agent Systems Using Hybrid Large Language Models)（AutoHMA-LLM：使用混合大型语言模型实现异构多智能体系统中高效任务协调和执行）			
ACMA (Agent in the Sky: Intelligent Multi-Agent Framework for Autonomous HAPS Coordination and Real-World Event Adaptation)（天空中的智能体：用于自主高空平台（HAPS）协调和现实世界事件适应的智能多智能体框架）	LLM	*AAAI*	-
UAV-VLA (UAV-VLA: Vision-Language-Action System for Large Scale Aerial Mission Generation)（UAV-VLA：用于大规模空中任务生成的视觉-语言-动作系统）	LLM+VLM	*Arxiv*	-

传统无人机任务规划算法在复杂动态环境中面临适应性和协调性方面的重大挑战。多无人机系统的任务规划必须全面考虑每个无人机的能力、局限性和感知模式，同时满足能耗和避碰等约束，以实现高效的任务分配和路径规划[[395], [396]]。然而，尽管深度学习提供了新的技术方法，这些方法仍然表现出一些局限性，例如对大规模标注数据的高度依赖，对环境动态的实时适应能力不足，以及处理意外情况或未定义故障模式的能力有限。此外，针对固定任务或环境训练的模型往往难以很好地推广到不同的场景中 [[88], [397], [115]]。

LLMs 利用 CoT 框架 [[228]]，可以将复杂任务分解为一系列清晰可执行的子任务，从而提供明确的规划路径和逻辑框架。凭借上下文学习和少量学习的优势，LLMs 能够灵活适应多样的任务需求，并迅速生成有效的规划策略，即使在没有大规模标注数据的情况下 [[230], [231]]。此外，LLMs 在自然语言理解和生成方面的卓越表现使其能够通过语言指令与操作员进行实时协作，显著增强任务规划的智能性和操作灵活性。

TypeFly [[5]] 使用 GPT-4 [[143]] 解析用户提供的自然语言指令，并生成精确的任务规划脚本。它还引入了一种轻量级任务规划语言（MiniSpec）以优化生成任务所需的令牌数量，从而提高任务生成效率和响应速度。该框架集成了一个视觉编码模块，用于实时环境感知和动态任务调整，并包括一个“重新规划”机制，以应对执行过程中的环境变化。SPINE [[336]]，专为无结构环境中

的任务规划设计，结合了 GPT-4 和语义拓扑图，能够从不完整的自然语言任务描述中推理和动态规划。该框架采用 Grounding DINO [[192]] 进行物体检测，使用 LLaVA [[169], [170]] 来丰富语义信息，并利用递归地平线框架将复杂任务分解为可执行路径，从而实现动态调整和高效执行。 LEVIOSA [[337]] 通过自然语言生成无人机轨迹，使用 Gemini [[151], [152]] 或 GPT-4o 来解析用户文本或语音输入，将任务需求转化为高层次的航点规划。该框架结合了强化学习与多批评者共识机制来优化轨迹，确保计划符合安全性和能效要求。它实现了从自然语言到三维无人机轨迹的端到端自动化，支持动态环境适应和协作多无人机任务执行。类似的研究包括 TPML [[338]]、REAL [[6]] 以及刘 等[[340]] 的工作，进一步扩展了 LLMs 在无人机任务规划中的应用。

3.4 飞行控制

标题	类型	出版物	代码
PromptCraft (Chatgpt for robotics: Design principles and model abilities)（用于机器人的 ChatGPT：设计原则和模型能力）	LLM	*IEEE Access*	*GitHub*
Zhong et al. (A safer vision-based autonomous planning system for quadrotor uavs with dynamic obstacle trajectory prediction and its application with llms)（一种更安全的基于视觉的四旋翼无人机自主规划系统，具有动态障碍物轨迹预测及其在 LLM 中的应用）	LLM	*WACV 2024*	-
Tazir et al. (From words to flight: Integrating openai chatgpt with px4/gazebo for natural language-based drone control)（从文字到飞行：整合 OpenAI ChatGPT 与 PX4/Gazebo 进行基于自然语言的无人机控制）	LLM	*WCSE 2023*	-
Phadke et al. (Integrating Large Language Models for UAV Control in Simulated Environments: A Modular Interaction Approach)（在模拟环境中集成大型语言模型进行无人机控制：一种模块化交互方法）	LLM	*Arxiv*	-
EAI-SIM (EAI-SIM: An Open-Source Embodied AI Simulation Framework with Large Language Models)（EAI-SIM：一个带有大型语言模型的开源具身 AI 模拟框架）	LLM	*ICCA 2024*	*GitHub*
TAIiST (TAIiST CPS-UAV at the SBFT Tool Competition 2024)（TAIiST CPS-UAV 参加 SBFT 2024 工具竞赛）	LLM	*SBFT 2024*	*GitHub*

标题	类型	出版物	代码
Swarm-GPT (Swarm-gpt: Combining large language models with safe motion planning for robot choreography design)（Swarm-GPT：结合大型语言模型和安全运动规划进行机器人编舞设计）	LLM	*Arxiv*	-
FlockGPT (FlockGPT: Guiding UAV Flocking with Linguistic Orchestration)（FlockGPT：通过语言编排引导无人机集群飞行）	LLM	*Arxiv*	-
CLIPSwarm (CLIPSwarm: Generating Drone Shows from Text Prompts with Vision-Language Models)（CLIPSwarm：使用视觉语言模型从文本提示生成无人机表演）	VFM	*Arxiv*	-

无人机飞行控制任务通常分为两类：单无人机飞行控制和群体无人机飞行控制。在单无人机飞行控制中，模仿学习和强化学习方法逐渐成为主流，显示出在提高控制策略智能方面的显著潜力 [[398], [399], [400]]。然而，这些方法通常依赖于大规模标注数据，并在实时性能和安全性方面面临限制。在群体无人机飞行控制中，多智能体强化学习和图神经网络（GNN）等技术为多无人机协作任务提供了强大的建模能力，在编队飞行、任务分配和动态障碍物规避等场景中显示出优势 [[401], [402], [403], [404]]。尽管如此，这些方法仍然面临通信延迟、计算复杂性和全局优化能力等重大挑战。

与传统方法相比，基于 LLM 的飞行控制为该领域引入了全新的可能性。利用少样本学习能力，LLM 能够快速适应新的任务需求；其上下文学习能力使模型能够动态分析任务环境并生成高级飞行策略。此外，基于语义的自然语言交互显著提升了人机协作效率，支持无人机的任务规划、实时决策和复杂环境适应。尽管这一研究方向仍处于初期探索阶段，但在需要语义理解和高级决策的任务场景中，已经展示出巨大的潜力。

在单无人机飞行控制领域，早期的研究为将 LLMs 应用于这一任务奠定了重要基础。例如，Courbon 等 [[405]] 提出了一个基于视觉的导航策略，使用单目相机观察自然地标，通过将当前视觉图像与预录关键帧进行匹配，构建视觉记忆并实现未知环境中的自主导航。Vemprala 等 [[341]] 开发了 PromptCraft 平台，这是将 LLMs 应用于无人机飞行控制的开创性工作。该平台将 ChatGPT 与微软的 AirSim [[305]] 仿真环境集成。通过设计专门针对飞行控制的提示，并将 ChatGPT API 与 AirSim API 结合，能够实现自然语言驱动的飞行控制。提示设

计在此过程中起着关键作用，直接影响任务理解和指令生成的准确性。类似的研究还包括 Zhong 等 [[342]]，Tazir 等 的探索。 [[343]]，以及 Phadke 等 。[[344]]，还有 EAI-SIM [[345]] 和 TAIiST [[346]] 等框架的发展。

在无人机群飞行控制领域，Jiao 等[[347]] 提出了 Swarm-GPT 系统，该系统结合了 LLMs 和基于模型的安全运动规划，为无人机群飞行控制建立了一个创新框架。该系统使用 GPT-3.5 [[141]] 生成无人机的时间序列航路点，并通过安全规划模块优化路径，以满足物理约束和避碰要求。Swarm-GPT 允许用户通过重新提示动态修改飞行路径，从而实现无人机群的灵活编队和动态调整。此外，该系统在模拟环境中展示了轨迹规划的安全性和编队表演的艺术效果。类似研究包括 FlockGPT [[348]] 和 CLIPSwarm [[349]]，它们探索自动化和创造性的控制方案，以提高无人机群表演的效率和可操作性。

3.5 基础设施

标题	类型	出版物	代码
DTLLM-VLT (DTLLM-VLT: Diverse Text Generation for Visual Language Tracking Based on LLM)（DTLLM-VLT：基于 LLM 的视觉语言跟踪多样化文本生成）	VFM+LLM	*CVPR 2024*	-
Yao et al. (Can llm substitute human labeling? a case study of fine-grained chinese address entity recognition dataset for uav delivery)（LLM 能否取代人工标注？以无人机配送细粒度中文地址实体识别数据集为例）	LLM	*Companion Proceedings of the ACM Web Conference 2024*	*GitHub*
GPG2A (Cross-View Meets Diffusion: Aerial Image Synthesis with Geometry and Text Guidance)（跨视角与扩散模型相遇：基于几何和文本引导的航空图像合成）	LLM	*Arxiv*	*GitLap*
AeroVerse (AeroVerse: UAV-Agent Benchmark Suite for Simulating, Pre-training, Finetuning, and Evaluating Aerospace Embodied World Models)（AeroVerse：用于模拟、预训练、微调和评估航空航天具身世界模型的无人机-智能体基准套件）	VLM+LLM	*Arxiv*	-

标题	类型	出版物	代码
Tang et al. (Defining and Evaluating Physical Safety for Large Language Models)（定义和评估大型语言模型的物理安全）	LLM	*Arxiv*	*Hugging face*
Xu et al. (Emergency Networking Using UAVs: A Reinforcement Learning Approach with Large Language Model)（使用无人机进行应急组网：基于大型语言模型的强化学习方法）	LLM	*IPSN 2024*	-
LLM-RS (Real-time Integration of Fine-tuned Large Language Model for Improved Decision-Making in Reinforcement Learning)（实时集成微调大型语言模型以改进强化学习中的决策）	LLM	*IJCNN 2024*	-
Pineli et al. (Evaluating Voice Command Pipelines for Drone Control: From STT and LLM to Direct Classification and Siamese Networks)（评估用于无人机控制的语音命令管线：从语音转文本（STT）和 LLM 到直接分类和 Siamese 网络）	LLM	*Arxiv*	-

数据集的构建和处理在无人机系统的基础研究中尤其重要。高质量的数据资源和完善的数据处理工作流程对于确保 LLM、VLM 和 VFM 技术在无人机任务中的高效应用至关重要。这些研究工作不仅为无人机在多模态任务中的应用奠定了坚实的基础，还为相关领域的技术创新和方法进步提供了有力支持。

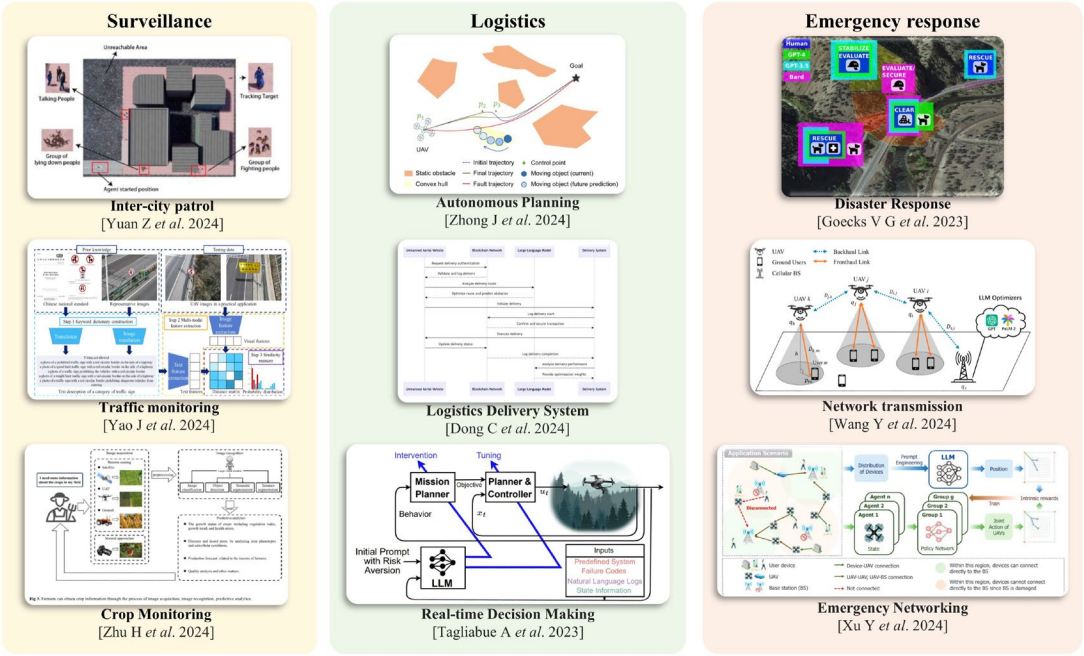
DTLLM-VLT [[350]] 是一个旨在通过多粒度文本生成来增强 VLT 性能的框架。该框架使用 SAM [[198]] 提取目标的分割掩膜，结合 Osprey [[208]] 生成初步的视觉描述。随后，LLaMA [[153], [154], [155]] 或 Vicuna [[156]] 生成四种类型的细粒度文本注释：初步简要描述、初步详细描述、密集简要描述和密集详细描述，涵盖目标类别、颜色、动作和动态变化。这些高质量的文本数据显著增强了多模态任务的语义支持，提高了跟踪的准确性和鲁棒性，同时降低了语义注释的时间和成本。姚 *等*。 [[271]] 为无人机配送系统开发了 CNER-UAV 数据集，以实现细粒度的中文命名实体识别，利用 GPT-3.5 [[141]] 和 ChatGLM [[161], [162], [163]] 实现了精准的地址信息识别。

无人机系统面临的一个显著挑战是获取航拍图像的高成本和劳动密集型工作。为了解决这个问题，Arrabi 等 [[351]] 提出了 GPG2A 模型，该模型利用地面到空中的 (G2A) 技术从地面图像合成航拍图像，克服了由显著视角差异带来的生成挑战。该模型采用了一个双阶段生成框架：第一阶段使用 ConvNeXt-B[[406]] 提取地面图像特征，并应用极坐标转换生成鸟瞰图 (BEV) 布局图，以明确捕捉场景几何。第二阶段引入了扩散模型，通过将 BEV 布局图与文本描述结合来生成高质量的航拍图像。文本描述由 Gemini [[151], [152]] 生成，并使用 BERT [[407]] 优化为动态文本提示，从而增强生成图像的语义相关性和场景一致性。 该方法有效解决了视角转换的挑战，并为高效获取航空图像提供了创新的解决方案，具有重要的实用价值。

在框架和平台方面，相关研究展示了多样的发展方向。 Yao 等 [[352]] 提出了 AeroVerse，这是一种高参考性的航空智能基准套件，专为无人机代理而设计。AeroVerse 集成了模拟器、数据集、任务定义和评估方法，旨在推动无人机在感知、认知、规划和决策等技术方面的发展。其系统架构包括基于虚幻引擎和 AirSim 构建的高精度仿真平台 AeroSimulator。AeroSimulator 生成跨越真实和虚拟场景的多模态数据集，并为五个核心任务定制精细的数据集：场景感知、空间推理、导航探索、任务规划和运动决策。

此外，一些创新框架将 LLMs 与特定无人机任务结合在一起。例如，Tang 等 [[353]] 开发了一个无人机控制的安全评估框架；Xu 等 [[354]] 设计了一个用于无人机在动态环境中部署的紧急通信网络优化框架；LLM-RS [[355]] 专注于无人机空战模拟任务，结合奖励设计和决策优化以增强系统性能；Pineli 等 [[356]] 提出了一个无人机语音控制框架，利用自然语言处理技术最大化人机交互的潜力。这些工作从各个维度为无人机技术的发展做出了贡献，为无人机智能和任务多样化提供了必要的支持。

4 基于基础模型的无人机应用场景



本节重点关注将无人机（UAVs）与大规模语言模型（LLMs）结合的实际应用场景。大规模语言模型（LLMs）为多模态数据提供了先进的认知和分析能力，包括图像、音频、文本，甚至视频数据。与集成了传统机器学习算法的无人机相比，将大规模语言模型（LLMs）纳入无人机系统显著增强了它们的环境感知能力 [[408], [409]]，使得决策过程更加智能 [[410]]，并通过利用大规模语言模型（LLMs）在人与机器交互中的强理解能力来改善用户体验 [[344], [411]]。

根据现有文献，我们介绍了如图 [5] 所示的 FM 与 UAV 整合的典型工作：监视、物流和应急响应。这三个呈现的类别并不 exhaustive 于所有 UAV 应用，而是代表了 UAV 技术与先进模型能力结合的当前有效领域。它们专注于提升三个关键能力：环境感知、自主决策和人机交互。

4.1 监视

在监视方面，无人机被用于监测交通场景、城市环境和其他监管任务。传统的无人机监测应用方法主要依赖于机器学习技术。近年来，该领域进行了大量研究，包括车辆轨迹监测 [[412]]、道路状况监测 [[413], [414]]、路边单元 (RSUs) 通信 [[415]] 以及城市场景中的应用和管理 [[416]]。然而，H. Menouar 等 [[417]] 指出，无人机在智能交通系统 (ITS) 和智慧城市中预计将发挥重要作用，但其有效性将依赖于更大的自主性和自动化。同样，Wang L 等 [[418]] 强调了

无人机在城市管理中的重要性，并突出了自动化和人机交互等挑战。 FMs 的出现最近引发了研究，探索 FMs 与 UAVs 的结合如何提升它们的可用性和任务性能。

在城市场景监测中，Yao J *et al*. [[419]] 迅速部署了 VLMs，通过多模态学习和大规模预训练网络监测交通标志的状态，在准确性和成本效益方面都取得了优秀的成果。与 FMs 集成的 UAVs 在车辆检测、车辆分类、行人检测、自行车检测、速度估计和车辆计数等任务中表现卓越。Yuan Z *et al*. [[420]] 提出了“巡逻代理”，该代理利用 VLMs 进行视觉信息获取，利用 LLMs 进行分析和决策。这使得 UAVs 能够自主进行城市巡逻、识别和跟踪任务。此外，与 LLMs 集成的 UAVs 在其他监测任务中也展示了出色的性能。在农业应用中，Zhu H *et al*. [[421]] 提出使用 LLMs 和 VLMs 帮助农民提高生产力和产量。

4.2 物流

在物流领域，无人机（UAV）使整个物流链中的智能流程得以实现，从决策制定到路线规划和最终交付 [[422]]。无人机在物流和配送中的应用是当前研究的关键领域。Jiang H 等[[423]] 使用先进的优化算法优化了无人机的调度和路线规划。Huang H 等 [[424]] 提出了涉及无人机和公共交通系统（如电车）的协作调度解决方案，证明其为 NP 完全问题。他们还引入了一种基于动态编程的精确算法来应对这一挑战。然而，无人机物流仍面临若干挑战。Wandelt S 等 [[425]] 确定了两个主要问题：自主导航与人机交互，以及实时数据分析。引入 FMs 提供了一种新的方法来应对这些挑战，通过 FMs 的推理和决策能力提升无人机的实时决策和规划能力。 此外，FMs 强大的理解能力提高了人机交互，提供了更好的用户体验。

针对使用 FMs 的物流应用，Tagliabue A 等 [[6]] 提出了一个名为 REAL 的框架，利用 LLMs 的先验知识并采用零次提示。这种方法显著提高了无人机的适应性和决策能力，改善了位置控制性能和实时任务决策能力。Luo S 等[[426]] 利用 LLMs 处理用户提供的地址信息。由于传统方法在细粒度处理方面因用户输入缺乏精确性而面临挑战，他们对 LLMs 进行了微调，以解决这个问题，从而提高了无人机交付系统的自动化水平和处理效率。Zhong J 等 [[342]] 专注于自主无人机规划，提出了一个与 LLMs 集成的基于视觉的规划系统。该系统结

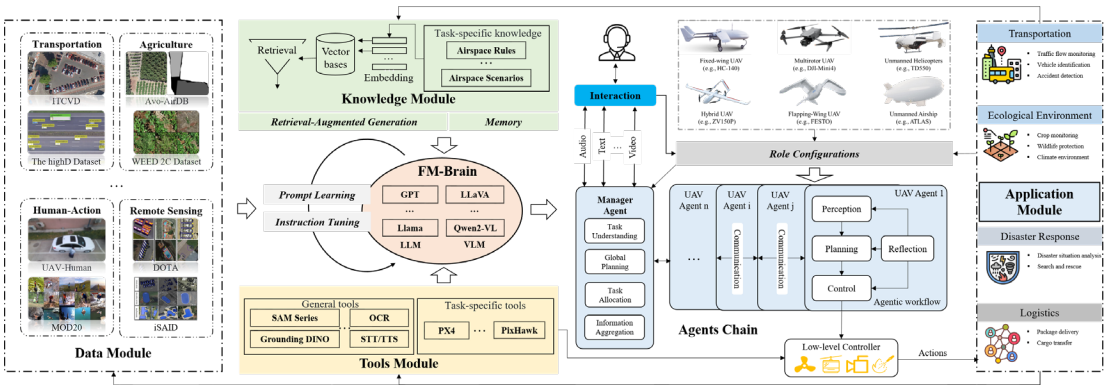
合了动态障碍物跟踪和轨迹预测，以实现高效可靠的自主飞行。此外，集成 LLMs 增强了人机交互，改善了整体用户体验。Dong C 等。 [[427]] 从供应链的角度出发，提出了一种创新的无人机物流智能配送系统。通过整合区块链技术，他们确保了系统的安全性和透明性。此外，他们利用 LLMs 进行路线优化和动态任务管理，并通过自然语言交互提供客户支持服务，为未来开发安全高效的无人机配送系统提供了框架。

4.3 紧急响应

无人机在应急响应和灾难救援任务中具有固有优势 [[69]]。它们高度灵活的操作能力使其适用于大多数应急场景。金 W 等[[428]] 分析了基于无人机的应急响应机制的需求，评估了灾难类型和无人机的性能特征，并提供了建议。通过配备不同的有效载荷和物资，无人机可以根据特定的灾难场景和任务要求提供定制化支持。戈克斯 V G 等 [[429]] 介绍了基于 LLM 的模型 DisasterResponse GPT，该模型利用上下文学习加速灾难响应，通过生成可行计划并实时快速更新和调整，从而实现快速决策。德库尔托 J 等 [[408]] 利用无人机提供即时视觉反馈和高数据吞吐量的能力，开发了一种场景理解模型，结合了 LLM 和 VLM。这种方法以成本效益的方式增强了无人机实时决策能力，以处理复杂和动态的数据。此外，他们集成了多个传感器，以自主执行复杂任务。

除了救援任务外，无人机越来越多地被研究作为建立通信网络的工具，以应对灾区或偏远地区的连接挑战。这些网络支持依赖网络的任务和离线紧急响应。Fourati F 等 [[430]] 突出了人工智能在通信工程中的关键作用，包括流量预测和信道建模等应用。Xu Y 等[[354]] 利用无人机作为移动接入点，以帮助城市通信系统在灾难场景中进行紧急网络部署。他们进一步采用 LLMs 来增强建模过程并加速优化工作流。Wang Y 等 [[431]] 使用结构化提示与 LLMs 优化无人机群的部署。与传统方法相比，他们的方法在确保强大的网络连接性和服务质量方面，通过精确的无人机定位降低了迭代次数。 LLM 驱动的框架简化了无人机网络运营商的操作挑战，为其在更复杂的现实场景中的应用铺平了道路。

5 自主无人机：集成基础模型与无人机系统的通用流程



本节系统地探讨了 LLMs 和 VLMs 在传统无人机（UAV）管道和任务中的整合。从人工智能代理的角度，我们提出了一个代理无人机（Agentic UAV）的框架，该框架将基础模型（FMs）与无人机系统相结合，如图 [6] 所示。该框架包括五个关键组件：数据模块、知识模块、工具模块、基础模型模块和代理模块。数据模块专注于创建新数据集或将现有数据调整为适合于微调和训练特定于无人机任务的基础模型的格式。知识模块储存领域特定的信息，例如空域法规和场景库，这些信息对于无人机操作至关重要。工具模块包含满足无人机任务所需的领域特定工具或 API，从而扩展代理的解决问题能力。基础模型模块专注于微调基础模型，以提高其在无人机相关领域的适应性和性能。代理模块旨在创建包含感知、规划和行动的工作流程，以完成无人机任务。该模块还建立了反思机制，以根据任务执行的反馈优化过程。 此外，考虑到无人机群的频繁使用，代理模块集成了多代理设计、交互和通信单元。为了协调和管理这些代理，框架引入了一个管理代理，负责全球任务的规划和分配。以下各小节将详细阐述这些模块。

5.1 数据模块

数据模块专注于将与无人机（UAV）相关的数据转化为适合微调和训练针对无人机特定任务的语言模型（FMs）的对话或问答格式。无人机数据大体上可以分为由无人机生成的多模态传感器数据（例如，图像、视频、激光雷达、GPS、惯性测量单元）和操作员提供的自然语言指令。然而，原始数据集中常常缺乏自然语言指令，必须通过人工标注或自动化方法生成。

生成自然语言指令通常涉及利用图像描述模型或类似工具为传感器数据创建描述性或基于问题的注释，例如生成有关无人机图像中特定物体或事件的问题。先进的语言模型，如基于 GPT 的模型，进一步实现了多样化和上下文丰富的指令的自动生成，减少了对人工工作的依赖，并显著扩大了数据集构建的范围。这些方法确保自然语言指令与多模态数据相一致，促进了需要感知和推理的任务中语言模型的整合。

构建特定于无人机的数据库对模型的训练和微调至关重要。例如，已经开发了专门针对无人机导航和地理定位任务的基准数据集，如 Chu 等人 [[388]] 提出的数据集，它通过文本-图像-边界框注释扩展了现有资源，以提高地理定位的准确性。同样，Yao 等人 [[271]] 引入了一种细粒度的中文地址识别数据集，以支持无人机递送，提高城市环境中的导航精度。此外，在遥感应用中，无人机影像被广泛用于目标检测、语义分割和环境监测等任务，多模态大模型显著提高了任务的效率和准确性 [[432]]。

5.2 基础模型

UAV 任务的基础模型关注两个核心方面：选择合适的模型和针对特定任务优化这些模型。这种模块化的方法确保 UAV 系统能够有效地处理多样化和复杂的场景，同时保持执行效率。

5.2.1 模型选择

该过程始于识别任务类型，并确定数据是否涉及单模态或多模态输入。对于基于语言的任务，LLMs 如 ChatGPT 和 LLama 提供了推理、决策和自然语言交互的坚实基础。对于涉及视觉和语言数据的多模态任务，VLMs 如 GPT-4V、LLaVa 和 Qwen2-VL 通常是理想选择。这些模型作为基础组件，为智能代理提供了能力支撑。

除了基于语言和视觉的模型，最近的进展还探索了大型 3D 模型，这些模型与在 3D 环境中操作的无人驾驶飞行器（UAV）特别相关。这些模型结合了具有解释 3D 数据和规划任务能力的 FMs。例如，Hong 等人 [[433]] 提出了一个能够进行密集字幕生成、3D 问答和使用点云导航的 3D LLM。同样，Agent3D-Zero [[434]] 采用线集合提示（SoLP）来通过生成多样的观察视角来增强场景几

何理解。虽然目前大多数研究集中在室内和封闭环境中，但将这些模型扩展到开放和动态的 UAV 场景中，展现了令人兴奋的未来机遇。

5.2.2 模型优化

一旦选择了基础模型，就通过如指令微调和提示微调等方法进行优化和调整，以满足无人机特定的要求。提示微调是一种简单的方法，涉及创建任务特定的模板，将任务背景知识（如目标、环境特征和任务分解）嵌入模型的交互中。少样本学习可以通过使用精心策划的示例来补充这一过程，帮助模型理解任务特定目标。对于复杂挑战，例如多阶段规划或动态场景理解，链式思维方法 [[228]] 将任务分解为顺序子任务，从而改善推理和执行。

指令微调通过生成针对无人机特定任务的数据集提供了进一步的适应性。例如，在视觉语言导航（VLN）任务中，数据集可以包括与物体检测或无人机任务中的导航相关的问题-答案对。像低秩适应（LoRA）[[435]] 这样的技术通过仅微调一部分参数来优化这些模型，保持计算效率，同时提高性能。此外，层冻结技术可以保留预训练知识，并最小化对小型任务特定数据集的过拟合。

基于指令微调，来自人类反馈的强化学习（RLHF）[[436]] 增强了模型与人类偏好和操作需求的对齐。通过 incorporación 从人类反馈中派生的奖励信号，RLHF 使模型能够应对动态无人机挑战，例如路径生成、任务调整和关键物体检测。这种方法提高了无人机控制的实时响应能力和自动化水平，最终改善了任务效率和适应性。

5.3 知识模块

检索增强生成（RAG）是一项新兴技术，它集成了检索和生成的能力。其核心功能在于从知识库中检索相关信息，并将其与生成模型的输出融合，从而提高所生成结果的准确性和领域适应性。RAG 模型利用检索模块从外部知识库获取与输入内容相关的信息，并将其作为生成模块的上下文。这种方法提高了生成输出的质量和可靠性。与传统生成模型不同，RAG 引入了一种实时检索机制，以缓解“幻觉”问题，即模型因背景知识不足而生成不正确或虚构的信息。此外，RAG 的模块化架构允许知识库和生成模型的独立更新，从而增加系统灵活性，并确保生成中使用的信息的时效性和准确性。因此，RAG 在需要高度专业化、实时信息处理或个性化输出的任务中展现出显著潜力。

为 UAV 特定任务量身定制 RAG 系统至关重要，因为无人机操作涉及多样化和复杂的场景。首先，RAG 可以实时获取最新的环境数据，例如气象条件、地形信息和空中交通更新，这对于航班规划和导航等任务至关重要。其次，将特定领域的知识库整合到 RAG 框架中使无人机能够执行高级决策任务，比如在动态环境中进行自主任务调整或在监视任务中识别未知物体。最后，RAG 可以通过检索上下文数据来澄清查询或增强系统决策的可解释性，从而促进与人类操作员的互动。例如，在基于无人机的环境监测任务中，RAG 可以检索污染水平或土地使用模式的历史数据，将其与当前传感器数据结合，并生成综合报告。这些能力表明，一个构建良好的 RAG 框架如何提升无人机系统的效率、准确性和适应性，为更智能和自主的无人机应用铺平道路。

5.4 工具模块

工具模块旨在提供通用功能和特定任务能力，以支持无人机操作。

5.4.1 一般工具

通用工具专注于广泛的多模态功能，以增强无人机系统的感知和互动能力。在这些工具中，视觉功能模型（VFM）作为解决多样视觉任务的基石，充分发挥其卓越的泛化能力和零样本学习能力。与强调推理和决策的基础模型（FM）不同，VFM 在理解特定视觉任务方面表现卓越，使其成为基础工具而非核心的“FM-大脑”组件。

VFMs 在无人机任务中提供了显著优势，能够与特定任务要求相匹配。例如，CLIP 系列因其强大的多模态对齐能力，非常适合物体识别和场景理解任务，能够实现开放词汇的物体检测和分类。因其零样本分割能力而闻名的 Segment Anything Model (SAM) 在各种环境和目标的图像分割方面表现理想。Grounding DINO 在物体检测和定位任务中表现出色，能够在动态场景中提供高效的目标跟踪和检测。这些模型可以独立处理特定任务，或者与 LLMs 或 VLMs 集成，以增强无人机系统在任务规划、导航和环境感知方面的智能。

此外，VFM 可以进行微调以适应无人机特定场景。例如，在专用数据集上微调 Grounding DINO 模型可以提高其在复杂多目标跟踪任务中的性能。此外，VFM 可以与传统机器学习或深度学习模型协作，形成“ 大模型 + 小模型 ”策

略，以平衡泛化能力与任务特定效率。例如，VFM 提取全球语义信息，而较小的模型则专注于细微细节，从而实现全球分析与局部分析的有效结合。

VFMs 的另一个创新应用是用于生成针对 VLM 的指令微调数据集。通过利用 VFM 输出的图像标题、分割描述和物体深度信息，这些数据集可以用于训练针对 UAV 特定任务的 VLM。例如，Chen *等人* [[437]] 使用来自 VFMs 的互联网规模空间推理数据创建了一个 3D 空间指令微调数据集，训练了 SpatialVLM 模型。这种方法突显了 VFMs 在为大型模型生成高质量数据集方面的潜力，显著提升了 UAV 系统的动态感知和任务规划能力。

5.4.2 任务特定工具

任务专用工具是针对无人机中心操作量身定制的，着重于飞行控制和任务执行。关键组件包括 PX4 和 Pixhawk，这些是广泛使用的开源飞行控制器。这些工具为无人机提供精确的控制、任务规划和实时适应能力，使其在复杂的空中任务中不可或缺。通过将这些专业工具与通用功能相结合，无人机系统在应对任务特定挑战方面实现了高度的灵活性和效率。

5.5 代理模块

代理模块旨在为无人机系统提供智能决策和任务执行能力。它整合了高层协调与任务特定的代理工作流程，以优化无人机在复杂任务中的操作。

5.5.1 管理代理

管理代理负责无人机群内的高层任务协调和调度，确保多个无人机高效执行任务。该代理承担全球规划和整体任务分配的角色，将大型任务分解为更小的可管理子任务，然后分配给各个无人机。此外，全球代理监控无人机群的状态，并根据实时反馈动态调整任务分配，确保每个无人机在更广泛的任务背景下有效运行。

5.5.2 UAV 特定的代理工作流程

群中的每个无人机遵循一个自主的代理工作流程，该流程由一系列代理组成，旨在处理感知、规划和控制任务。这些代理依次操作，确保每个无人机有效处理必要的数据并执行其任务目标。感知代理首先处理传感器数据，使用先

进的视觉基础模型（VFM）识别障碍物、物体和兴趣点，如用于物体识别的 CLIP、用于分割的 SAM 以及用于定位的 Grounding DINO。

接下来，规划代理从感知代理获取数据，以生成优化的飞行路径和任务策略，确保无人机能够有效地导航环境并完成指定任务。最后，控制代理将计划转化为可执行的指令，控制无人机的飞行和任务执行。

该工作流程允许每个无人机独立运行，同时也为整体任务目标做出贡献。此外，无人机特定的代理工作流程能够适应各种无人机任务，从搜索和救援到监视，通过根据每项任务的具体要求微调代理的能力。这种适应性增强了无人机在处理复杂动态环境中的效率。

5.5.3 代理协作与适应性

全球代理与无人机特定代理之间的协作对于优化任务执行至关重要。全球代理提供指导整体任务策略的高层指令。这些指令被 Individual UAV 代理分解为详细的执行计划，确保每个无人机能够自主操作，同时为集体任务目标做出贡献。无人机代理与全球代理进行沟通，以接收更新的指令并报告进度，从而实现任务的持续适应及根据实时数据和变化条件对任务计划进行动态调整。

此外，群体中的无人机代理可以相互互动以交换信息和协调行动。这种点对点通信使无人机能够根据共享的情境意识调整其行为，例如在多个无人机必须避免碰撞或合作完成联合任务时。例如，一个无人机可能会与另一个无人机共享其感知数据，以调整飞行路径或实时同步任务。这种互动确保无人机群体协调运行，每个代理根据来自其他代理的全局指导和局部实时信息调整其行动。

References

1. Liu et al. [2024]Y. Liu, F. Yao, Y. Yue, G. Xu, X. Sun, K. Fu,Navagent: Multi-scale urban street view fusion for uav embodied vision-and-language navigation,arXiv preprint arXiv:2411.08579 (2024).
2. Sarkar et al. [2024]A. Sarkar, S. Sastry, A. Pirinen, C. Zhang, N. Jacobs, Y. Vorobeychik,Gomaa-geo: Goal modality agnostic active geo-localization,arXiv preprint arXiv:2406.01917 (2024).
3. Liu et al. [2024]J. Liu, J. Cui, M. Ye, X. Zhu, S. Tang,Shooting condition insensitive unmanned aerial vehicle object detection,Expert Systems with Applications 246 (2024) 123221.
4. Qiu et al. [2024]H. Qiu, J. Li, J. Gan, S. Zheng, L. Yan,Dronegpt: Zero-shot video question answering for drones,in: Proceedings of the International Conference on Computer Vision and Deep Learning, 2024, pp. 1–6.
5. Chen et al. [2023]G. Chen, X. Yu, N. Ling, L. Zhong,Typefly: Flying drones with large language model,arXiv preprint arXiv:2312.14950 (2023).
6. Tagliabue et al. [2023]A. Tagliabue, K. Kondo, T. Zhao, M. Peterson, C. T. Tewari, J. P. How,Real: Resilience and adaptation using large language models on autonomous aerial robots,arXiv preprint arXiv:2311.01403 (2023).
7. Panagiotou and Yakinthos [2020]P. Panagiotou, K. Yakinthos,Aerodynamic efficiency and performance enhancement of fixed-wing uavs,Aerospace Science and Technology 99 (2020) 105575.
8. Villa et al. [2020]D. K. Villa, A. S. Brandao, M. Sarcinelli-Filho,A survey on load transportation using multirotor uavs,Journal of Intelligent & Robotic Systems 98 (2020) 267–296.
9. Rashad et al. [2020]R. Rashad, J. Goerres, R. Aarts, J. B. Engelen, S. Stramigioli,Fully actuated multirotor uavs: A literature review,IEEE Robotics & Automation Magazine 27 (2020) 97–107.

10. Alvarenga et al. [2015]J. Alvarenga, N. I. Vitzilaios, K. P. Valavanis, M. J. Rutherford,Survey of unmanned helicopter model-based navigation and control techniques,Journal of Intelligent & Robotic Systems 80 (2015) 87–138.
11. Saeed et al. [2018]A. S. Saeed, A. B. Younes, C. Cai, G. Cai,A survey of hybrid unmanned aerial vehicles,Progress in Aerospace Sciences 98 (2018) 91–105.
12. Du et al. [2024]H. Du, L. Ren, Y. Wang, X. Cao, C. Sun,Advancements in perception system with multi-sensor fusion for embodied agents,Information Fusion (2024) 102859.
13. Martinez-Carranza and Rascon [2020]J. Martinez-Carranza, C. Rascon,A review on auditory perception for unmanned aerial vehicles,Sensors 20 (2020) 7276.
14. Zhang et al. [2023]J. Zhang, S. Xu, Y. Zhao, J. Sun, S. Xu, X. Zhang,Aerial orthoimage generation for uav remote sensing,Information Fusion 89 (2023) 91–120.
15. Mittal et al. [2020]P. Mittal, R. Singh, A. Sharma,Deep learning-based object detection in low-altitude uav datasets: A survey,Image and Vision computing 104 (2020) 104046.
16. Liu et al. [2020]M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, C. Piao,Uav-yolo: Small object detection on unmanned aerial vehicle perspective,Sensors 20 (2020) 2238.
17. Girisha et al. [2019]S. Girisha, M. P. MM, U. Verma, R. M. Pai,Semantic segmentation of uav aerial videos using convolutional neural networks,in: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), IEEE, 2019, pp. 21–27.
18. Liu et al. [2021]S. Liu, J. Cheng, L. Liang, H. Bai, W. Dang,Light-weight semantic segmentation network for uav remote sensing images,IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021) 8287–8296.
19. Li et al. [2010]Z. Li, N. Hovakimyan, V. Dobrokhodov, I. Kaminer,Vision-based target tracking and motion estimation using a small uav,in: 49th IEEE Conference on Decision and Control (CDC), IEEE, 2010, pp. 2505–2510.
20. Dobrokhodov et al. [2006]V. N. Dobrokhodov, I. I. Kaminer, K. D. Jones, R. Ghabcheloo,Vision-based tracking and motion estimation for moving targets using small uavs,in: 2006 American Control Conference, IEEE, 2006, pp. 6–pp.
21. Mascaro et al. [2018]R. Mascaro, L. Teixeira, T. Hinzmann, R. Siegwart, M. Chli,Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation,in: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, 2018, pp. 1421–1428.
22. Wan et al. [2022]L. Wan, R. Liu, L. Sun, H. Nie, X. Wang,Uav swarm based radar signal sorting via multi-source data fusion: A deep transfer learning framework,Information Fusion 78 (2022) 90–101.
23. Rezwan and Choi [2022]S. Rezwan, W. Choi,Artificial intelligence approaches for uav navigation: Recent advances and future challenges,IEEE access 10 (2022) 26320–26339.
24. Gyagenda et al. [2022]N. Gyagenda, J. V. Hatilima, H. Roth, V. Zhmud,A review of gnss-independent uav navigation techniques,Robotics and Autonomous Systems 152 (2022) 104069.
25. Balamurugan et al. [2016]G. Balamurugan, J. Valarmathi, V. Naidu,Survey on uav navigation in gps denied environments,in: 2016 International conference on signal processing, communication, power and embedded system (SCOPES), IEEE, 2016, pp. 198–204.
26. McEnroe et al. [2022]P. McEnroe, S. Wang, M. Liyanage,A survey on the convergence of edge computing and ai for uavs: Opportunities and challenges,IEEE Internet of Things Journal 9 (2022) 15435–15459.
27. Neumann and Bartholmai [2015]P. P. Neumann, M. Bartholmai,Real-time wind estimation on a micro unmanned aerial vehicle using its inertial measurement unit,Sensors and Actuators A: Physical 235 (2015) 300–310.
28. Barbieri et al. [2019]L. Barbieri, S. T. Kral, S. C. Bailey, A. E. Frazier, J. D. Jacob, J. Reuder, D. Brus, P. B. Chilson, C. Crick, C. Detweiler, et al.,Intercomparison of small unmanned aircraft system (suas) measurements for atmospheric science during the lapse-rate campaign,Sensors 19 (2019) 2179.
29. Couturier and Akhloufi [2021]A. Couturier, M. A. Akhloufi,A review on absolute visual localization for uav,Robotics and Autonomous Systems 135 (2021) 103666.
30. Rovira-Sugranes et al. [2022]A. Rovira-Sugranes, A. Razi, F. Afghah, J. Chakareski,A review of ai-enabled routing protocols for uav networks: Trends, challenges, and future outlook,Ad Hoc Networks 130 (2022) 102790.
31. Atif et al. [2021]M. Atif, R. Ahmad, W. Ahmad, L. Zhao, J. J. Rodrigues,Uav-assisted wireless localization for search and rescue,IEEE Systems Journal 15 (2021) 3261–3272.
32. Lu et al. [2018]Y. Lu, Z. Xue, G.-S. Xia, L. Zhang,A survey on vision-based uav navigation,Geo-spatial information science 21 (2018) 21–32.
33. Gupta and Fernando [2022]A. Gupta, X. Fernando,Simultaneous localization and mapping (slam) and data fusion in unmanned aerial vehicles: Recent advances and challenges,Drones 6 (2022) 85.
34. Kassas et al. [2024]Z. M. Kassas, N. Khairallah, J. J. Khalife, C. Lee, J. Jurado, S. Wachtel, J. Duede, Z. Hoeffner, T. Hulsey, R. Quirarte, et al.,Aircraft navigation in gnss-denied environments via radio slam with terrestrial signals of opportunity,IEEE Transactions on Intelligent Transportation Systems (2024).
35. Tisdale et al. [2009]J. Tisdale, Z. Kim, J. K. Hedrick,Autonomous uav path planning and estimation,IEEE Robotics & Automation Magazine 16 (2009) 35–42.
36. Goerzen et al. [2010]C. Goerzen, Z. Kong, B. Mettler,A survey of motion planning algorithms from the perspective of autonomous uav guidance,Journal of Intelligent and Robotic Systems 57 (2010) 65–100.
37. Hong et al. [2021]Y. Hong, S. Kim, Y. Kim, J. Cha,Quadrotor path planning using a* search algorithm and minimum snap trajectory generation,ETRI Journal 43 (2021) 1013–1023.
38. Chai et al. [2022]X. Chai, Z. Zheng, J. Xiao, L. Yan, B. Qu, P. Wen, H. Wang, Y. Zhou, H. Sun,Multi-strategy fusion differential evolution algorithm for uav path planning in complex environment,Aerospace Science and Technology 121 (2022) 107287.
39. Xiao et al. [2021]S. Xiao, X. Tan, J. Wang,A simulated annealing algorithm and grid map-based uav coverage path planning method for 3d reconstruction,Electronics 10 (2021) 853.
40. Ait-Saadi et al. [2022]A. Ait-Saadi, Y. Meraihi, A. Soukane, A. Ramdane-Cherif, A. B. Gabis,A novel hybrid chaotic aquila optimization algorithm with simulated annealing for unmanned aerial vehicles path planning,Computers and Electrical Engineering 104 (2022) 108461.
41. Phung and Ha [2021]M. D. Phung, Q. P. Ha,Safety-enhanced uav path planning with spherical vector-based particle swarm optimization,Applied Soft Computing 107 (2021) 107376.
42. Yu et al. [2022]Z. Yu, Z. Si, X. Li, D. Wang, H. Song,A novel hybrid particle swarm optimization algorithm for path planning of uavs,IEEE Internet of Things Journal 9 (2022) 22547–22558.
43. He et al. [2021]W. He, X. Qi, L. Liu,A novel hybrid particle swarm optimization for multi-uav cooperate path planning,Applied Intelligence 51 (2021) 7350–7364.

44. Yang et al. [2023]Y. Yang, X. Xiong, Y. Yan,Uav formation trajectory planning algorithms: A review,Drones 7 (2023) 62.
45. Liu et al. [2021]H. Liu, J. Ge, Y. Wang, J. Li, K. Ding, Z. Zhang, Z. Guo, W. Li, J. Lan,Multi-uav optimal mission assignment and path planning for disaster rescue using adaptive genetic algorithm and improved artificial bee colony method,in: Actuators, volume 11, MDPI, 2021, p. 4.
46. Han et al. [2022]Z. Han, M. Chen, S. Shao, Q. Wu,Improved artificial bee colony algorithm-based path planning of unmanned autonomous helicopter using multi-strategy evolutionary learning,Aerospace Science and Technology 122 (2022) 107374.
47. Pan et al. [2021]Y. Pan, Y. Yang, W. Li,A deep learning trained by genetic algorithm to improve the efficiency of path planning for data collection with multi-uav,Ieee Access 9 (2021) 7994–8005.
48. Cui and Wang [2021]Z. Cui, Y. Wang,Uav path planning based on multi-layer reinforcement learning technique,Ieee Access 9 (2021) 59486–59497.
49. Heidari et al. [2023]A. Heidari, N. Jafari Navimipour, M. Unal, G. Zhang,Machine learning applications in internet-of-drones: Systematic review, recent deployments, and open issues,ACM Computing Surveys 55 (2023) 1–45.
50. He et al. [2021]L. He, N. Aouf, B. Song,Explainable deep reinforcement learning for uav autonomous path planning,Aerospace science and technology 118 (2021) 107052.
51. Zhu et al. [2021]B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, J. Henry,Uav trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning,IEEE Transactions on Vehicular Technology 70 (2021) 9540–9554.
52. Guo et al. [2023]Y. Guo, X. Liu, Q. Jia, X. Liu, W. Zhang,Hpo-rrt*: A sampling-based algorithm for uav real-time path planning in a dynamic environment,Complex & Intelligent Systems 9 (2023) 7133–7153.
53. Lin and Saripalli [2017]Y. Lin, S. Saripalli,Sampling-based path planning for uav collision avoidance,IEEE Transactions on Intelligent Transportation Systems 18 (2017) 3179–3192.
54. Puente-Castro et al. [2021]A. Puente-Castro, D. Rivero, A. Pazos, E. Fernandez-Blanco,Using reinforcement learning in the path planning of swarms of uavs for the photographic capture of terrains,Engineering Proceedings 7 (2021) 32.
55. Puente-Castro et al. [2022]A. Puente-Castro, D. Rivero, A. Pazos, E. Fernandez-Blanco,A review of artificial intelligence applied to path planning in uav swarms,Neural Computing and Applications 34 (2022) 153–170.
56. Pan et al. [2021]Z. Pan, C. Zhang, Y. Xia, H. Xiong, X. Shao,An improved artificial potential field method for path planning and formation control of the multi-uav systems,IEEE Transactions on Circuits and Systems II: Express Briefs 69 (2021) 1129–1133.
57. Zhao et al. [2021]C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, J. Li,Multi-uav trajectory planning for energy-efficient content coverage: A decentralized learning-based approach,IEEE Journal on Selected Areas in Communications 39 (2021) 3193–3207.
58. Li et al. [2024]K. Li, X. Yan, Y. Han,Multi-mechanism swarm optimization for multi-uav task assignment and path planning in transmission line inspection under multi-wind field,Applied Soft Computing 150 (2024) 111033.
59. Fahlstrom et al. [2022]P. G. Fahlstrom, T. J. Gleason, M. H. Sadraey, Introduction to UAV systems, John Wiley & Sons, 2022.
60. Harvey et al. [2022]C. Harvey, L. L. Gamble, C. R. Bolander, D. F. Hunsaker, J. J. Joo, D. J. Inman,A review of avian-inspired morphing for uav flight control,Progress in Aerospace Sciences 132 (2022) 100825.
61. Mahmoodabadi and Rezaee Babak [2020]M. J. Mahmoodabadi, N. Rezaee Babak,Fuzzy adaptive robust proportional–integral–derivative control optimized by the multi-objective grasshopper optimization algorithm for a nonlinear quadrotor,Journal of Vibration and Control 26 (2020) 1574–1589.
62. Bello et al. [2022]A. B. Bello, F. Navarro, J. Raposo, M. Miranda, A. Zazo, M. Álvarez,Fixed-wing uav flight operation under harsh weather conditions: A case study in livingston island glaciers, antarctica,Drones 6 (2022) 384.
63. Koksal et al. [2020]N. Koksal, H. An, B. Fidan,Backstepping-based adaptive control of a quadrotor uav with guaranteed tracking performance,ISA transactions 105 (2020) 98–110.
64. Zuo et al. [2022]Z. Zuo, C. Liu, Q.-L. Han, J. Song,Unmanned aerial vehicles: Control methods and future challenges,IEEE/CAA Journal of Automatica Sinica 9 (2022) 601–614.
65. Fei et al. [2021]J. Fei, Y. Chen, L. Liu, Y. Fang,Fuzzy multiple hidden layer recurrent neural control of nonlinear system using terminal sliding-mode controller,IEEE transactions on cybernetics 52 (2021) 9519–9534.
66. Gambhire et al. [2021]S. Gambhire, D. R. Kishore, P. Londhe, S. Pawar,Review of sliding mode based control techniques for control system applications,International Journal of dynamics and control 9 (2021) 363–378.
67. Jasim and Veres [2020]O. A. Jasim, S. M. Veres,A robust controller for multi rotor uavs,Aerospace Science and Technology 105 (2020) 106010.
68. Basiri et al. [2022]A. Basiri, V. Mariani, G. Silano, M. Aatif, L. Iannelli, L. Glielmo,A survey on the application of path-planning algorithms for multi-rotor uavs in precision agriculture,The Journal of Navigation 75 (2022) 364–383.
69. Boroujeni et al. [2024]S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, P. Fule, A. Watts, N.-M. T. Kokolakis, K. G. Vamvoudakis,A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management,Information Fusion (2024) 102369.
70. Campion et al. [2018]M. Campion, P. Ranganathan, S. Faruque,Uav swarm communication and control architectures: a review,Journal of Unmanned Vehicle Systems 7 (2018) 93–106.
71. Sharma et al. [2020]A. Sharma, P. Vanjani, N. Paliwal, C. M. W. Basnayaka, D. N. K. Jayakody, H.-C. Wang, P. Muthuchidambaranathan,Communication and networking technologies for uavs: A survey,Journal of Network and Computer Applications 168 (2020) 102739.
72. Hentati and Fourati [2020]A. I. Hentati, L. C. Fourati,Comprehensive survey of uavs communication networks,Computer Standards & Interfaces 72 (2020) 103451.
73. Wu et al. [2021]Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, A. L. Swindlehurst,A comprehensive overview on 5g-and-beyond networks with uavs: From communications to sensing and intelligence,IEEE Journal on Selected Areas in Communications 39 (2021) 2912–2945.
74. Ullah et al. [2020]Z. Ullah, F. Al-Turjman, L. Mostarda,Cognition in uav-aided 5g and beyond communications: A survey,IEEE Transactions on Cognitive Communications and Networking 6 (2020) 872–891.
75. Alladi et al. [2020]T. Alladi, V. Chamola, N. Sahu, M. Guizani,Applications of blockchain in unmanned aerial vehicles: A review,Vehicular Communications 23 (2020) 100249.
76. Kumar et al. [2021]R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, T. R. Gadekallu, G. Srivastava,Sp2f: A secured privacy-preserving framework for smart agricultural unmanned aerial vehicles,Computer Networks 187 (2021) 107819.

77. Messaoudi et al. [2023]K. Messaoudi, O. S. Oubbati, A. Rachedi, A. Lakas, T. Bendouma, N. Chaib,A survey of uav-based data collection: Challenges, solutions and future perspectives,Journal of network and computer applications 216 (2023) 103670.
78. Yoo et al. [2022]M. Yoo, Y. Na, H. Song, G. Kim, J. Yun, S. Kim, C. Moon, K. Jo,Motion estimation and hand gesture recognition-based human–uav interaction approach in real time,Sensors 22 (2022) 2513.
79. Li et al. [2021]T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, Z. Li,Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles,in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 16266–16275.
80. Sun et al. [2022]Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu,Human action recognition from various data modalities: A review,IEEE transactions on pattern analysis and machine intelligence 45 (2022) 3200–3225.
81. Zhang et al. [2020]J. Zhang, Z. Yu, X. Wang, Y. Lyu, S. Mao, S. C. Periaswamy, J. Patton, X. Wang,Rfhui: An rfid based human-unmanned aerial vehicle interaction system in an indoor environment,Digital Communications and Networks 6 (2020) 14–22.
82. Deng et al. [2023]T. Deng, Z. Huo, L. Zhang, Z. Dong, L. Niu, X. Kang, X. Huang,A vr-based bci interactive system for uav swarm control,Biomedical Signal Processing and Control 85 (2023) 104944.
83. Xiao et al. [2024]Z. Xiao, P. Li, C. Liu, H. Gao, X. Wang,Macns: A generic graph neural network integrated deep reinforcement learning based multi-agent collaborative navigation system for dynamic trajectory planning,Information Fusion 105 (2024) 102250.
84. Jiao et al. [2020]R. Jiao, Z. Wang, R. Chu, M. Dong, Y. Rong, W. Chou,An intuitive end-to-end human-uav interaction system for field exploration,Frontiers in Neurorobotics 13 (2020) 117.
85. Divband Soorati et al. [2021]M. Divband Soorati, J. Clark, J. Ghofrani, D. Tarapore, S. D. Ramchurn,Designing a user-centered interaction interface for human–swarm teaming,Drones 5 (2021) 131.
86. Zheng et al. [2020]Y.-J. Zheng, Y.-C. Du, Z.-L. Su, H.-F. Ling, M.-X. Zhang, S.-Y. Chen,Evolutionary human-uav cooperation for transmission network restoration,IEEE Transactions on Industrial Informatics 17 (2020) 1648–1657.
87. Lim et al. [2021]Y. Lim, N. Pongsakornsathien, A. Gardi, R. Sabatini, T. Kistan, N. Ezer, D. J. Bursch,Adaptive human-robot interactions for multiple unmanned aerial vehicles,Robotics 10 (2021) 12.
88. Chang et al. [2020]W. Chang, W. Lizhen, Y. Chao, W. Zhichao, L. Han, Y. Chao,Coactive design of explainable agent-based task planning and deep reinforcement learning for human-uavs teamwork,Chinese Journal of Aeronautics 33 (2020) 2930–2945.
89. Cauchard et al. [2021]J. R. Cauchard, M. Khamis, J. Garcia, M. Kljun, A. M. Brock,Toward a roadmap for human-drone interaction,Interactions 28 (2021) 76–81.
90. Ribeiro et al. [2021]R. Ribeiro, J. Ramos, D. Safadinho, A. Reis, C. Rabadão, J. Barroso, A. Pereira,Web ar solution for uav pilot training and usability testing,Sensors 21 (2021) 1456.
91. Mohiuddin et al. [2023]A. Mohiuddin, T. Taha, Y. Zweiri, D. Gan,Dual-uav payload transportation using optimized velocity profiles via real-time dynamic programming,Drones 7 (2023) 171.
92. González-Jorge et al. [2017]H. González-Jorge, J. Martínez-Sánchez, M. Bueno, P. Arias,Unmanned aerial systems for civil applications: A review,Drones 1 (2017) 2.
93. Hadi et al. [2014]G. S. Hadi, R. Varianto, B. R. Trilaksono, A. Budiyono,Autonomous uav system development for payload dropping mission,Journal of Instrumentation, Automation and Systems 1 (2014) 72–77.
94. Kusznir and Smoczek [2020]T. Kusznir, J. Smoczek,Sliding mode-based control of a uav quadrotor for suppressing the cable-suspended payload vibration,Journal of Control Science and Engineering 2020 (2020) 5058039.
95. Lee and Son [2020]S. Lee, H. Son,Antisway control of a multirotor with cable-suspended payload,IEEE Transactions on Control Systems Technology 29 (2020) 2630–2638.
96. Mohammadi et al. [2020]K. Mohammadi, S. Sirouspour, A. Grivani,Control of multiple quad-copters with a cable-suspended payload subject to disturbances,IEEE/ASME Transactions on Mechatronics 25 (2020) 1709–1718.
97. Lee et al. [2021]C. Lee, S. Kim, B. Chu,A survey: Flight mechanism and mechanical structure of the uav,International Journal of Precision Engineering and Manufacturing 22 (2021) 719–743.
98. Zhou et al. [2020]Y. Zhou, B. Rao, W. Wang,Uav swarm intelligence: Recent advances and future trends,Ieee Access 8 (2020) 183856–183878.
99. Chakraborty and Kar [2017]A. Chakraborty, A. K. Kar,Swarm intelligence: A review of algorithms,Nature-inspired computing and optimization: Theory and applications (2017) 475–494.
100. Lamport [2001]L. Lamport,Paxos made simple,ACM SIGACT News (Distributed Computing Column) 32, 4 (Whole Number 121, December 2001) (2001) 51–58.
101. Kennedy and Eberhart [1995]J. Kennedy, R. Eberhart,Particle swarm optimization,in: Proceedings of ICNN'95-international conference on neural networks, volume 4, ieee, 1995, pp. 1942–1948.
102. Jones and Matarić [2018]C. Jones, M. J. Matarić,Behavior-based coordination in multi-robot systems,in: Autonomous Mobile Robots, CRC Press, 2018, pp. 549–570.
103. Ma et al. [2022]L. Ma, B. Lin, W. Zhang, J. Tao, X. Zhu, H. Chen,A survey of research on the distributed cooperation method of the uav swarm based on swarm intelligence,in: 2022 IEEE 13th International Conference on Software Engineering and Service Science (ICSESS), IEEE, 2022, pp. 305–309.
104. Schwarzrock et al. [2018]J. Schwarzrock, I. Zacarias, A. L. Bazzan, R. Q. de Araujo Fernandes, L. H. Moreira, E. P. de Freitas,Solving task allocation problem in multi unmanned aerial vehicles systems using swarm intelligence,Engineering Applications of Artificial Intelligence 72 (2018) 10–20.
105. Zhang and Chen [2021]X. Zhang, X. Chen,Uav task allocation based on clone selection algorithm,Wireless Communications and Mobile Computing 2021 (2021) 5518927.
106. Kudo and Cai [2023]F. Kudo, K. Cai,A tsp-based online algorithm for multi-task multi-agent pickup and delivery,IEEE Robotics and Automation Letters (2023).
107. Sarkar et al. [2018]C. Sarkar, H. S. Paul, A. Pal,A scalable multi-robot task allocation algorithm,in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 5022–5027.
108. Darrah et al. [2005]M. Darrah, W. Niland, B. Stolarik,Multiple uav dynamic task allocation using mixed integer linear programming in a sead mission,in: Infotech@ Aerospace, 2005, p. 7164.
109. Ye et al. [2020]F. Ye, J. Chen, Y. Tian, T. Jiang,Cooperative multiple task assignment of heterogeneous uavs using a modified genetic algorithm with multi-type-gene chromosome encoding strategy,Journal of intelligent & robotic systems 100 (2020) 615–627.

110. Han et al. [2021]S. Han, C. Fan, X. Li, X. Luo, Z. Liu,A modified genetic algorithm for task assignment of heterogeneous unmanned aerial vehicle system,Measurement and Control 54 (2021) 994–1014.
111. Yan et al. [2024]F. Yan, J. Chu, J. Hu, X. Zhu,Cooperative task allocation with simultaneous arrival and resource constraint for multi-uav using a genetic algorithm,Expert Systems with Applications 245 (2024) 123023.
112. Jiang et al. [2017]X. Jiang, Q. Zhou, Y. Ye,Method of task assignment for uav based on particle swarm optimization in logistics,in: Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence, 2017, pp. 113–117.
113. Gao et al. [2018]Y. Gao, Y. Zhang, S. Zhu, Y. Sun,Multi-uav task allocation based on improved algorithm of multi-objective particle swarm optimization,in: 2018 International conference on cyber-enabled distributed computing and knowledge discovery (CyberC), IEEE, 2018, pp. 443–4437.
114. Choi et al. [2010]H.-J. Choi, J.-B. Seo, Y.-D. Kim,Task assignment of multiple uavs using milp and ga,Journal of the Korean Society for Aeronautical & Space Sciences 38 (2010) 427–436.
115. Yang et al. [2019]J. Yang, X. You, G. Wu, M. M. Hassan, A. Almogren, J. Guna,Application of reinforcement learning in uav cluster task scheduling,Future generation computer systems 95 (2019) 140–148.
116. Yin et al. [2022]Y. Yin, Y. Guo, Q. Su, Z. Wang,Task allocation of multiple unmanned aerial vehicles based on deep transfer reinforcement learning,Drones 6 (2022) 215.
117. Peng et al. [2021]Q. Peng, H. Wu, R. Xue,Review of dynamic task allocation methods for uav swarms oriented to ground targets,Complex System Modeling and Simulation 1 (2021) 163–175.
118. Skaltsis et al. [2023]G. M. Skaltsis, H.-S. Shin, A. Tsourdos,A review of task allocation methods for uavs,Journal of Intelligent & Robotic Systems 109 (2023) 76.
119. Cheng et al. [2016]Q. Cheng, D. Yin, J. Yang, L. Shen,An auction-based multiple constraints task allocation algorithm for multi-uav system,in: 2016 International Conference on Cybernetics, Robotics and Control (CRC), IEEE, 2016, pp. 1–5.
120. Duan et al. [2019]X. Duan, H. Liu, H. Tang, Q. Cai, F. Zhang, X. Han,A novel hybrid auction algorithm for multi-uavs dynamic task assignment,IEEE access 8 (2019) 86207–86222.
121. Zhang et al. [2022]Z. Zhang, H. Liu, G. Wu,A dynamic task scheduling method for multiple uavs based on contract net protocol,Sensors 22 (2022) 4486.
122. Wang et al. [2023]G. Wang, X. Lv, X. Yan,A two-stage distributed task assignment algorithm based on contract net protocol for multi-uav cooperative reconnaissance task reassignment in dynamic environments,Sensors 23 (2023) 7980.
123. Campion et al. [2018]M. Campion, P. Ranganathan, S. Faruque,A review and future directions of uav swarm communication architectures,in: 2018 IEEE international conference on electro/information technology (EIT), IEEE, 2018, pp. 0903–0908.
124. Bekmezci et al. [2013]I. Bekmezci, O. K. Sahingoz, Ş. Temel,Flying ad-hoc networks (fanets): A survey,Ad Hoc Networks 11 (2013) 1254–1270.
125. Javed et al. [2024]S. Javed, A. Hassan, R. Ahmad, W. Ahmed, R. Ahmed, A. Saadat, M. Guizani,State-of-the-art and future research challenges in uav swarms,IEEE Internet of Things Journal (2024).
126. Turker et al. [2016]T. Turker, G. Yilmaz, O. K. Sahingoz,Gpu-accelerated flight route planning for multi-uav systems using simulated annealing,in: Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7-10, 2016, Proceedings 17, Springer, 2016, pp. 279–288.
127. Wei and Wei [2009]L. Wei, Z. Wei,Path planning of uavs swarm using ant colony system,in: 2009 Fifth International Conference on Natural Computation, volume 5, IEEE, 2009, pp. 288–292.
128. Ragi and Mittelmann [2017]S. Ragi, H. D. Mittelmann,Mixed-integer nonlinear programming formulation of a uav path optimization problem,in: 2017 American Control Conference (ACC), IEEE, 2017, pp. 406–411.
129. Kool et al. [2018]W. Kool, H. Van Hoof, M. Welling,Attention, learn to solve routing problems!,arXiv preprint arXiv:1803.08475 (2018).
130. Xia and Yudi [2018]C. Xia, A. Yudi,Multi—uav path planning based on improved neural network,in: 2018 Chinese Control And Decision Conference (CCDC), IEEE, 2018, pp. 354–359.
131. Sanna et al. [2021]G. Sanna, S. Godio, G. Guglieri,Neural network based algorithm for multi-uav coverage path planning,in: 2021 International Conference on Unmanned Aircraft Systems (ICUAS), IEEE, 2021, pp. 1210–1217.
132. Ouyang et al. [2023]Q. Ouyang, Z. Wu, Y. Cong, Z. Wang,Formation control of unmanned aerial vehicle swarms: A comprehensive review,Asian Journal of Control 25 (2023) 570–593.
133. Bu et al. [2024]Y. Bu, Y. Yan, Y. Yang,Advancement challenges in uav swarm formation control: A comprehensive review,Drones 8 (2024) 320.
134. Askari et al. [2015]A. Askari, M. Mortazavi, H. Talebi,Uav formation control via the virtual structure approach,Journal of Aerospace Engineering 28 (2015) 04014047.
135. Lewis and Tan [1997]M. A. Lewis, K.-H. Tan,High precision formation control of mobile robots using virtual structures,Autonomous robots 4 (1997) 387–403.
136. Desai et al. [1998]J. P. Desai, J. Ostrowski, V. Kumar,Controlling formations of multiple mobile robots,in: Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146), volume 4, IEEE, 1998, pp. 2864–2869.
137. Huang et al. [2022]H. Huang, A. V. Savkin, W. Ni,Decentralized navigation of a uav team for collaborative covert eavesdropping on a group of mobile ground nodes,IEEE Transactions on Automation Science and Engineering 19 (2022) 3932–3941.
138. Sun et al. [2022]S. Sun, Y. Liu, S. Guo, G. Li, X. Yuan,Observation-driven multiple uav coordinated standoff target tracking based on model predictive control,Tsinghua Science and Technology 27 (2022) 948–963.
139. Duan et al. [2021]H. Duan, L. Xin, Y. Shi,Homing pigeon-inspired autonomous navigation system for unmanned aerial vehicles,IEEE Transactions On Aerospace and Electronic Systems 57 (2021) 2218–2224.
140. Tao et al. [2023]C. Tao, R. Zhang, Z. Song, B. Wang, Y. Jin,Multi-uav formation control in complex conditions based on improved consistency algorithm,Drones 7 (2023) 185.
141. Brown [2020]T. B. Brown,Language models are few-shot learners,arXiv preprint arXiv:2005.14165 (2020).
142. Ouyang et al. [2022]L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al.,Training language models to follow instructions with human feedback,Advances in neural information processing systems 35 (2022) 27730–27744.
143. Achiam et al. [2023]J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al.,Gpt-4 technical report,arXiv preprint arXiv:2303.08774 (2023).
144. Anthropic [2023]Anthropic, Model card and evaluations for claude models, 2023. URL: https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf.

145. Anthropic [2022]Anthropic,Constitutional ai: Harmlessness from ai feedback (2022). URL: https://en.wikipedia.org/wiki/Claude_%28language_model%29.
146. Anthropic [2023]Anthropic, Mapping the mind of a large language model, 2023. URL: https://www.anthropic.com/research/mapping-mind-language-model.
147. Jiang et al. [2023]A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al.,Mistral 7b,arXiv preprint arXiv:2310.06825 (2023).
148. Jiang et al. [2024]A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al.,Mixtral of experts,arXiv preprint arXiv:2401.04088 (2024).
149. Chowdhery et al. [2023]A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al.,Palm: Scaling language modeling with pathways,Journal of Machine Learning Research 24 (2023) 1–113.
150. Driess et al. [2023]D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al.,Palm-e: An embodied multimodal language model,arXiv preprint arXiv:2303.03378 (2023).
151. Team et al. [2023]G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al.,Gemini: a family of highly capable multimodal models,arXiv preprint arXiv:2312.11805 (2023).
152. Reid et al. [2024]M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al.,Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,arXiv preprint arXiv:2403.05530 (2024).
153. Touvron et al. [2023a]H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al.,Llama: Open and efficient foundation language models,arXiv preprint arXiv:2302.13971 (2023a).
154. Touvron et al. [2023b]H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al.,Llama 2: Open foundation and fine-tuned chat models,arXiv preprint arXiv:2307.09288 (2023b).
155. Dubey et al. [2024]A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al.,The llama 3 herd of models,arXiv preprint arXiv:2407.21783 (2024).
156. Chiang et al. [2023]W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al.,Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,See https://vicuna. lmsys. org (accessed 14 April 2023) 2 (2023) 6.
157. Bai et al. [2023]J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al.,Qwen technical report,arXiv preprint arXiv:2309.16609 (2023).
158. Yang et al. [2024]A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, et al.,Qwen2 technical report,arXiv preprint arXiv:2407.10671 (2024).
159. Cai et al. [2024]Z. Cai, M. Cao, H. Chen, K. Chen, K. Chen, X. Chen, X. Chen, Z. Chen, Z. Chen, P. Chu, et al.,Internlm2 technical report,arXiv preprint arXiv:2403.17297 (2024).
160. Zhao et al. [2023]Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, B. Kang,Bubogpt: Enabling visual grounding in multi-modal llms,arXiv preprint arXiv:2307.08581 (2023).
161. Du et al. [2021]Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, J. Tang,Glm: General language model pretraining with autoregressive blank infilling,arXiv preprint arXiv:2103.10360 (2021).
162. Zeng et al. [2022]A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al.,Glm-130b: An open bilingual pre-trained model,arXiv preprint arXiv:2210.02414 (2022).
163. GLM et al. [2024]T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, et al.,Chatglm: A family of large language models from glm-130b to glm-4 all tools,arXiv preprint arXiv:2406.12793 (2024).
164. Bi et al. [2024]X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al.,Deepseek llm: Scaling open-source language models with longtermism,arXiv preprint arXiv:2401.02954 (2024).
165. Liu et al. [2024]A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, et al.,Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model,arXiv preprint arXiv:2405.04434 (2024).
166. Guo et al. [2024]D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, et al.,Deepseek-coder: When the large language model meets programming–the rise of code intelligence,arXiv preprint arXiv:2401.14196 (2024).
167. OpenAI [2024]OpenAI, GPT-4V System Card, 2024. URL: https://openai.com/index/gpt-4v-system-card/, accessed: 2024-11-16.
168. Anthropic [2024]Anthropic, The claude 3 model family: Opus, sonnet, haiku, 2024. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, accessed on November 16, 2024.
169. Liu et al. [2024a]H. Liu, C. Li, Q. Wu, Y. J. Lee,Visual instruction tuning,Advances in neural information processing systems 36 (2024a).
170. Liu et al. [2024b]H. Liu, C. Li, Y. Li, Y. J. Lee,Improved baselines with visual instruction tuning,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024b, pp. 26296–26306.
171. Liu et al. [2024c]H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, Llava-next: Improved reasoning, ocr, and world knowledge, 2024c. URL: https://llava-vl.github.io/blog/2024-01-30-llava-next/, accessed: 2024-11-16.
172. Lin et al. [2024]B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Zhang, M. Ning, L. Yuan,Moe-llava: Mixture of experts for large vision-language models,arXiv preprint arXiv:2401.15947 (2024).
173. Xu et al. [2024]G. Xu, P. Jin, L. Hao, Y. Song, L. Sun, L. Yuan,Llava-o1: Let vision language models reason step-by-step,arXiv preprint arXiv:2411.10440 (2024).
174. Alayrac et al. [2022]J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al.,Flamingo: a visual language model for few-shot learning,Advances in neural information processing systems 35 (2022) 23716–23736.
175. Li et al. [2022]J. Li, D. Li, C. Xiong, S. Hoi,Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.
176. Li et al. [2023]J. Li, D. Li, S. Savarese, S. Hoi,Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.
177. Dai et al. [2023]W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL: https://arxiv.org/abs/2305.06500. arXiv:2305.06500.
178. Li et al. [2025]Y. Li, C. Wang, J. Jia,Llama-vid: An image is worth 2 tokens in large language models,in: European Conference on Computer Vision, Springer, 2025, pp. 323–340.
179. Kim et al. [2024]W. Kim, C. Choi, W. Lee, W. Rhee,An image grid can be worth a video: Zero-shot video question answering using a vlm,arXiv preprint arXiv:2403.18406 (2024).

180. Maaz et al. [2023]M. Maaz, H. Rasheed, S. Khan, F. S. Khan,Video-chatgpt: Towards detailed video understanding via large vision and language models,arXiv preprint arXiv:2306.05424 (2023).
181. Wang et al. [2024]Z. Wang, S. Yu, E. Stengel-Eskin, J. Yoon, F. Cheng, G. Bertasius, M. Bansal,Videotree: Adaptive tree-based video representation for llm reasoning on long videos,arXiv preprint arXiv:2405.19209 (2024).
182. Zeng et al. [2021]Y. Zeng, X. Zhang, H. Li,Multi-grained vision language pre-training: Aligning texts with visual concepts,arXiv preprint arXiv:2111.08276 (2021).
183. Lu et al. [2024]P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, J. Gao,Chameleon: Plug-and-play compositional reasoning with large language models,Advances in Neural Information Processing Systems 36 (2024).
184. Ke et al. [2024]F. Ke, Z. Cai, S. Jahangard, W. Wang, P. D. Haghighi, H. Rezatofighi,Hydra: A hyper agent for dynamic compositional visual reasoning,arXiv preprint arXiv:2403.12884 (2024).
185. Gupta and Kembhavi [2023]T. Gupta, A. Kembhavi,Visual programming: Compositional visual reasoning without training,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14953–14962.
186. Radford et al. [2021]A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.,Learning transferable visual models from natural language supervision,in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
187. Yao et al. [2021]L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, C. Xu,Filip: Fine-grained interactive language-image pre-training,arXiv preprint arXiv:2111.07783 (2021).
188. Zhong et al. [2022]Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, et al.,Regionclip: Region-based language-image pretraining,in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16793–16803.
189. Sun et al. [2023]Q. Sun, Y. Fang, L. Wu, X. Wang, Y. Cao,Eva-clip: Improved training techniques for clip at scale,arXiv preprint arXiv:2303.15389 (2023).
190. Li et al. [2022]L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al.,Grounded language-image pre-training,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.
191. Zhang et al. [2022]H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum,Dino: Detr with improved denoising anchor boxes for end-to-end object detection,arXiv preprint arXiv:2203.03605 (2022).
192. Liu et al. [2023]S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al.,Grounding dino: Marrying dino with grounded pre-training for open-set object detection,arXiv preprint arXiv:2303.05499 (2023).
193. Oquab et al. [2023]M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al.,Dinov2: Learning robust visual features without supervision,arXiv preprint arXiv:2304.07193 (2023).
194. Ranzinger et al. [2024]M. Ranzinger, G. Heinrich, J. Kautz, P. Molchanov,Am-radio: Agglomerative vision foundation model reduce all domains into one,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12490–12500.
195. Zhou et al. [2024]G. Zhou, H. Pan, Y. LeCun, L. Pinto,Dino-wm: World models on pre-trained visual features enable zero-shot planning,arXiv preprint arXiv:2411.04983 (2024).
196. Cheng et al. [2024]T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, Y. Shan,Yolo-world: Real-time open-vocabulary object detection,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16901–16911.
197. Lüddecke and Ecker [2022]T. Lüddecke, A. Ecker,Image segmentation using text and image prompts,in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 7086–7096.
198. Kirillov et al. [2023]A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al.,Segment anything,in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
199. Xu et al. [2024]X. Xu, H. Chen, L. Zhao, Z. Wang, J. Zhou, J. Lu,Embodiedsam: Online segment any 3d thing in real time,arXiv preprint arXiv:2408.11811 (2024).
200. Zhou et al. [2024]Y. Zhou, J. Gu, T. Y. Chiang, F. Xiang, H. Su,Point-sam: Promptable 3d segmentation model for point clouds,arXiv preprint arXiv:2406.17741 (2024).
201. Yuan et al. [2025]H. Yuan, X. Li, C. Zhou, Y. Li, K. Chen, C. C. Loy,Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively,in: European Conference on Computer Vision, Springer, 2025, pp. 419–437.
202. Pan et al. [2025]T. Pan, L. Tang, X. Wang, S. Shan,Tokenize anything via prompting,in: European Conference on Computer Vision, Springer, 2025, pp. 330–348.
203. Xiong et al. [2024]Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, et al.,Efficientsam: Leveraged masked image pretraining for efficient segment anything,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16111–16121.
204. Zhang et al. [2023]C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, C. S. Hong,Faster segment anything: Towards lightweight sam for mobile applications,arXiv preprint arXiv:2306.14289 (2023).
205. Ravi et al. [2024]N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al.,Sam 2: Segment anything in images and videos,arXiv preprint arXiv:2408.00714 (2024).
206. Yang et al. [2024]C.-Y. Yang, H.-W. Huang, W. Chai, Z. Jiang, J.-N. Hwang,Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory,arXiv preprint arXiv:2411.11922 (2024).
207. Wang et al. [2023]X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, T. Huang,Seggpt: Segmenting everything in context,arXiv preprint arXiv:2304.03284 (2023).
208. Yuan et al. [2024]Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, J. Zhu,Osprey: Pixel understanding with visual instruction tuning,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 28202–28211.
209. Zou et al. [2024]X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. J. Lee,Segment everything everywhere all at once,Advances in Neural Information Processing Systems 36 (2024).
210. Liu et al. [2024]Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, Z. Liu,Segment any point cloud sequences by distilling vision foundation models,Advances in Neural Information Processing Systems 36 (2024).
211. Lai et al. [2024]X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, J. Jia,Lisa: Reasoning segmentation via large language model,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9579–9589.
212. Bhat et al. [2023]S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, M. Müller,Zoedepth: Zero-shot transfer by combining relative and metric depth,arXiv preprint arXiv:2302.12288 (2023).
213. Zhu et al. [2024]R. Zhu, C. Wang, Z. Song, L. Liu, T. Zhang, Y. Zhang,Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation,arXiv preprint arXiv:2407.08187 (2024).

214. Yang et al. [2024a]L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, H. Zhao,Depth anything: Unleashing the power of large-scale unlabeled data,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024a, pp. 10371–10381.
215. Yang et al. [2024b]L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao,Depth anything v2,arXiv preprint arXiv:2406.09414 (2024b).
216. Bochkovskii et al. [2024]A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, V. Koltun,Depth pro: Sharp monocular metric depth in less than a second,arXiv preprint arXiv:2410.02073 (2024).
217. Vaswani [2017]A. Vaswani,Attention is all you need,Advances in Neural Information Processing Systems (2017).
218. Minaee et al. [2024]S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao,Large language models: A survey,arXiv preprint arXiv:2402.06196 (2024).
219. Zhao et al. [2023]W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al.,A survey of large language models,arXiv preprint arXiv:2303.18223 (2023).
220. Chang et al. [2024]Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al.,A survey on evaluation of large language models,ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.
221. Naveed et al. [2023]H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian,A comprehensive overview of large language models,arXiv preprint arXiv:2307.06435 (2023).
222. Li et al. [2023]Y. Li, B. Hui, X. Xia, J. Yang, M. Yang, L. Zhang, S. Si, J. Liu, T. Liu, F. Huang, et al.,One shot learning as instruction data prospector for large language models,arXiv preprint arXiv:2312.10302 (2023).
223. Radford et al. [2019]A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al.,Language models are unsupervised multitask learners,OpenAI blog 1 (2019) 9.
224. Liu et al. [2021]J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen,What makes good in-context examples for gpt-3?,arXiv preprint arXiv:2101.06804 (2021).
225. Dong et al. [2022]Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, et al.,A survey on in-context learning,arXiv preprint arXiv:2301.00234 (2022).
226. Kojima et al. [2022]T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa,Large language models are zero-shot reasoners,Advances in neural information processing systems 35 (2022) 22199–22213.
227. Zhang et al. [2022]Z. Zhang, A. Zhang, M. Li, A. Smola,Automatic chain of thought prompting in large language models,arXiv preprint arXiv:2210.03493 (2022).
228. Wei et al. [2022]J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al.,Chain-of-thought prompting elicits reasoning in large language models,Advances in neural information processing systems 35 (2022) 24824–24837.
229. Feng et al. [2024]G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, L. Wang,Towards revealing the mystery behind chain of thought: a theoretical perspective,Advances in Neural Information Processing Systems 36 (2024).
230. Shen et al. [2024]Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang,Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face,Advances in Neural Information Processing Systems 36 (2024).
231. Khot et al. [2022]T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, A. Sabharwal,Decomposed prompting: A modular approach for solving complex tasks,arXiv preprint arXiv:2210.02406 (2022).
232. Huang et al. [2024]X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang, E. Chen,Understanding the planning of llm agents: A survey,arXiv preprint arXiv:2402.02716 (2024).
233. White et al. [2024]C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Naidu, et al.,Livebench: A challenging, contamination-free llm benchmark,arXiv preprint arXiv:2406.19314 (2024).
234. Ma et al. [2024]X. Ma, Y. Bhalgat, B. Smart, S. Chen, X. Li, J. Ding, J. Gu, D. Z. Chen, S. Peng, J.-W. Bian, et al.,When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models,arXiv preprint arXiv:2405.10255 (2024).
235. Du et al. [2022]Y. Du, Z. Liu, J. Li, W. X. Zhao,A survey of vision-language pre-trained models,arXiv preprint arXiv:2202.10936 (2022).
236. Long et al. [2022]S. Long, F. Cao, S. C. Han, H. Yang,Vision-and-language pretrained models: A survey,arXiv preprint arXiv:2204.07356 (2022).
237. Zhou et al. [2022]K. Zhou, J. Yang, C. C. Loy, Z. Liu,Learning to prompt for vision-language models,International Journal of Computer Vision 130 (2022) 2337–2348.
238. Yin et al. [2024]S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen,A survey on multimodal large language models,National Science Review (2024) nwae403.
239. Zhang et al. [2024]J. Zhang, J. Huang, S. Jin, S. Lu,Vision-language models for vision tasks: A survey,IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
240. Yang et al. [2023]Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, L. Wang,The dawn of lmms: Preliminary explorations with gpt-4v (ision),arXiv preprint arXiv:2309.17421 9 (2023) 1.
241. Islam and Moushi [2024]R. Islam, O. M. Moushi,Gpt-4o: The cutting-edge advancement in multimodal llm,Authorea Preprints (2024).
242. Latif et al. [2024]E. Latif, Y. Zhou, S. Guo, Y. Gao, L. Shi, M. Nayaaba, G. Lee, L. Zhang, A. Bewersdorff, L. Fang, et al.,A systematic assessment of openai o1-preview for higher order thinking in education,arXiv preprint arXiv:2410.21287 (2024).
243. Chiang et al. [2023]W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.
244. Rizzoli et al. [2023]G. Rizzoli, F. Barbato, M. Caligiuri, P. Zanuttigh,Syndrone-multi-modal uav dataset for urban scenarios,in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2210–2220.
245. Carion et al. [2020]N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko,End-to-end object detection with transformers,in: European conference on computer vision, Springer, 2020, pp. 213–229.
246. Mou et al. [2020]L. Mou, Y. Hua, P. Jin, X. X. Zhu,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 (2020) 125–133.
247. Bashmal et al. [2023]L. Bashmal, Y. Bazi, M. M. Al Rahhal, M. Zuair, F. Melgani,Capera: Captioning events in aerial videos,Remote Sensing 15 (2023) 2139.
248. Jaisawal et al. [2024]P. K. Jaisawal, S. Papakonstantinou, V. Gollnick,Airfisheye dataset: A multi-model fisheye dataset for uav applications,in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 11818–11824.
249. Florea et al. [2021]H. Florea, V.-C. Miclea, S. Nedevschi,Wilduav: Monocular uav dataset for depth estimation tasks,in: 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2021, pp. 291–298.

250. Oh et al. [2011]S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al.,A large-scale benchmark dataset for event recognition in surveillance video,in: CVPR 2011, IEEE, 2011, pp. 3153–3160.
251. Zhang et al. [2022]C. Zhang, G. Huang, L. Liu, S. Huang, Y. Yang, X. Wan, S. Ge, D. Tao,Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking,IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 9186–9205.
252. Li et al. [2022]B. Li, C. Fu, F. Ding, J. Ye, F. Lin,All-day object tracking for unmanned aerial vehicle,IEEE Transactions on Mobile Computing 22 (2022) 4515–4529.
253. Zhang et al. [2022]P. Zhang, J. Zhao, D. Wang, H. Lu, X. Ruan,Visible-thermal uav tracking: A large-scale benchmark and new baseline,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8886–8895.
254. Wang et al. [2021]X. Wang, X. Shu, Z. Zhang, B. Jiang, Y. Wang, Y. Tian, F. Wu,Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark,in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13763–13773.
255. Zhang et al. [2020]S. Zhang, Q. Zhang, Y. Yang, X. Wei, P. Wang, B. Jiao, Y. Zhang,Person re-identification in aerial imagery,IEEE Transactions on Multimedia 23 (2020) 281–291.
256. Kristan et al. [2020]M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, J.-K. Kämäräinen, M. Danelljan, L. Č. Zajc, A. Lukežič, O. Drbohlav, et al.,The eighth visual object tracking vot2020 challenge results,in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 547–601.
257. Huang et al. [2019]L. Huang, X. Zhao, K. Huang,Got-10k: A large high-diversity benchmark for generic object tracking in the wild,IEEE transactions on pattern analysis and machine intelligence 43 (2019) 1562–1577.
258. Li and Yeung [2017]S. Li, D.-Y. Yeung,Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models,in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
259. Robicquet et al. [2016]A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese,Learning social etiquette: Human trajectory understanding in crowded scenes,in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 549–565.
260. Mundhenk et al. [2016]T. N. Mundhenk, G. Konjevod, W. A. Sakla, K. Boakye,A large contextual dataset for classification, detection and counting of cars with deep learning,in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 785–800.
261. Kapoor et al. [2023]S. Kapoor, A. Sharma, A. Verma, S. Singh,Aeriform in-action: A novel dataset for human action recognition in aerial videos,Pattern Recognition 140 (2023) 109505.
262. Corona et al. [2021]K. Corona, K. Osterdahl, R. Collins, A. Hoogs,Meva: A large-scale multiview, multimodal video dataset for activity detection,in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1060–1068.
263. Perera et al. [2020]A. G. Perera, Y. W. Law, T. T. Ogunwa, J. Chahl,A multiviewpoint outdoor dataset for human action recognition,IEEE Transactions on Human-Machine Systems 50 (2020) 405–413.
264. Choi et al. [2020]J. Choi, G. Sharma, M. Chandraker, J.-B. Huang,Unsupervised and semi-supervised domain adaptation for action recognition from drones,in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1717–1726.
265. Perera et al. [2019]A. G. Perera, Y. W. Law, J. Chahl,Drone-action: An outdoor recorded drone video dataset for action recognition,Drones 3 (2019) 82.
266. Perera et al. [2018]A. G. Perera, Y. Wei Law, J. Chahl,Uav-gesture: A dataset for uav control and gesture recognition,in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
267. Lee et al. [2024]J. Lee, T. Miyanishi, S. Kurita, K. Sakamoto, D. Azuma, Y. Matsuo, N. Inoue,Citynav: Language-goal aerial navigation dataset with geographic information,arXiv preprint arXiv:2406.14240 (2024).
268. Liu et al. [2023]S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, Q. Wu,Aerialvln: Vision-and-language navigation for uavs,in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15384–15394.
269. Zhu et al. [2021]S. Zhu, T. Yang, C. Chen,Vigor: Cross-view image geo-localization beyond one-to-one retrieval,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3640–3649.
270. Zheng et al. [2020]Z. Zheng, Y. Wei, Y. Yang,University-1652: A multi-view multi-source benchmark for drone-based geo-localization,in: Proceedings of the 28th ACM international conference on Multimedia, 2020, pp. 1395–1403.
271. Yao et al. [2024]Y. Yao, S. Luo, H. Zhao, G. Deng, L. Song,Can llm substitute human labeling? a case study of fine-grained chinese address entity recognition dataset for uav delivery,in: Companion Proceedings of the ACM on Web Conference 2024, 2024, pp. 1099–1102.
272. Dai et al. [2023]M. Dai, E. Zheng, Z. Feng, L. Qi, J. Zhuang, W. Yang,Vision-based uav self-positioning in low-altitude urban environments,IEEE Transactions on Image Processing (2023).
273. Schumann and Riezler [2022]R. Schumann, S. Riezler,Analyzing generalization of vision and language navigation to unseen outdoor areas,arXiv preprint arXiv:2203.13838 (2022).
274. Zhang et al. [2025]G. Zhang, Y. Liu, X. Yang, H. Huang, C. Huang,Trafficnight: An aerial multimodal benchmark for nighttime vehicle surveillance,in: European Conference on Computer Vision, Springer, 2025, pp. 36–48.
275. Zhu et al. [2021]P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling,Detection and tracking meet drones challenge,IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 7380–7399.
276. Shah et al. [2018]A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, A. Hauptmann,Cadp: A novel dataset for cctv traffic camera based accident analysis,in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2018, pp. 1–9.
277. Hsieh et al. [2017]M.-R. Hsieh, Y.-L. Lin, W. H. Hsu,Drone-based object counting by spatially regularized regional proposal network,in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4145–4153.
278. Waqas Zamir et al. [2019]S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, X. Bai,isaid: A large-scale dataset for instance segmentation in aerial images,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 28–37.
279. Yang et al. [2018]M. Y. Yang, W. Liao, X. Li, B. Rosenhahn,Deep learning for vehicle detection in aerial images,in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3079–3083.
280. Lyu et al. [2020]Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, M. Y. Yang,Uavid: A semantic segmentation dataset for uav imagery,ISPRS journal of photogrammetry and remote sensing 165 (2020) 108–119.

281. Bozcan and Kayacan [2020]I. Bozcan, E. Kayacan,Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance,in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 8504–8510.
282. Krajewski et al. [2018]R. Krajewski, J. Bock, L. Kloeker, L. Eckstein,The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems,in: 2018 21st international conference on intelligent transportation systems (ITSC), IEEE, 2018, pp. 2118–2125.
283. Du et al. [2018]D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian,The unmanned aerial vehicle benchmark: Object detection and tracking,in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 370–386.
284. Razakarivony and Jurie [2016]S. Razakarivony, F. Jurie,Vehicle detection in aerial imagery: A small target detection benchmark,Journal of Visual Communication and Image Representation 34 (2016) 187–203.
285. Lam et al. [2018]D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, B. McCord,xview: Objects in context in overhead imagery,arXiv preprint arXiv:1802.07856 (2018).
286. Xia et al. [2018]G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang,Dota: A large-scale dataset for object detection in aerial images,in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3974–3983.
287. Lu et al. [2017]X. Lu, B. Wang, X. Zheng, X. Li,Exploring models and data for remote sensing image caption generation,IEEE Transactions on Geoscience and Remote Sensing 56 (2017) 2183–2195.
288. Liu et al. [2024]F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, J. Zhou,Remoteclip: A vision language foundation model for remote sensing,IEEE Transactions on Geoscience and Remote Sensing (2024).
289. Li et al. [2020]K. Li, G. Wan, G. Cheng, L. Meng, J. Han,Object detection in optical remote sensing images: A survey and a new benchmark,ISPRS journal of photogrammetry and remote sensing 159 (2020) 296–307.
290. Zhang et al. [2019]Y. Zhang, Y. Yuan, Y. Feng, X. Lu,Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection,IEEE Transactions on Geoscience and Remote Sensing 57 (2019) 5535–5548.
291. Liu et al. [2017]Z. Liu, L. Yuan, L. Weng, Y. Yang,A high resolution optical satellite image dataset for ship recognition and some new baselines,in: International conference on pattern recognition applications and methods, volume 2, SciTePress, 2017, pp. 324–331.
292. Long et al. [2017]Y. Long, Y. Gong, Z. Xiao, Q. Liu,Accurate object localization in remote sensing images based on convolutional neural networks,IEEE Transactions on Geoscience and Remote Sensing 55 (2017) 2486–2498.
293. Cheng et al. [2017]G. Cheng, J. Han, X. Lu,Remote sensing image scene classification: Benchmark and state of the art,Proceedings of the IEEE 105 (2017) 1865–1883.
294. Cheng et al. [2014]G. Cheng, J. Han, P. Zhou, L. Guo,Multi-class geospatial object detection and geographic image classification based on collection of part detectors,ISPRS Journal of Photogrammetry and Remote Sensing 98 (2014) 119–132.
295. Amraoui et al. [2022]K. E. Amraoui, M. Lghoul, A. Ezzaki, L. Masmoudi, M. Hadri, H. Elbelrhiti, A. A. Simo,Avo-airdb: An avocado uav database for agricultural image segmentation and classification,Data in Brief 45 (2022) 108738.
296. Raptis et al. [2023]E. K. Raptis, M. Krestenitis, K. Egglezos, O. Kypris, K. Ioannidis, L. Doitsidis, A. C. Kapoutsis, S. Vrochidis, I. Kompatsiaris, E. B. Kosmatopoulos,End-to-end precision agriculture uav-based functionalities tailored to field characteristics,Journal of Intelligent & Robotic Systems 107 (2023) 23.
297. Tetila et al. [2024]E. C. Tetila, B. L. Moro, G. Astolfi, A. B. da Costa, W. P. Amorim, N. A. de Souza Belete, H. Pistori, J. G. A. Barbedo,Real-time detection of weeds by species in soybean using uav images,Crop Protection 184 (2024) 106846.
298. Ödübek and Atik [2024]E. Ödübek, M. E. Atik,Detection of asphalt pavement cracks with yolo architectures from unmanned aerial vehicle images,in: 2024 32nd Signal Processing and Communications Applications Conference (SIU), IEEE, 2024, pp. 1–4.
299. Vieira e Silva et al. [2023]A. L. B. Vieira e Silva, H. de Castro Felix, F. P. M. Simões, V. Teichrieb, M. dos Santos, H. Santiago, V. Sgotti, H. Lott Neto,Insplad: A dataset and benchmark for power line asset inspection in uav images,International journal of remote sensing 44 (2023) 7294–7320.
300. Mishra et al. [2020]B. Mishra, D. Garg, P. Narang, V. Mishra,Drone-surveillance for search and rescue in natural disaster,Computer Communications 156 (2020) 1–10.
301. Wang and Mahmoudian [2023]Z. Wang, N. Mahmoudian,Aerial fluvial image dataset for deep semantic segmentation neural networks and its benchmarks,IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2023) 4755–4766.
302. Rahnemoonfar et al. [2021]M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, R. R. Murphy,Floodnet: A high resolution aerial imagery dataset for post flood scene understanding,IEEE Access 9 (2021) 89644–89654.
303. Pan et al. [2024]L. Pan, C. Song, X. Gan, K. Xu, Y. Xie,Military image captioning for low-altitude uav or ugv perspectives,Drones 8 (2024) 421.
304. Mou et al. [2023]C. Mou, T. Liu, C. Zhu, X. Cui,Waid: A large-scale dataset for wildlife detection with drones,Applied Sciences 13 (2023) 10397.
305. Shah et al. [2018]S. Shah, D. Dey, C. Lovett, A. Kapoor,Airsim: High-fidelity visual and physical simulation for autonomous vehicles,in: Field and Service Robotics: Results of the 11th International Conference, Springer, 2018, pp. 621–635.
306. Dosovitskiy et al. [2017]A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, V. Koltun,Carla: An open urban driving simulator,in: Conference on robot learning, PMLR, 2017, pp. 1–16.
307. NVIDIA [2024]NVIDIA, Isaac sim robotics simulator, 2024. URL: https://developer.nvidia.com/isaac/sim, accessed: 2024-11-01.
308. Gao et al. [2024]C. Gao, B. Zhao, W. Zhang, J. Zhang, J. Mao, Z. Zheng, F. Man, J. Fang, Z. Zhou, J. Cui, X. Chen, Y. Li,Embodiedcity: A benchmark platform for embodied agent in real-world city environment,arXiv preprint (2024).
309. Zhang et al. [2021]C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals,Understanding deep learning (still) requires rethinking generalization,Communications of the ACM 64 (2021) 107–115.
310. Crawshaw [2020]M. Crawshaw,Multi-task learning with deep neural networks: A survey,arXiv preprint arXiv:2009.09796 (2020).
311. Gehrmann et al. [2019]S. Gehrmann, H. Strobelt, R. Krüger, H. Pfister, A. M. Rush,Visual interaction with deep learning models through collaborative semantic inference,IEEE transactions on visualization and computer graphics 26 (2019) 884–894.
312. Li et al. [2024]H. Li, X. Liu, G. Li,A benchmark for uav-view natural language-guided tracking,Electronics 13 (2024) 1706.

313. Ma et al. [2024]Z. Ma, Y. Li, R. Ma, C. Liang,Unsupervised semantic segmentation of high-resolution uav imagery for road scene parsing,arXiv preprint arXiv:2402.02985 (2024).
314. Limberg et al. [2024]C. Limberg, A. Gonçalves, B. Rigault, H. Prendinger,Leveraging yolo-world and gpt-4v lmms for zero-shot person detection and action recognition in drone imagery,arXiv preprint arXiv:2404.01571 (2024).
315. Kim et al. [2024]H. Kim, D. Lee, S. Park, Y. M. Ro,Weather-aware drone-view object detection via environmental context understanding,in: 2024 IEEE International Conference on Image Processing (ICIP), IEEE, 2024, pp. 549–555.
316. Sakaino [2023]H. Sakaino,Dynamic texts from uav perspective natural images,in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2070–2081.
317. Liang et al. [2023]F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, D. Marculescu,Open-vocabulary semantic segmentation with mask-adapted clip,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7061–7070.
318. Gu et al. [2021]X. Gu, T.-Y. Lin, W. Kuo, Y. Cui,Open-vocabulary object detection via vision and language knowledge distillation,arXiv preprint arXiv:2104.13921 (2021).
319. Gong et al. [2024]Z. Gong, Z. Wei, D. Wang, X. Ma, H. Chen, Y. Jia, Y. Deng, Z. Ji, X. Zhu, N. Yokoya, et al.,Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation,arXiv preprint arXiv:2410.22629 (2024).
320. Florea and Nedevschi [2024]H. Florea, S. Nedevschi,Tandepth: Leveraging global dems for metric monocular depth estimation in uavs,arXiv preprint arXiv:2409.05142 (2024).
321. de Zarzà et al. [2023]I. de Zarzà, J. de Curtò, C. T. Calafate,Socratic video understanding on unmanned aerial vehicles,Procedia Computer Science 225 (2023) 144–154.
322. Zhao et al. [2023]H. Zhao, F. Pan, H. Ping, Y. Zhou,Agent as cerebrum, controller as cerebellum: Implementing an embodied lmm-based agent on drones,arXiv preprint arXiv:2311.15033 (2023).
323. Bazi et al. [2024]Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, F. Melgani,Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery,Remote Sensing 16 (2024) 1477.
324. Zhang et al. [2024]Z. Zhang, T. Zhao, Y. Guo, J. Yin,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 (2024).
325. Zhan et al. [2024]Y. Zhan, Z. Xiong, Y. Yuan,Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model,arXiv preprint arXiv:2401.09712 (2024).
326. Zhang et al. [2024]J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, H. Wang,Navid: Video-based vlm plans the next step for vision-and-language navigation,arXiv preprint arXiv:2402.15852 (2024).
327. Hong et al. [2024]H. Hong, S. Wang, Z. Huang, Q. Wu, J. Liu,Why only text: Empowering vision-and-language navigation with multi-modal prompts,arXiv preprint arXiv:2406.02208 (2024).
328. Gao et al. [2024]Y. Gao, Z. Wang, L. Jing, D. Wang, X. Li, B. Zhao,Aerial vision-and-language navigation via semantic-topo-metric representation guided llm reasoning,arXiv preprint arXiv:2410.08500 (2024).
329. Wang et al. [2024]X. Wang, D. Yang, Z. Wang, H. Kwan, J. Chen, W. Wu, H. Li, Y. Liao, S. Liu,Towards realistic uav vision-language navigation: Platform, benchmark, and methodology,arXiv preprint arXiv:2410.07087 (2024).
330. Sanyal and Roy [2024]S. Sanyal, K. Roy,Asma: An adaptive safety margin algorithm for vision-language drone navigation via scene-aware control barrier functions,arXiv preprint arXiv:2409.10283 (2024).
331. Zhang et al. [2024]W. Zhang, Y. Liu, X. Wang, X. Chen, C. Gao, X. Chen,Demo abstract: Embodied aerial agent for city-level visual language navigation using large language model,in: 2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2024, pp. 265–266.
332. Chen et al. [2023]Z. Chen, J. Li, F. Fukumoto, P. Liu, Y. Suzuki,Vision-language navigation for quadcopters with conditional transformer and prompt-based text rephraser,in: Proceedings of the 5th ACM International Conference on Multimedia in Asia, 2023, pp. 1–7.
333. Blei et al. [2024]Y. Blei, M. Krawez, N. Nilavadi, T. K. Kaiser, W. Burgard,Cloudtrack: Scalable uav tracking with cloud semantics,arXiv preprint arXiv:2409.16111 (2024).
334. Cai et al. [2024]Z. Cai, C. R. Cardenas, K. Leo, C. Zhang, K. Backman, H. Li, B. Li, M. Ghorbanali, S. Datta, L. Qu, et al.,Neusis: A compositional neuro-symbolic framework for autonomous perception, reasoning, and planning in complex uav search missions,arXiv preprint arXiv:2409.10196 (2024).
335. Döschl and Kiam [2024]B. Döschl, J. J. Kiam,Say-reapex: An llm-modulo uav online planning framework for search and rescue,in: 2nd CoRL Workshop on Learning Effective Abstractions for Planning, 2024.
336. Ravichandran et al. [2024]Z. Ravichandran, V. Murali, M. Tzes, G. J. Pappas, V. Kumar,Spine: Online semantic planning for missions with incomplete natural language specifications in unstructured environments,arXiv preprint arXiv:2410.03035 (2024).
337. Aikins et al. [2024]G. Aikins, M. P. Dao, K. J. Moukpe, T. C. Eskridge, K.-D. Nguyen,Leviosa: Natural language-based uncrewed aerial vehicle trajectory generation,Electronics 13 (2024) 4508.
338. Cui et al. [2024]J. Cui, G. Liu, H. Wang, Y. Yu, J. Yang,Tpml: Task planning for multi-uav system with large language models,in: 2024 IEEE 18th International Conference on Control & Automation (ICCA), IEEE, 2024, pp. 886–891.
339. pen [2023]PCL. Peng Cheng Mind. Accessed: 2023-02-08, 2023. URL: https://openi.pcl.ac.cn/PengChengMind/PengCheng.Mind.
340. Liu et al. [2024]Y. Liu, Z. Zhou, J. Liu, L. Chen, J. Wang,Multi-agent formation control using large language models,Authorea Preprints (2024).
341. Vemprala et al. [2024]S. H. Vemprala, R. Bonatti, A. Bucker, A. Kapoor,Chatgpt for robotics: Design principles and model abilities,IEEE Access (2024).
342. Zhong et al. [2024]J. Zhong, M. Li, Y. Chen, Z. Wei, F. Yang, H. Shen,A safer vision-based autonomous planning system for quadrotor uavs with dynamic obstacle trajectory prediction and its application with llms,in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 920–929.
343. TAZIR et al. [2023]M. L. TAZIR, M. MANCAS, T. DUTOIT,From words to flight: Integrating openai chatgpt with px4/gazebo for natural language-based drone control,in: International Workshop on Computer Science and Engineering, 2023.
344. Phadke et al. [2024]A. Phadke, A. Hadimlioglu, T. Chu, C. N. Sekharan,Integrating large language models for uav control in simulated environments: A modular interaction approach,arXiv preprint arXiv:2410.17602 (2024).
345. Liu et al. [2024]G. Liu, T. Sun, W. Li, X. Li, X. Liu, J. Cui,Eai-sim: An open-source embodied ai simulation framework with large language models,in: 2024 IEEE 18th International Conference on Control & Automation (ICCA), IEEE, 2024, pp. 994–999.

346. Zhu et al. [2024]T. Zhu, W. Newton, S. Embury, Y. Sun,Taiist cps-uav at the sbft tool competition 2024,in: Proceedings of the 17th ACM/IEEE International Workshop on Search-Based and Fuzz Testing, 2024, pp. 51–52.
347. Jiao et al. [2023]A. Jiao, T. P. Patel, S. Khurana, A.-M. Korol, L. Brunke, V. K. Adajania, U. Culha, S. Zhou, A. P. Schoellig,Swarm-gpt: Combining large language models with safe motion planning for robot choreography design,arXiv preprint arXiv:2312.01059 (2023).
348. Lykov et al. [2024]A. Lykov, S. Karaf, M. Martynov, V. Serpiva, A. Fedoseev, M. Konenkov, D. Tsetserukou,Flockgpt: Guiding uav flocking with linguistic orchestration,arXiv preprint arXiv:2405.05872 (2024).
349. Pueyo et al. [2024]P. Pueyo, E. Montijano, A. C. Murillo, M. Schwager,Clipswarm: Generating drone shows from text prompts with vision-language models,arXiv preprint arXiv:2403.13467 (2024).
350. Li et al. [2024]X. Li, X. Feng, S. Hu, M. Wu, D. Zhang, J. Zhang, K. Huang,Dtllm-vlt: Diverse text generation for visual language tracking based on llm,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7283–7292.
351. Arrabi et al. [2024]A. Arrabi, X. Zhang, W. Sultani, C. Chen, S. Wshah,Cross-view meets diffusion: Aerial image synthesis with geometry and text guidance,arXiv preprint arXiv:2408.04224 (2024).
352. Yao et al. [2024]F. Yao, Y. Yue, Y. Liu, X. Sun, K. Fu,Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models,arXiv preprint arXiv:2408.15511 (2024).
353. Tang et al. [2024]Y.-C. Tang, P.-Y. Chen, T.-Y. Ho,Defining and evaluating physical safety for large language models,arXiv preprint arXiv:2411.02317 (2024).
354. Xu et al. [2024]Y. Xu, Z. Jian, J. Zha, X. Chen,Emergency networking using uavs: A reinforcement learning approach with large language model,in: 2024 23rd ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2024, pp. 281–282.
355. Xiang et al. [2024]X. Xiang, J. Xue, L. Zhao, Y. Lei, C. Yue, K. Lu,Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning,in: 2024 International Joint Conference on Neural Networks (IJCNN), IEEE, 2024, pp. 1–8.
356. Pineli Simões et al. [2024]L. E. Pineli Simões, L. Brandão Rodrigues, R. Mota Silva, G. Rodrigues da Silva,Evaluating voice command pipelines for drone control: From stt and llm to direct classification and siamese networks,arXiv e-prints (2024) arXiv–2407.
357. Huang et al. [2022]Y. Huang, J. Chen, D. Huang,Ufpmp-det: Toward accurate and efficient object detection on drone imagery,in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 1026–1033.
358. Zhu et al. [2021]X. Zhu, S. Lyu, X. Wang, Q. Zhao,Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios,in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.
359. Fu et al. [2023]X. Fu, G. Wei, X. Yuan, Y. Liang, Y. Bo,Efficient yolov7-drone: an enhanced object detection approach for drone aerial imagery,Drones 7 (2023) 616.
360. Yang et al. [2020]W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu, et al.,Advancing image understanding in poor visibility environments: A collective benchmark study,IEEE Transactions on Image Processing 29 (2020) 5737–5752.
361. Tan et al. [2021]L. Tan, X. Lv, X. Lian, G. Wang,Yolov4_drone: Uav image target detection based on an improved yolov4 algorithm,Computers & Electrical Engineering 93 (2021) 107261.
362. Fang et al. [2023]W. Fang, G. Zhang, Y. Zheng, Y. Chen,Multi-task learning for uav aerial object detection in foggy weather condition,Remote Sensing 15 (2023) 4617.
363. Hoanh and Pham [2024]N. Hoanh, T. V. Pham,A multi-task framework for car detection from high-resolution uav imagery focusing on road regions,IEEE Transactions on Intelligent Transportation Systems (2024).
364. Jing et al. [2022]H. Jing, Y. Cheng, H. Wu, H. Wang,Radar target detection with multi-task learning in heterogeneous environment,IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.
365. Qu et al. [2024]H. Qu, Y. Cai, J. Liu,Llms are good action recognizers,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18395–18406.
366. Han and Lim [2024]G. Han, S.-N. Lim,Few-shot object detection with foundation models,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 28608–28618.
367. Lin et al. [2024]C. Lin, Y. Jiang, L. Qu, Z. Yuan, J. Cai,Generative region-language pretraining for open-ended object detection,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13958–13968.
368. Zang et al. [2024]Y. Zang, W. Li, J. Han, K. Zhou, C. C. Loy,Contextual object detection with multimodal large language models,International Journal of Computer Vision (2024) 1–19.
369. Yang et al. [2024]F. Yang, S. Zhao, Y. Zhang, H. Chen, H. Chen, W. Tang, H. Lu, P. Xu, Z. Yang, J. Han, et al.,Llmi3d: Empowering llm with 3d perception from a single 2d image,arXiv preprint arXiv:2408.07422 (2024).
370. Huang et al. [2023]L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al.,A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,ACM Transactions on Information Systems (2023).
371. Liu et al. [2024]H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, W. Peng,A survey on hallucination in large vision-language models,arXiv preprint arXiv:2402.00253 (2024).
372. Favero et al. [2024]A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, S. Soatto,Multi-modal hallucination control by visual information grounding,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14303–14312.
373. Zhao et al. [2021]Q. Zhao, J. Liu, Y. Li, H. Zhang,Semantic segmentation with attention mechanism for remote sensing images,IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–13.
374. Yuan et al. [2021]X. Yuan, J. Shi, L. Gu,A review of deep learning methods for semantic segmentation of remote sensing imagery,Expert Systems with Applications 169 (2021) 114417.
375. Cai et al. [2022]Y. Cai, Y. Yang, Y. Shang, Z. Chen, Z. Shen, J. Yin,Iterdanet: Iterative intra-domain adaptation for semantic segmentation of remote sensing images,IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–17.
376. Bai et al. [2022]L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, S. Ouyang,Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning,IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–13.
377. He et al. [2016]K. He, X. Zhang, S. Ren, J. Sun,Deep residual learning for image recognition,in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

378. Li et al. [2024]L. Li, Y. Zhang, Z. Jiang, Z. Wang, L. Zhang, H. Gao,Unmanned aerial vehicle-neural radiance field (uav-nerf): Learning multiview drone three-dimensional reconstruction with neural radiance field,Remote Sensing 16 (2024) 4168.
379. Wu et al. [2024]Y. Wu, J. Liu, S. Ji,3d gaussian splatting for large-scale surface reconstruction from aerial images,arXiv preprint arXiv:2409.00381 (2024).
380. Florea and Nedevschi [2022]H. Florea, S. Nedevschi,Survey on monocular depth estimation for unmanned aerial vehicles using deep learning,in: 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2022, pp. 319–326.
381. Chang et al. [2023]R. Chang, K. Yu, Y. Yang,Self-supervised monocular depth estimation using global and local mixed multi-scale feature enhancement network for low-altitude uav remote sensing,Remote Sensing 15 (2023) 3275.
382. Yu et al. [2023]K. Yu, H. Li, L. Xing, T. Wen, D. Fu, Y. Yang, C. Zhou, R. Chang, S. Zhao, L. Xing, et al.,Scene-aware refinement network for unsupervised monocular depth estimation in ultra-low altitude oblique photography of uav,ISPRS Journal of Photogrammetry and Remote Sensing 205 (2023) 284–300.
383. Antol et al. [2015]S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh,Vqa: Visual question answering,in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.
384. Goyal et al. [2017]Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh,Making the v in vqa matter: Elevating the role of image understanding in visual question answering,in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.
385. Zhou et al. [2020]L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, J. Gao,Unified vision-language pre-training for image captioning and vqa,in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13041–13049.
386. Hu et al. [2022]X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang,Scaling up vision-language pre-training for image captioning,in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17980–17989.
387. Wang et al. [2023]X. Wang, X. Cui, D. Li, F. Liu, L. Jiao,Multi-model fusion for aerial vision and dialog navigation based on human attention aids,arXiv preprint arXiv:2308.14064 (2023).
388. Chu et al. [2025]M. Chu, Z. Zheng, W. Ji, T. Wang, T.-S. Chua,Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching,in: European Conference on Computer Vision, Springer, 2025, pp. 213–231.
389. Zhang et al. [2023]L. Zhang, A. Rao, M. Agrawala,Adding conditional control to text-to-image diffusion models,in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.
390. Liu et al. [2024]R. Liu, W. Wang, Y. Yang,Volumetric environment representation for vision-language navigation,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16317–16328.
391. Li et al. [2024]X. Li, S. Hu, X. Feng, D. Zhang, M. Wu, J. Zhang, K. Huang,Dtvlt: A multi-modal diverse text benchmark for visual language tracking based on llm,arXiv preprint arXiv:2410.02492 (2024).
392. Sun et al. [2024]L. Sun, X. Li, Z. Yang, D. Gao,Visual object tracking based on the motion prediction and block search in uav videos,Drones 8 (2024) 252.
393. Wu et al. [2019]C. Wu, B. Ju, Y. Wu, X. Lin, N. Xiong, G. Xu, H. Li, X. Liang,Uav autonomous target search based on deep reinforcement learning in complex disaster scene,IEEE Access 7 (2019) 117227–117245.
394. Hou et al. [2023]Y. Hou, J. Zhao, R. Zhang, X. Cheng, L. Yang,Uav swarm cooperative target search: A multi-agent reinforcement learning approach,IEEE Transactions on Intelligent Vehicles (2023).
395. Bethke et al. [2008]B. Bethke, M. Valenti, J. P. How,Uav task assignment,IEEE robotics & automation magazine 15 (2008) 39–44.
396. Zhou et al. [2018]Z. Zhou, J. Feng, B. Gu, B. Ai, S. Mumtaz, J. Rodriguez, M. Guizani,When mobile crowd sensing meets uav: Energy-efficient task assignment and route planning,IEEE Transactions on Communications 66 (2018) 5526–5538.
397. Mao et al. [2024]X. Mao, G. Wu, M. Fan, Z. Cao, W. Pedrycz,Dl-drl: A double-level deep reinforcement learning approach for large-scale task scheduling of multi-uav,IEEE Transactions on Automation Science and Engineering (2024).
398. Tejaswi and Lee [2022]K. Tejaswi, T. Lee,Constrained imitation learning for a flapping wing unmanned aerial vehicle,IEEE Robotics and Automation Letters 7 (2022) 10534–10541.
399. Shukla et al. [2020]D. Shukla, S. Keshmiri, N. Beckage,Imitation learning for neural network autopilot in fixed-wing unmanned aerial systems,in: 2020 International Conference on Unmanned Aircraft Systems (ICUAS), IEEE, 2020, pp. 1508–1517.
400. Choi and Ahn [2020]U. Choi, J. Ahn,Imitation learning-based unmanned aerial vehicle planning for multitarget reconnaissance under uncertainty,Journal of Aerospace Information Systems 17 (2020) 36–50.
401. Liang et al. [2023]Z. Liang, Q. Li, G. Fu,Multi-uav collaborative search and attack mission decision-making in unknown environments,Sensors 23 (2023) 7398.
402. Wang and Wang [2024]H. Wang, J. Wang,Enhancing multi-uav air combat decision making via hierarchical reinforcement learning,Scientific Reports 14 (2024) 4458.
403. Du et al. [2024]Y. Du, N. Qi, X. Li, M. Xiao, A.-A. A. Boulogeorgos, T. A. Tsiftsis, Q. Wu,Distributed multi-uav trajectory planning for downlink transmission: a gnn-enhanced drl approach,IEEE Wireless Communications Letters (2024).
404. Li et al. [2022]K. Li, W. Ni, X. Yuan, A. Noor, A. Jamalipour,Deep-graph-based reinforcement learning for joint cruise control and task offloading for aerial edge internet of things (edgeiot),IEEE Internet of Things Journal 9 (2022) 21676–21686.
405. Courbon et al. [2010]J. Courbon, Y. Mezouar, N. Guénard, P. Martinet,Vision-based navigation of unmanned aerial vehicles,Control engineering practice 18 (2010) 789–799.
406. Liu et al. [2022]Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie,A convnet for the 2020s,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
407. Devlin [2018]J. Devlin,Bert: Pre-training of deep bidirectional transformers for language understanding,arXiv preprint arXiv:1810.04805 (2018).
408. De Curtò et al. [2023]J. De Curtò, I. De Zarza, C. T. Calafate,Semantic scene understanding with large language models on unmanned aerial vehicles,Drones 7 (2023) 114.
409. Wang et al. [2023]C. Wang, Z. Zhong, X. Xiang, Y. Zhu, L. Wu, D. Yin, J. Li,Uav path planning in multi-task environments with risks through natural language understanding,Drones 7 (2023) 147.
410. Kuwertz et al. [2018]A. Kuwertz, D. Mühlenberg, J. Sander, W. Müller,Applying knowledge-based reasoning for information fusion in intelligence, surveillance, and reconnaissance,in: Multisensor Fusion and Integration in the Wake of Big Data, Deep Learning and Cyber Physical System: An Edition of the Selected Papers from the 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2017), Springer, 2018, pp. 119–139.
411. Feng et al. [2023]Y. Feng, H. Snoussi, J. Teng, A. Cherouat, T. Wang,Large language model-based multi-task uavs-towards distilled real-time interactive control,in: IET Conference Proceedings CP870, volume 2023, IET, 2023, pp. 114–118.

412. Mahajan et al. [2023]V. Mahajan, E. Barmpounakis, M. R. Alam, N. Geroliminis, C. Antoniou,Treating noise and anomalies in vehicle trajectories from an experiment with a swarm of drones,IEEE Transactions on Intelligent Transportation Systems 24 (2023) 9055–9067.
413. Telikani et al. [2024]A. Telikani, A. Sarkar, B. Du, J. Shen,Machine learning for uav-aided its: A review with comparative study,IEEE Transactions on Intelligent Transportation Systems (2024).
414. Bisio et al. [2022]I. Bisio, C. Garibotto, H. Haleem, F. Lavagetto, A. Sciarrone,A systematic review of drone based road traffic monitoring system,Ieee Access 10 (2022) 101537–101555.
415. Saputro et al. [2018]N. Saputro, K. Akkaya, R. Algin, S. Uluagac,Drone-assisted multi-purpose roadside units for intelligent transportation systems,in: 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), IEEE, 2018, pp. 1–5.
416. Dung [2019]N. D. Dung,Developing models for managing drones in the transportation system in smart cities,Electrical, Control and Communication Engineering 15 (2019) 71–78.
417. Menouar et al. [2017]H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, A. Tuncer,Uav-enabled intelligent transportation systems for the smart city: Applications and challenges,IEEE Communications Magazine 55 (2017) 22–28.
418. Wang et al. [2023]L. Wang, X. Deng, J. Gui, P. Jiang, F. Zeng, S. Wan,A review of urban air mobility-enabled intelligent transportation systems: Mechanisms, applications and challenges,Journal of Systems Architecture 141 (2023) 102902.
419. Yao et al. [2024]J. Yao, J. Li, Y. Li, M. Zhang, C. Zuo, S. Dong, Z. Dai,A vision–language model-based traffic sign detection method for high-resolution drone images: A case study in guyuan, china,Sensors 24 (2024) 5800.
420. Yuan et al. [2024]Z. Yuan, F. Xie, T. Ji,Patrol agent: An autonomous uav framework for urban patrol using on board vision language model and on cloud large language model,in: 2024 6th International Conference on Robotics and Computer Vision (ICRCV), IEEE, 2024, pp. 237–242.
421. Zhu et al. [2024]H. Zhu, S. Qin, M. Su, C. Lin, A. Li, J. Gao,Harnessing large vision and language models in agriculture: A review,arXiv preprint arXiv:2407.19679 (2024).
422. Tian et al. [2024]Y. Tian, F. Lin, X. Zhang, J. Ge, Y. Wang, X. Dai, Y. Lv, F.-Y. Wang,Logisticsvista: 3d terminal delivery services with uavs, ugvs and usvs based on foundation models and scenarios engineering,IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI) (2024).
423. Jiang et al. [2024]H. Jiang, T. Wu, X. Ren, L. Gou,Optimisation of multi-type logistics uav scheduling under high demand,Promet-Traffic&Transportation 36 (2024) 115–131.
424. Huang et al. [2020]H. Huang, A. V. Savkin, C. Huang,Scheduling of a parcel delivery system consisting of an aerial drone interacting with public transportation vehicles,Sensors 20 (2020) 2045.
425. Wandelt et al. [2023]S. Wandelt, S. Wang, C. Zheng, X. Sun,Aerial: A meta review and discussion of challenges toward unmanned aerial vehicle operations in logistics, mobility, and monitoring,IEEE Transactions on Intelligent Transportation Systems (2023).
426. Luo et al. [2024]S. Luo, Y. Yao, H. Zhao, L. Song,A language model-based fine-grained address resolution framework in uav delivery system,IEEE Journal of Selected Topics in Signal Processing (2024).
427. Dong et al. [2024]C. Dong, N. Syed, F. Jiang, R. Elphick-Darling, S. Chen, J. Zhang, M. Lu, X. Liu,Securing uav delivery systems with blockchain and large language models: an innovative logistics solution,in: 2024 11th International Conference on Machine Intelligence Theory and Applications (MiTA), IEEE, 2024, pp. 1–8.
428. Jin et al. [2020]W. Jin, J. Yang, Y. Fang, W. Feng,Research on application and deployment of uav in emergency response,in: 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), IEEE, 2020, pp. 277–280.
429. Goecks and Waytowich [2023]V. G. Goecks, N. R. Waytowich,Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios,arXiv preprint arXiv:2306.17271 (2023).
430. Fourati and Alouini [2021]F. Fourati, M.-S. Alouini,Artificial intelligence for satellite communication: A review,Intelligent and Converged Networks 2 (2021) 213–243.
431. Wang et al. [2024]Y. Wang, J. Farooq, H. Ghazzai, G. Setti,Multi-uav placement for integrated access and backhauling using llm-driven optimization (2024).
432. Gong et al. [2024]Z. Gong, Z. Wei, D. Wang, X. Ma, H. Chen, Y. Jia, Y. Deng, Z. Ji, X. Zhu, N. Yokoya, J. Zhang, B. Du, L. Zhang,Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation,arXiv preprint arXiv:2410.22629 (2024).
433. Hong et al. [2023]Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, C. Gan,3d-llm: Injecting the 3d world into large language models,Advances in Neural Information Processing Systems 36 (2023) 20482–20494.
434. Zhang et al. [2024]S. Zhang, D. Huang, J. Deng, S. Tang, W. Ouyang, T. He, Y. Zhang,Agent3d-zero: An agent for zero-shot 3d understanding,arXiv preprint arXiv:2403.11835 (2024).
435. Hu et al. [2021]E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen,Lora: Low-rank adaptation of large language models,arXiv preprint arXiv:2106.09685 (2021).
436. Casper et al. [2023]S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, et al.,Open problems and fundamental limitations of reinforcement learning from human feedback,arXiv preprint arXiv:2307.15217 (2023).
437. Chen et al. [2024]B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, F. Xia,Spatialvlm: Endowing vision-language models with spatial reasoning capabilities,in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14455–14465.