

# 对话起飞：实现基于 PX4 无人机代理的自然语言控制

Shoon Kit Lim 

University of Southampton Malaysia,  
Iskandar Puteri, Johor, Malaysia  
skl1g14@soton.ac.uk

Jing Huey Khor 

Connected Intelligence Research Group,  
University of Southampton Malaysia,  
Iskandar Puteri, Johor, Malaysia  
j.khor@soton.ac.uk

Melissa Jia Ying Chong 

University of Southampton Malaysia,  
Iskandar Puteri, Johor, Malaysia  
m.j.y.chong@soton.ac.uk

Ting Yang Ling 

Sustainable Electronic Technologies,  
University of Southampton,  
Southampton SO17 1BJ U.K.  
ivan.ling@soton.ac.uk

**Abstract**—近期在代理和物理人工智能 (AI) 方面的进展主要集中在地面平台，例如人形和轮式机器人，而对空中机器人的研究相对较少。同时，最先进的无人机多模态视觉语言系统通常依赖于仅对资源丰富的组织开放的闭源模型。为实现自主无人机的自然语言控制的普及，本文提出了一个开源代理框架，该框架集成了基于 PX4 的飞行控制、机器人操作系统 2 (ROS2) 中间件以及使用 Ollama 的本地托管模型。性能在模拟和定制四旋翼平台上进行了评估，基准测试了四个大语言模型 (LLM) 系列用于命令生成，以及三个视觉语言模型 (VLM) 系列用于场景理解。结果表明，LLM，特别是 Gemma3、Qwen2.5 和 Llama-3.2，持续产生了 100% 有效的飞行指令，而 DeepSeek-LLM 的表现则显著较低，仅为 38%。此外，所有评估的 VLM，包括 Gemma3、Llama3.2-Vision 以及 Llava1.6，均能够检测到指定对象的存在，并给出有效的二进制响应，范围从 97% 到 100%。任务成功率根据模型配对而异，其中 Gemma3 LLM 和 VLM 组合观察到的最高成功率为 40%。源码、模型配置和提示模板可在 <https://github.com/limshoonkit/ros2-agent-ws> 公开获取。

**Index Terms**—PX4, UAV, VLM, LLM

## I. 引言

工业 5.0 标志着制造业的重大转型，从数字自动化向一种协作方法迈进，在这种方法中，人类与智能机器一起工作 [?]。自欧洲联盟于 2021 年引入这一范式转变以来，它已在各个行业迅速获得关注。它推动了自动化的边界，强调在智能工厂中，人类与机器人代理并肩工作以优化生产过程的协作环境。

最近在加速计算和机器学习方面的进展使机器人能够实现更高的自主性和更灵活的控制，从而解锁与人类及其周围环境的更直观交互 [?]。实现这一目标的一个关键挑战在于赋予机器人感知、推理和根据镜像人类交流的指令行动的能力。为此，推出了开放 X-Embodiment 项目 [?]，旨在通过来自各种机器人体现和环境的大规模多样数据集来增强机器人的学习。该计划整合了来自 34 个研究实验室的 60 个现有机器人数据集的数据，创建了一个标准化的数据集，其中包含来自 21 个机构的 22 种不同机器人体现的超过一百万条机器人轨迹，以推动跨体现的机器人学习。

在这些技术进步的核心是大语言模型 (LLMs)，特别是生成式预训练转换器 (GPTs) [?] 的类人语言理解。通过

利用 LLMs，用户可以自然直观地交流复杂的指令，消除了对刚性编码命令和低级指令集的需求。在此基础上，视觉语言模型 (VLMs) 通过统一视觉和语言理解引入了多模态智能。经过成对图像和文本的大量数据集训练，VLMs 可以对世界发展出丰富的语义理解。因此，它们在图像字幕、视觉问答和复杂场景解释等视觉语言任务中表现出色 [?]

尽管大型语言模型 (LLMs) 和视觉语言模型 (VLMs) 在各个领域展示了泛化和适应能力，如 [?] 和 [?] 所证明的那样，它们在自主导航背景下的有效性仍然是一个开放的研究问题。无人驾驶飞行器 (UAVs) 所面临的挑战与地面机器人系统（如自动驾驶汽车、机器人操控器、多足机器人和类人机器人）截然不同。其主要挑战源自于运动引起的图像退化增加、图像采集与处理之间的时间不匹配，以及无人机对实时控制的严格要求，这些都进一步使得安全的人机协作复杂化。

本文提出了一种面向自然语言控制基于 PX4 的无人机的具身智能体框架，该框架建立在 `ros-agents`<sup>1</sup> 之上。该框架结合了 NVIDIA Isaac Sim 和预设环境，包括户外停车场、联合办公空间、医院、仓库和数据中心，用于软件在环 (SITL) 模拟。机器人操作系统 2 (ROS2) 封装器封装了 Ollama，以服务不同的 LLMs 和 VLMs，例如 Gemma、DeepSeek、Qwen 和 Llama 家族。模块化任务由各个节点管理，其中包括一个视觉问答节点，用于处理配对图像和用户查询以生成文本响应；路径规划节点，将目标点和当前位置转换为使用低级 PX4 飞行动作的无碰撞轨迹；以及地图编码节点，该节点通过将当前位置和图像中的语义信息嵌入到文本标记中进行地图表示。整体框架如图 1 所示。

总之，本文的贡献如下：

- 1) 一个基于 ROS 的代理框架，将 PX4 飞行控制栈与 Ollama 平台连接起来。
- 2) 在 Ollama 上对开源 LLMs 和 VLMs 进行的空中导航对比分析，通过仿真 (Isaac Sim) 和实际定制四旋翼飞行器实验进行评估。

<sup>1</sup><https://github.com/automatika-robotics/ros-agents>

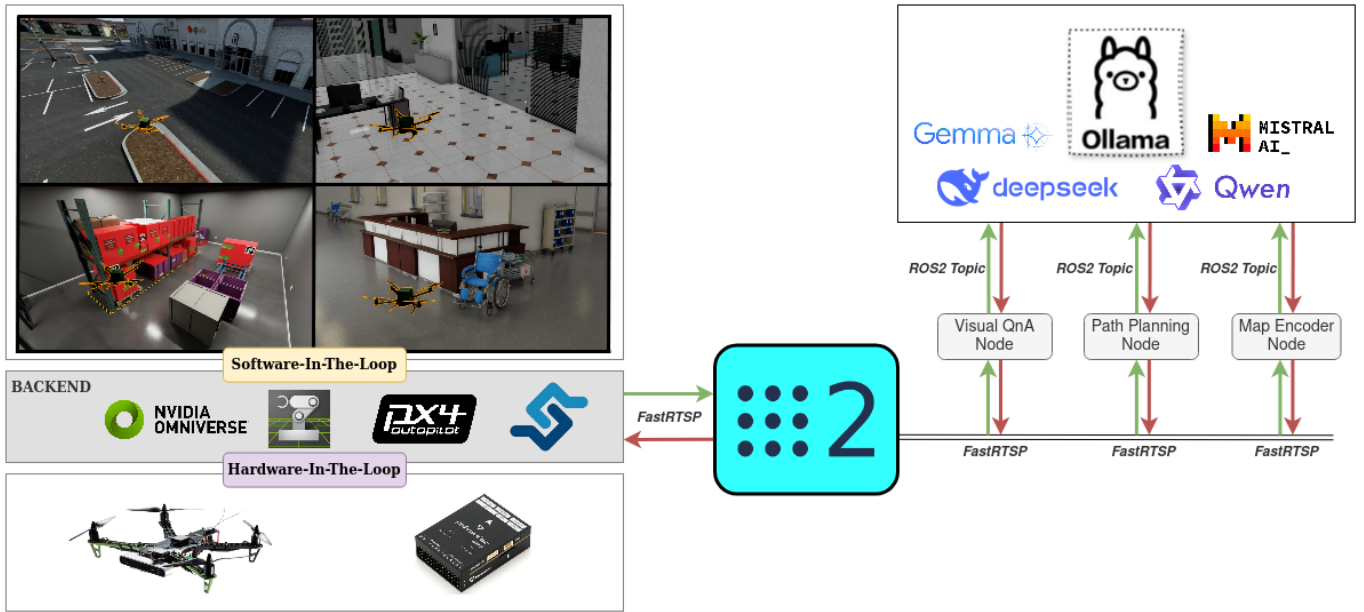


Fig. 1. 针对基于 PX4 的无人机代理的自然语言控制提出的框架。

本文的其余部分结构如下。第 II 节概述了该领域的相关工作。在第 ?? 节中, 描述了实验设置, 第 III 节中展示了引导无人机搜索和靠近目标时体现化代理的评估结果和相关讨论。最后, 第 IV 节对该工作进行了总结, 包括对潜在未来工作的回顾。

## II. 相关工作

之前的机器人研究在处理任务的广度和在不同的机体之间泛化的能力方面面临限制。虽然将语言条件整合到深度学习模型中显示出希望, 但许多现有的数据集仍然特定于特定的机器人平台。例如, RT-1 [?] 使用了 Transformer 架构和模仿学习, 而 RT-2 [?] 则通过将动作表示为文本标记的互联网规模预训练扩展了这一方法, 将其与自然语言标记统一在一个视觉-语言-行动 (VLA) 建模框架中。同时, SayCan [?] 采用语言模型来解释用户指令, 并为移动操纵器生成一系列可行步骤, 确保每个步骤都可以在现实世界中执行, 从而使机器人能够完成给定的高级人类任务。与此同时, SayTap [?] 提出了一个用于基于语言的四足机器人控制的新接口, 基于早期的语言调制技术。

多项研究已开发出专门针对无人机应用中的 LLM 和 VLM 的新型数据集和基准, 包括 AerialVLN [?]、CityNav [?]、AeroVerse [?]、AVDN [?] 和 UAV-Need-Help [?]。这些数据集旨在通过结合连续动作空间、城市和户外环境, 以及类似对话的复杂语言指令来捕捉逼真的空中动态。然而, 文献中反复出现的批评之一是缺乏在仿真到现实部署中的验证。此外, 解决自然语言指令固有的歧义性和可变性仍然存在挑战。考虑到无人机有限的电池寿命和操作范围, 一些数据集还涉及到可能不实际的大规模区域。此外, 对于带有噪声传感器数据的杂乱环境的考虑也不足。

OpenAI 的 ChatGPT 的显著成功推动了自然语言驱动的无人机编队的发展。Colosseum (前称微软 AirSim) 通过在 Unreal Engine 5 中将 ChatGPT 与 PX4 自动驾驶仪结合, 展示了这一点。这套系统通过 MAVLink 消息将

用户提示序列转换为飞行指令, 从而简化了逼真的 PX4 SITL 仿真。同样, 利用 Gazebo 模拟器的一项研究展示了 ChatGPT 生成 PX4 指挥命令以执行基本功能, 如武装、解除武装、起飞、着陆和飞行模式切换。扩展到更广泛的 ROS 生态系统, ROSA 利用 LLMs 通过推理-行动-观察循环进行一般机器人控制, 使其能够解释自然语言, 计划和执行动态可调用工具的操作, 并反复优化响应直到任务解决。此项工作与现有的方法不同, 避免仅依赖于闭源的 OpenAI ChatGPT 作为 LLM 框架。相反, 它使用 Ollama 作为 LLM 和 VLM 提供平台, 提供了集成更广泛模型的灵活性。

所提出的接口在模拟和真实环境中进行了评估和验证。实验设置如图 2 所示, 其中包括为一个定制建造的四旋翼飞行器和室内测试环境构建数字孪生。四旋翼飞行器安装了用于车载计算的 NVIDIA Jetson Orin Nano 开发套件、用于视觉惯性测程 (VIO) 的 ZED Mini 摄像头, 以及用于电机控制的 Pixhawk 6c Mini 飞行控制器。物理测试区域用尼龙网围成, 尺寸为  $7\text{m} \times 4.5\text{m} \times 2.2\text{m}$ , 整个空间内放置了纸箱作为静态障碍物。

在虚拟环境中使用 Isaac Sim 作为渲染和物理引擎进行了 SITL 模拟。这些模拟是在包含 NVIDIA RTX 3080Ti (16 GB) GPU 的 Ubuntu 22.04 工作站上运行的。另一台具有 RTX 3080Ti (16 GB) GPU 的 Ubuntu 22.04 工作站被用来承载 Ollama 模型, 它作为一个远程服务器, 通过标准局域网连接为物理和模拟系统提供语言推理能力。

体现化的智能体负责引导四旋翼飞行器从已知的初始配置到预定义的目标位置  $G$ , 同时遵循一组操作约束。正式地, 智能体必须生成一个离散时间轨迹:

$$\text{Traj}_{0:K} = (S_0, S_1, \dots, S_K) \quad (1)$$

其中  $S_0$  表示已知的初始状态,  $S_k$  表示期望的终端状态。状态转换由 PX4 自动驾驶仪的动态决定:



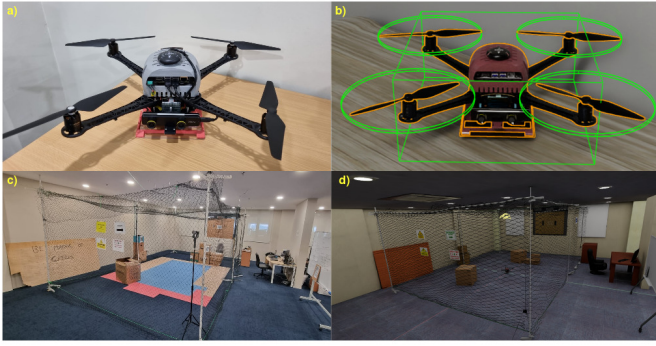


Fig. 2. 实验设置显示: (a) 定制四旋翼飞行器, (b) 四旋翼飞行器的数字孪生, (c) 室内飞行环境, 和 (d) 环境的数字孪生。

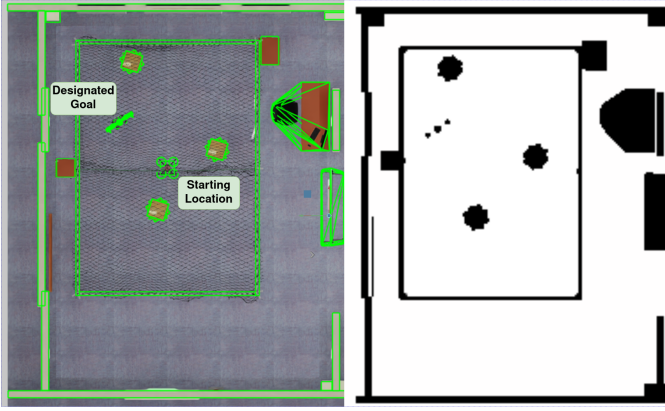


Fig. 3. 从自上而下的视角概述四旋翼飞行器的操作边界及其等效的二维占用网格图。

$$S_{k+1} = \text{PX4}(S_k, A_k) \quad (2)$$

其中,  $A_k$  是代理在离散步骤  $k$  时发出的控制输入。

代理的动作空间由两个具有受限参数范围的确定性运动原语定义:

- 如果  $A_k = \text{Turn}(\theta)$ ; , 那么  $\theta \in [\pm 90^\circ]$ 。  $\theta$  的负值对应左偏航 (逆时针), 正值对应右偏航 (顺时针)。
- 如果  $A_k = \text{Move}(d)$ ; , 则  $d \in [\pm 3.0]$  米。这表示四轴飞行器沿其机身框架的前轴前后移动了  $d$  的距离。

在每个时间索引  $k$ , 代理计算

$$A_k = \mathcal{LLM}(C_k, \mathcal{H}_k^{(N)}, \text{VLM}(\mathcal{V}_k)) \quad (3)$$

, 其中  $C_k$  代表特定任务的上下文查询 (例如, 状态元数据、任务指令、环境提示),  $\mathcal{H}_k^{(N)}$  表示代理执行的最近  $N$  个动作命令的历史,  $\mathcal{V}_k$  是从四旋翼飞行器机载摄像头获取的当前视觉观测集。一个 LLM 在给定上下文输入和最近动作历史的情况下生成动作命令, 而一个 VLM 评估视觉输入以确定当前场景中是否存在与任务相关的对象。LLM 和 VLM 协同工作以实现任务目标。

除了动作空间的限制外, 四旋翼直升机的操作还受到其物理边界和任务长度的进一步限制。四旋翼直升机的位置信息状态  $S$  必须在整个任务过程中保持在一个预定义的

有界三维空间区域  $\mathcal{B} \in \mathbb{R}^3$  内。此外, 智能体不能超过允许的最大时间步长  $K_{\max} \in \mathbb{R}^+$ :

$$\mathcal{S}[x, y, z] \in \mathcal{B}, \quad t \in [0, K_{\max}] \quad (4)$$

在每次仿真过程中, 目标物体及其位置在四旋翼飞行器的操作边界内随机生成, 并随机布置障碍, 如图 3 所示。四旋翼飞行器总是从相同的初始位置开始。障碍物的高度均设置为 1.5 米, 其宽度和长度按四旋翼飞行器翼展 0.56 米进行缩放。所有障碍物与目标之间保持至少 1 米的净空距离。

在每次实验的开始, 四轴飞行器被命令转换到 OFF-BOARD 模式, 起飞并上升到大约 1 米的标称高度。在导航试验期间, 这个高度会被保持, 限制飞行器在一个固定的垂直平面内。当达到预定高度后, 嵌入式代理将自主控制空中平台。

随后, 人工操作员通过发布 ROS2 字符串消息, 指定要搜索的目标对象和目标坐标的文本描述, 以启动任务。实验中指定的目标是一个视觉上可识别的对象, 例如另一个机器人或无人机。如果四旋翼机在 0.5 米范围内找到并接近目标, 则任务被视为成功。

构成任务失败的额外条件包括与障碍物碰撞或违反操作边界, 以及超过最大允许执行时间。在真实的实验中, 人类操作员作为安全后备机制存在。如果即将发生潜在的撞击, 操作员会将飞行模式切换到 POSITION 模式, 使四旋翼无人机在原地悬停, 从而允许操作员立即重新获得手动控制。

### III. 结果与讨论

本研究评估了撰写时可用的最新开源 LLM, 包括 Google Gemma3、阿里巴巴 Qwen2.5、Meta Llama 3.2 和 Deepseek-LLM。需要付费 API 访问的专有模型, 如 OpenAI 的 ChatGPT-4 和 Anthropic 的 Claude, 被排除在分析之外。由于硬件内存的限制, 研究无法评估所有模型配置的全谱, 包括不同的量化水平、指令嵌入以及扩展的参数规模。

为了增强确定性模型的输出并最小化随机变化, 每个模型的内部思维链推理通过系统级提示被抑制。所有大型语言模型 (LLMs) 和视觉语言模型 (VLMs) 的采样温度也被保持在低值 0.2。模型内置的对话历史记录也被禁用。相反, 每个评估试验接收到一个固定的历史记录, 其中包括了最近五个有效命令, 这些命令被明确作为用户提示提供。每个试验被分配最多 20 个推理步骤以完成其任务。每个组合运行总共 20 个回合。

LLM 的评估过程主要集中在它们通过生成语法正确的导航命令来遵循系统级指令的能力, 具体来说是“转向”或“移动”, 其幅度保持在预定义范围内。任何包含多余文本元素、符号、markdown 格式或标准 AI 免责声明的输出均被归类为无效。重要的是, LLM 的评估不基于角度计算或根据四旋翼飞机相对于目标坐标的当前位置和方向的移动距离估计的算术精度。

视觉语言模型的性能是通过它们对给定物体类别是否出现在相机视野中给出的二元“是”或“否”响应来衡量的。三类物体被用于验证: 人形机器人、无人机和四旋翼飞行器。未考虑目标检测的正确性。

评估结果的总结见于表格 I, 其中 Gemma LLM-VLM 配对的任务成功率最高, 为 40 %。结果还显示出语言模

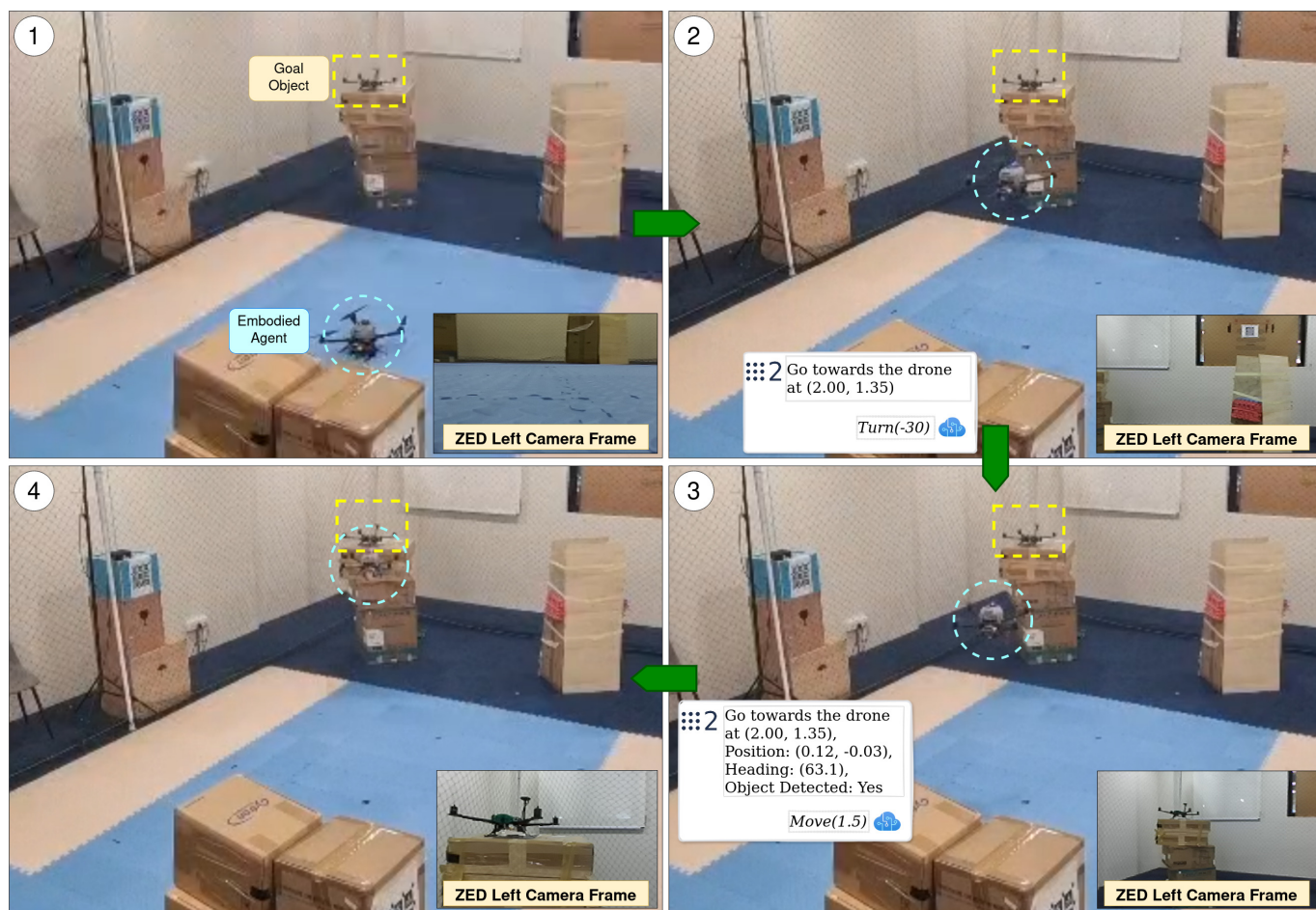


Fig. 4. 具身体代理在真实飞行中的演示。

型生成的有效导航命令比例与相应的任务完成率之间的强关联。这种关系在 DeepSeek 模型中尤为明显，该模型的命令有效率仅为 38 %，因此未能实现有意义的任务成功。值得注意的是，即便在生成了有效命令并准确检测到 100 % 物体存在的情况下，仍有一些任务失败是由于命令不理想，导致四轴飞行器低于或超过 0.5 米目标半径。另外，误报和漏报的物体检测也被认为是任务成功率低的原因之一。在三个采样的 VLM 模型中，Llava1.6 是唯一一个在实验过程中会产生超出“是”和“否”的响应的模型。然而，由于每个模型配对的评价都是基于相对较少的集数，Gemma, Qwen 和 Llama LLM 之间观察到的成功率的微小差异可能反映了随机变化，因此可能无法达到统计学意义。

在实验过程中进行的定性观察为所选模型的行为提供了额外的见解。首先，LLM 组件偶尔会连续重复相同的命令和类似的数值。这种行为可能是由于低温设置导致的，因为低温设置不鼓励探索。此外，DeepSeek 模型不断产生不稳定或基于推理的响应，因此即使系统级提示要求它只以“Move”和“Turn”这两种格式生成响应，它还是会产生无效命令。至于 VLM 组件，任务成功率与摄像机视口与目标物体之间有清晰无遮挡的视线似乎有很强的相关性。当在仿真重置时，目标被模拟障碍物遮挡时，成功率显著

降低。对于实际部署，只有 Gemma 被选为 LLM 和 VLM 组件，因为这种组合在整个仿真中表现出最高的任务成功率。此外，Gemma 有一种固有的倾向，输出的运动值比较保守，从而降低了碰撞的风险。图 4 展示了一个成功的序列，其中展示了通过生成的命令四轴飞行器执行的查询、响应和行动的互动过程。

当前研究有几个限制需要充分注意。首先，PX4 飞行控制系统在东北下（NED）坐标框架内运行。因此，需要显式转换为东北上（ENU）惯例，这是机器人系统中常用的，以确保智能体指令和机器人平台的向量正确对齐。其次，智能体仅限于发布前进和偏航速率指令，因此未能充分利用航空机器人平台通常具备的全向机动能力。此外，通过 ROS 传输的图像数据采用 RGB8 格式编码，这可能引入潜在差异，因为一些模型可能使用其他色彩通道格式（如 BGR8，首选于 OpenCV 库）的训练。在当前评估中，还缺乏对几项关键系统指标的定量分析，包括端到端延迟、语言模型的令牌使用和路径最优性，这些可以使用如 A\* 的最短路径算法来评估在完全可观察的地图上。此外，缺乏消融研究来检查对话历史或少样例对任务成功率的影响。虽然已经应用领域随机化来考虑模拟环境中的传感器和控制噪声，但没有合成成像伪影，如运动模糊或电子快门引起的曝光不规则性，这可能进一步解决从模拟到真实的差



TABLE I  
使用不同 LLM 和 VLM 的具身导航代理的评估

| LLM Model    | Parameter Size | LLM Valid Commands ( % ) | VLM Model       | Parameter Size | VLM Valid Detections ( % ) | Mission Success Rate ( % ) |
|--------------|----------------|--------------------------|-----------------|----------------|----------------------------|----------------------------|
| Gemma3       | 4B             | 100                      | Gemma3          | 12B            | 100                        | 40                         |
|              |                |                          | Llama3.2-Vision | 11B            | 100                        | 30                         |
|              |                |                          | Llava1.6        | 7B             | 98                         | 30                         |
| Qwen2.5      | 3B             | 100                      | Gemma3          | 12B            | 100                        | 30                         |
|              |                |                          | Llama3.2-Vision | 11B            | 100                        | 35                         |
|              |                |                          | Llava1.6        | 7B             | 97                         | 30                         |
| Llama-3.2    | 3B             | 100                      | Gemma3          | 12B            | 100                        | 30                         |
|              |                |                          | Llama3.2-Vision | 11B            | 100                        | 30                         |
|              |                |                          | Llava1.6        | 7B             | 98                         | 35                         |
| DeepSeek-LLM | 7B             | 38                       | Gemma3          | 12B            | 100                        | 0                          |
|              |                |                          | Llama3.2-Vision | 11B            | 100                        | 5                          |
|              |                |                          | Llava1.6        | 7B             | 98                         | 0                          |

距。

#### IV. 结论

为了通过自然语言和多模态视觉理解推动现实的无人机控制，这项工作提出了一个基于 Ollama 服务平台的代理框架，实现了针对特定资源和任务需求的开源 LLMs 和 VLMs 的灵活集成。虽然现实世界的部署仍然面临挑战，但领先的开源模型在场景理解和准确的自然语言命令执行方面表现出强大的潜力。引入专门的目标检测模块如 YOLO 或 DINO 可能进一步增强 VLM 的检测能力。这些能力对于工业应用如常规库存跟踪和设备检查尤其具有前景。未来的研究可能会集中在策划领域特定数据集，并应用如从人类反馈中进行强化学习（RLHF）和知识蒸馏等技术提高任务性能。此外，开发一个参数少于十亿的单端到端 VLA 模型提供了一条通向高效、完全边缘可部署系统的有前途的途径，减少了延迟和对云基础设施的最低依赖。