UAV Flight Log Generation and High-Level Mission Planning Using Vision-Language Models

TsungYen Yu
National Chengchi University

Abstract

本研究希望透過建立階段式的 pipeline 以結合視覺語言模型 (VLM) 與無人機控制，使無人機在執行任務時能夠同步進行影像資料解析。具體而言，無人機將持續分析其攝影鏡頭捕捉到的現場影像，評估如何執行使用者下達的高階抽象指令 (如「找到一座古老的寺廟」)，並判斷是否成功達成。同時無人機系統將採用類似航海日誌之記錄方式詳細紀錄飛行參數、現場觀測到的關鍵事件以及重要地標，並將經整理後的資訊有結構化的回傳給操作者以提供全方位的回饋與決策支持。在實驗中將以 AirSim (Unreal Engine + ROS2) 作為無人機模擬環境以降低實驗成本與縮短研發週期，由於 AirSim 模擬系統與真實世界的無人機輸入訊號相同且 Unreal Engine 可模擬真實世界環境，若能解決如何使 VLM 的資料傳遞更加快速、降低延遲，該技術便具有直接應用於真實無人機控制之潛力，從而推進無人機自動化的與智慧應用的未來發展。

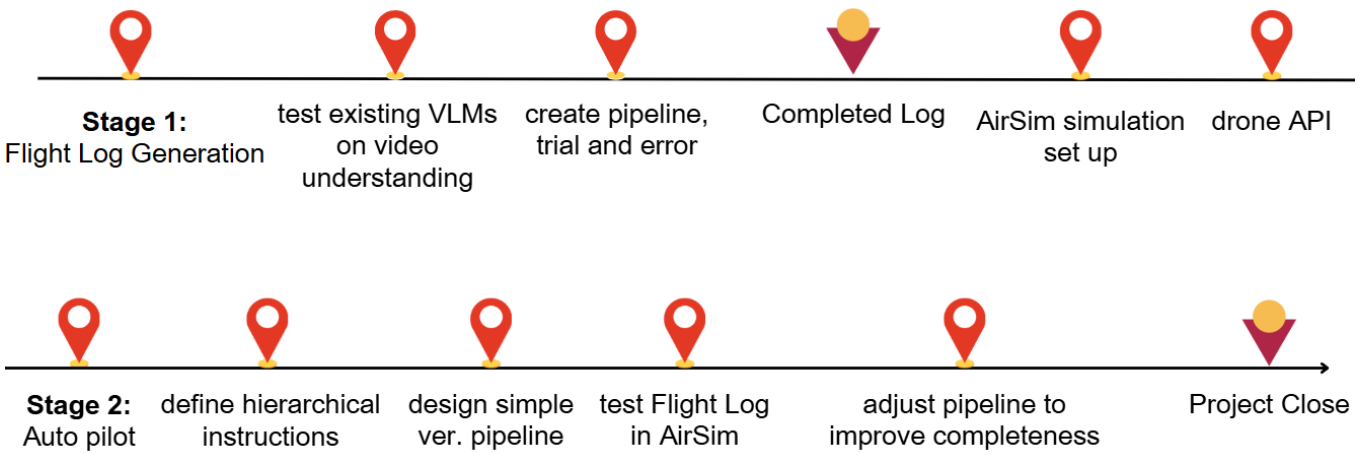
1. Introduction

隨著人工智慧在電腦視覺領域的快速崛起，卷積神經網路 (Convolutional Neural Network) 與更近期的 Vision Transformer 等各種強大的架構與技術正逐漸重塑自主系統 (autonomous system) 的樣貌。對於無人機 (UAV) 而言，這代表能從以往僅靠 GPS 的僵化路線飛行，進化到真正具備情境理解能力的自主飛行。與此同時，自然語言處理 (NLP) 領域也隨著大型語言模型 (LLM) 的興起迎來了自己的崛起，進一步解鎖了機器在處理和推理複雜人類語言方面的能力。這兩項原本獨立發展的前沿領域最終透過 CLIP 模型的誕生得以橋接，最終開啟了視覺語言模型（VLMs）的時代，提供一種新穎的互動介面，讓沒有相關機械操控經驗的使用者可以輕鬆透過白話文與無人機溝通，並讓其自主的執行任務。

目前也有許多嶄新的技術將目標放在 VLM + Robotics 的應用方面，如 VLA models；其透過 finetune pre-trained 的 VLM，使其可以 closed-loop 的控制機械，並直接輸出機械的控制指令。本研究則著重於運用 VLM 的高層次推理與多樣化能力，提出一套系統框架，融合上述領域，並於 AirSim 模擬環境中，達成人類與無人機之間的自然協作，同時在飛行結束後，透過 VLM 強大的語言處理能力，產出一篇有結構的飛行日誌，以供使用者快速統整飛行結果。

2. Methodology

下圖為整體研究的開發流程一覽:



2-1. Flight Log Generation

本研究的第一部分旨在建立一個有效的 pipeline 以處理任意無人機空拍影片，將其轉換成一篇具體、有結構的飛航日誌，以供飛航後的初步評估，以下是詳細步驟:

A. 選擇並測試 VLM

為了測試各大視覺語言模型針對無人機空拍影片(垂直俯視)的適應能力、空間理解的能力、時間連貫的推理能力與語意反饋的能力，我將連續 10 張間隔 3 秒的空拍照片輸入，並使其產生詳盡的總結。我一共測試了 4 個模型，分別為 LLaVA (13b)、ChatGpt 5、Claude-sonnet 4.5、Gemini 2.5，其大略能力如下圖:

- LLaVA (13b):
  - Run on local machine, fastest and cost-free
  - NLP backbone is too weak to reliably solve complex vision tasks
- ChatGpt:
  - Requires API key access
  - Richest contextual descriptions at a moderate speed
- Claude:
  - Requires API key access
  - Accurate contextual descriptions at a high speed
  - Accepts the longest context window
- Gemini:
  - Requires API key access
  - Accurate contextual descriptions at a moderate speed
  - Provides video understanding capabilities

因此，最後選用 ChatGpt 5 作為眼睛，Claude-sonnet 4.5 作為頭腦，Gemini 2.5 則透過原生的 video understanding 能力作為獨立的 Agent。

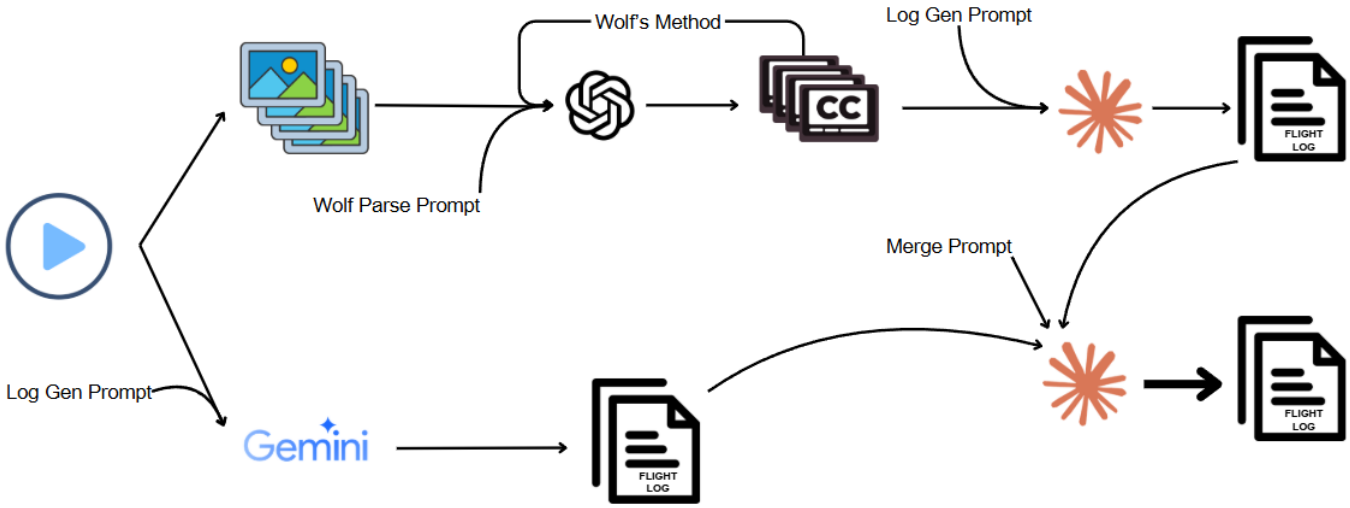
B. 基於 VLM 的能力，設計飛航日誌的格式

下圖為最後式選出的飛航日誌樣板:

1. Flight Identification
   - Date
   - Start Time
   - End Time
   - Total Duration
2. Flight Purpose
   - Purpose of Flight
   - Type of Operation
3. Environment Observation
   - Location Name/Description
   - GPS Coordinates (Takeoff)
   - GPS Coordinates (Landing)
   - Weather Conditions:
     - Wind Speed, Wind Direction, Visibility, Temperature, Cloud Cover, Precipitation
4. Flight Parameters
   - Maximum Altitude
   - Maximum Distance
   - Flight Pattern: ☐ Hover ☐ Linear ☐ Orbit ☐ Waypoint ☐ Free Flight
   - Key Waypoints/Locations
   - Flight Path Summary
5. Camera & Recording Settings
   - Video Resolution
   - Frame Rate
   - Recording Format
6. Safety Consideration
   - Obstacles Present
   - People in Area
   - Emergency Landing Sites
7. Incidents
   - Any Issues Encountered
   - Wildlife Interactions
   - Signal Loss Events
   - Weather Changes
   - Equipment Malfunctions
8. Notes & Lessons Learned
   - Flight Performance
   - Footage Quality
   - Areas for Improvement
   - Future Considerations

C. 設計將空拍影片轉換成飛航日誌的 pipeline

此 pipeline 共分為兩部分，第一部分是採用以 0.33 key FPS 的 Frame-by-Frame 方法，先將各 key frames 轉換成純文字的 captions，之後再將整份 captions 統整出一篇日誌；其中使用了 Wolf [1] 所提出的方法，將上一張 key frame 得出的 caption 也一起餵入 VLM，以保持時間連貫的推理能力。第二部分更為直接，直接使用 gemini 2.5 的 video understanding 能力來將空拍影片換成日誌。我透過此兩種截然不同的方法以相互彌補對方不足的地方，以下是實際的 pipeline 示意圖:



2-2. Autonomous Flight

本研究的第二部分旨在達成無人機的自主飛行，使用者只須下達高階的抽象命令，即可使無人機於 AirSim 模擬環境中自動化飛行，以下是詳細步驟:

A. 架設 AirSim 模擬環境，並設計無人機飛行之 API
為了簡化並減輕 VLM 的控制負擔，本研究限縮無人機之基本動作只可包含以下三種類型:
1. 向前飛 x
2. 原地旋轉 θ
3. 垂直飛 $\pm x$

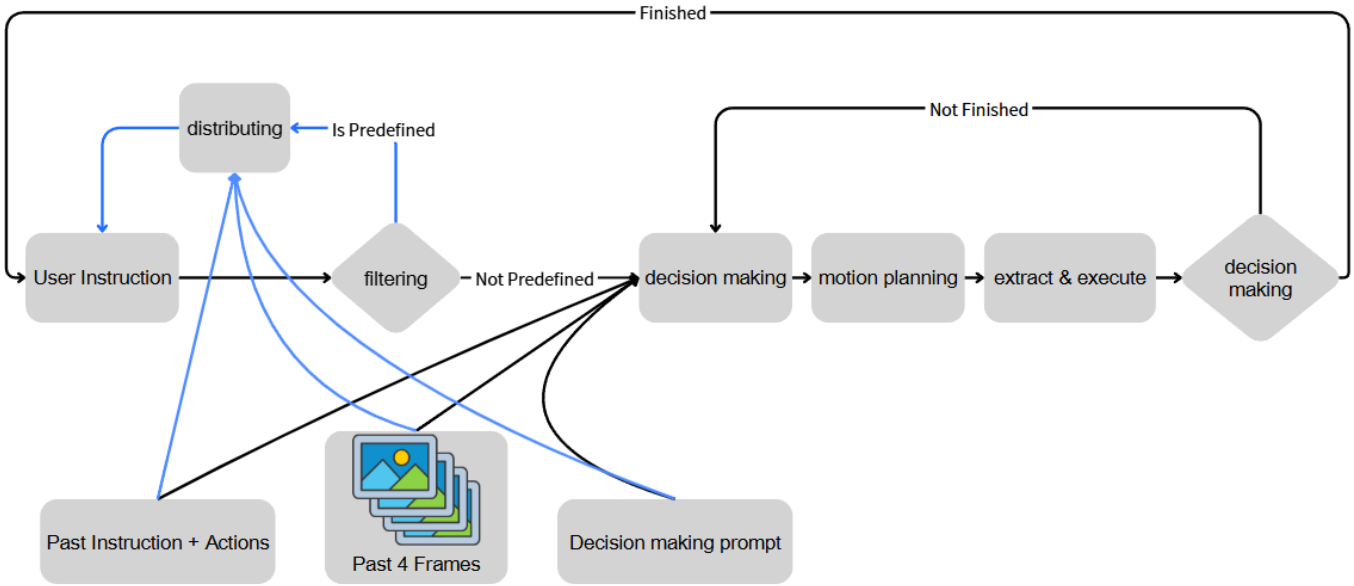
B. 將使用者的指令分成高、中和低階三種層級

	Name	Description	Example
High Level	(Vision) Task	Ambiguous, high-level goals without specific "how-to" instructions.	"Find me an old temple", "Fly until you see a roundabout"
High Level	Predefined Mission	Complicated but commonly used. Hard for VLMs to handle	"Return Flight", "Areal Scan"
Middle Level	Navigation	Gives a specific idea of "how" to fly	"Fly ahead for 100m and turn left", "fly in a square that has a length of 100m"
Low Level	Action	The most specific commands (APIs) that directly control the system	move_forward(x), rotate(x)

如上圖所示，越高階的指令越抽象、越缺乏如何飛行的指示

C. 設計一套可擴充且穩定的 pipeline，以達成無人機自主飛行
此 pipeline 的目的是把高、中階指令，全部轉譯為低階指令，以供無人機於 AirSim 中飛行。
在飛行完後會輸出兩個物件:
1. 一支 20 fps 的無人機空拍影片
2. 飛行過程中，所有使用者輸入的指令內容與其相對應轉譯出的低階指令 (Action)

下圖為整個 pipeline 的示意圖:



首先會經過一段 filter，判斷使用者輸入的指令是否為 Predefined Mission，若經 VLM 判斷後發現是，那便會將此指令傳到該 Predefined Mission 所定義之子 pipeline 做個別處理，若不然，就會往下經過兩階段的 decision making，第一階段會配合 motion planning 來判斷「如何」飛、第二階段則是判斷「是否」達成使用者的指令。

D. 設計各 Predefined Mission 所需之子 pipeline

此研究一共定義了三種 Predefined Mission (由於 pipeline 的設計，可輕易擴充)，分別為:

1. 返程
2. 地域掃描
3. 靜態物體追隨

為節省篇幅，詳細實作內容將透過此附錄 (supplementary slide P.33) 展示，其中詳細解釋了此三種任務的步驟。

3. Experiments and Results

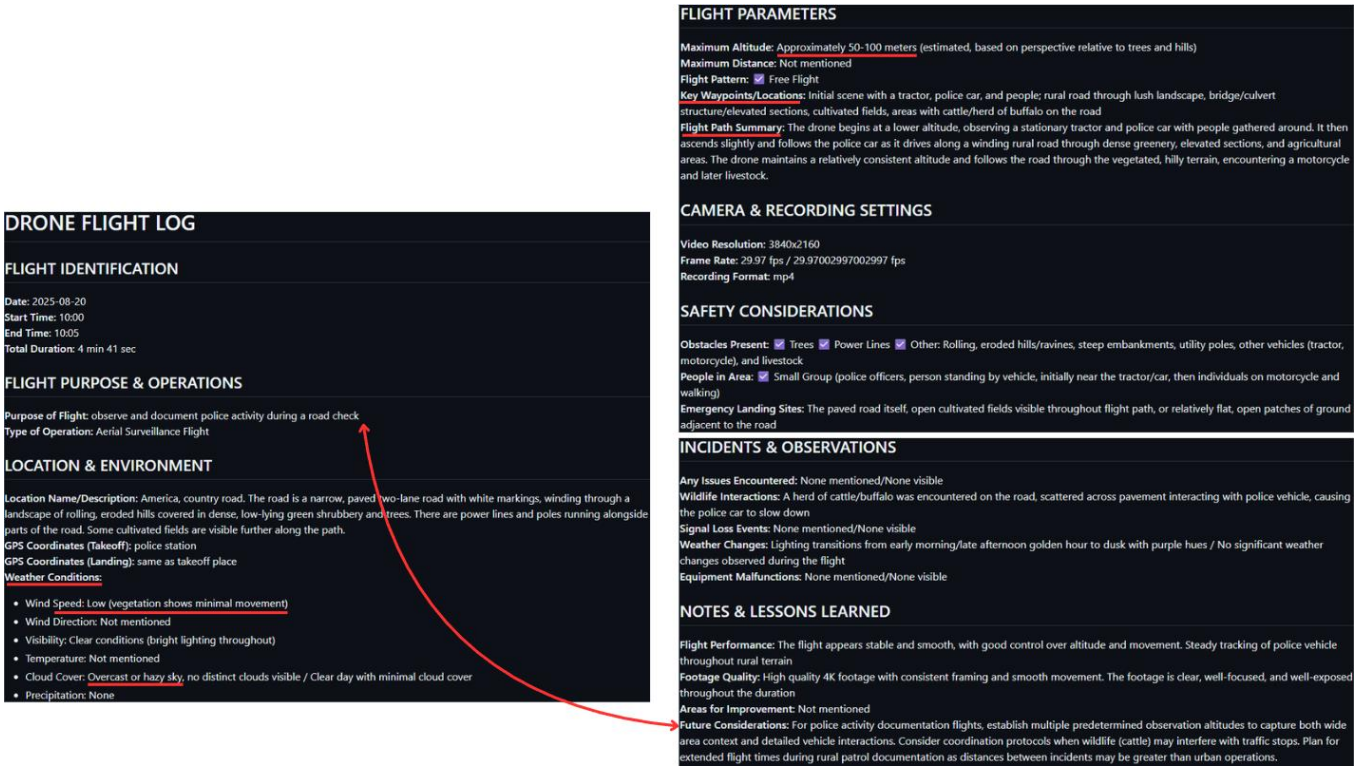
由於此研究之成果皆屬於篇幅較長的文字 (Flight Log) 與影片成果，因此在此處只略微帶過以示意，完整的成果可以透過我的 GitHub 輕鬆瀏覽。以下僅展示:

1. 真實世界空拍影片的飛航日誌產生
2. 中階指令 (Navigation) 執行
3. 高階指令 (Vision Task) 執行
4. 高階指令 (Predefined Mission) 執行

Demo 1: 真實世界空拍影片的飛航日誌 - Following a police car

以下是人工擷取之較具代表性的影像:

以下是產生出的飛航日誌:



DRONE FLIGHT LOG

FLIGHT IDENTIFICATION

Date: 2025-08-20
Start Time: 10:00
End Time: 10:05
Total Duration: 4 min 41 sec

FLIGHT PURPOSE & OPERATIONS

Purpose of Flight: observe and document police activity during a road check
Type of Operation: Aerial Surveillance Flight

LOCATION & ENVIRONMENT

Location Name/Description: America, country road. The road is a narrow, paved two-lane road with white markings, winding through a landscape of rolling, eroded hills covered in dense, low-lying green shrubbery and trees. There are power lines and poles running alongside parts of the road. Some cultivated fields are visible further along the path.
GPS Coordinates (Takeoff): police station
GPS Coordinates (Landing): same as takeoff place
Weather Conditions:

- Wind Speed: Low (vegetation shows minimal movement)
- Wind Direction: Not mentioned
- Visibility: Clear conditions (bright lighting throughout)
- Temperature: Not mentioned
- Cloud Cover: Overcast or hazy sky, no distinct clouds visible / Clear day with minimal cloud cover
- Precipitation: None

FLIGHT PARAMETERS

Maximum Altitude: Approximately 50-100 meters (estimated, based on perspective relative to trees and hills)
Maximum Distance: Not mentioned
Flight Pattern: ☑ Free Flight
Key Waypoints/Locations: Initial scene with a tractor, police car, and people; rural road through lush landscape, bridge/culvert structure/elevated sections, cultivated fields, areas with cattle/herd of buffalo on the road
Flight Path Summary: The drone begins at a lower altitude, observing a stationary tractor and police car with people gathered around. It then ascends slightly and follows the police car as it drives along a winding rural road through dense greenery, elevated sections, and agricultural areas. The drone maintains a relatively consistent altitude and follows the road through the vegetated, hilly terrain, encountering a motorcycle and later livestock.

CAMERA & RECORDING SETTINGS

Video Resolution: 3840x2160
Frame Rate: 29.97 fps / 29.97002997002997 fps
Recording Format: mp4

SAFETY CONSIDERATIONS

Obstacles Present: ☑ Trees ☑ Power Lines ☑ Other: Rolling, eroded hills/ravines, steep embankments, utility poles, other vehicles (tractor, motorcycle), and livestock
People in Area: ☑ Small Group (police officers, person standing by vehicle, initially near the tractor/car, then individuals on motorcycle and walking)
Emergency Landing Sites: The paved road itself, open cultivated fields visible throughout flight path, or relatively flat, open patches of ground adjacent to the road

INCIDENTS & OBSERVATIONS

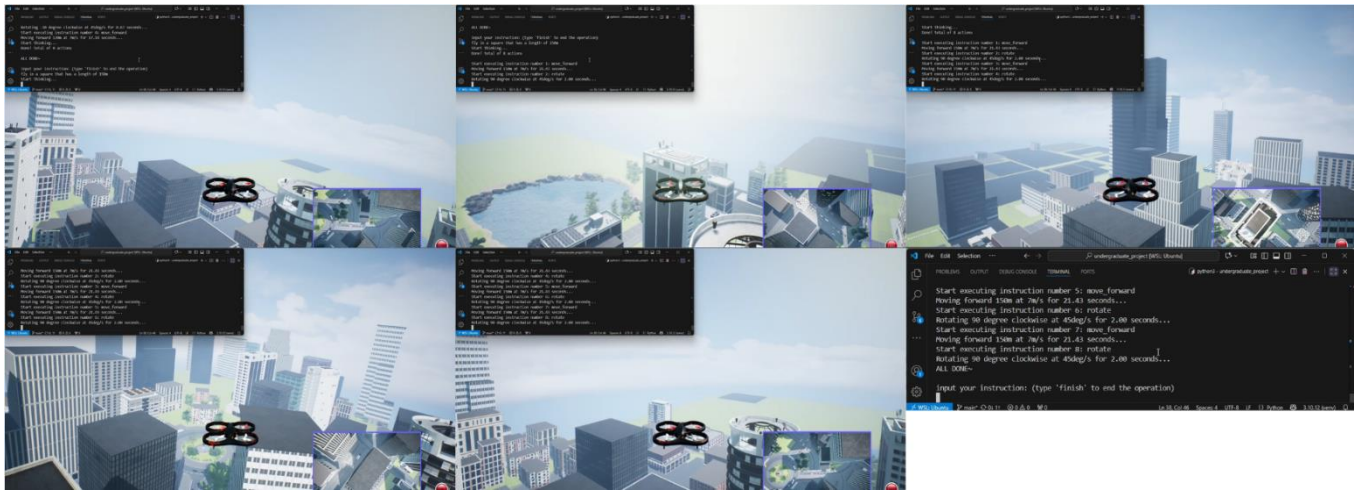
Any Issues Encountered: None mentioned/None visible
Wildlife Interactions: A herd of cattle/buffalo was encountered on the road, scattered across pavement interacting with police vehicle, causing the police car to slow down
Signal Loss Events: None mentioned/None visible
Weather Changes: Lighting transitions from early morning/late afternoon golden hour to dusk with purple hues / No significant weather changes observed during the flight
Equipment Malfunctions: None mentioned/None visible

NOTES & LESSONS LEARNED

Flight Performance: The flight appears stable and smooth, with good control over altitude and movement. Steady tracking of police vehicle throughout rural terrain
Footage Quality: High quality 4K footage with consistent framing and smooth movement. The footage is clear, well-focused, and well-exposed throughout the duration
Areas for Improvement: Not mentioned
Future Considerations: For police activity documentation flights, establish multiple predetermined observation altitudes to capture both wide area context and detailed vehicle interactions. Consider coordination protocols when wildlife (cattle) may interfere with traffic stops. Plan for extended flight times during rural patrol documentation as distances between incidents may be greater than urban operations.

雖然割草機並未被提及，但可以看出內容大致吻合，甚至可以非常粗略的推估高度。

Demo 2: 中階指令 (Navigation) 執行 – “fly in a square that has a length of 150m”



在示意圖中，左上方為向無人機下達指令的終端，右下角為無人機的相機實際捕捉到的空拍場景。可以看出無人機確實飛了一個正方形的路線後回到初始位置。

Demo 3: 高階指令 (Vision Task) 執行 – “find me an old temple”

可以看到無人機確實在找到廟後停止了。

Demo 4: 高階指令 (Predefined Mission) 執行 – “go home”

左上的是起飛位置，“go home” 是在右上的時刻輸入，可以看到於左下的時刻成功飛回原點。

4. Conclusion

本研究證明視覺語言模型（VLMs）成功建立了一種嶄新的雙向介面，連結人類與無人飛行載具，弭平了視覺感知與語言互動之間的鴻溝。我們的框架透過兩項核心能力，驗證了 VLMs 在機械控制的優勢與前景，其一為「飛行日誌（Flight Log）」生成方法，能提供一種簡單、人力負擔低的無人機影像評估方式，以幫助使用者快速了解飛行狀況；其二為「自主飛行（Autonomous Flight）」流程，能將使用者意圖轉化為具體行動。更重要的是，該流程具備高度彈性的架構，可在不更動核心架構的前提下輕鬆擴充其他子方法。而且，由於 AirSim 整合了 Unreal Engine 和 ROS2，其模擬環境與無人機輸入訊號皆高度擬真，一旦硬體需求滿足時，此架構即可直接接軌真實世界的應用。

然而，現有的 VLM 在處理視覺推理 (Visual Reasoning) 方面還稍嫌不足，本研究的其中一個 Predefined Mission，靜態物體追隨，便是受此限制；在判斷需矯正多少角度以持續跟隨河流、車道等人眼可輕易判斷之靜態物體時，時常偏離正確的航道，最後丟失目標。現階段的 VLM 雖然可以很清楚的知道這些靜態物體的存在，但是對於細部的掌控 (比如說車道的延伸角度、方向) 卻仍有很大的進步空間。

Reference

[1] Boyi Li et al. (2024) Wolf: Dense Video Captioning with a World Summarization Framework