Experiment Report — Vision-Language Web Application (LandMarkFinder)

Jerry Wang
CNIC
CAS
Beijing, China
jerry2003w@gmail.com

I. Introduction

This experiment aims to implement a vision-language web application powered by a large language model (LLM) capable of image understanding.

The project utilizes Ollama to locally serve a multimodal model (LLaVA) that can process both images and text.

The final goal is to build a simple yet practical web system that accepts an image input and automatically generates a descriptive text output related to the content of the image.

II. Application Scenario Description

The designed application, LandMarkFinder, focuses on landmark and architectural recognition.

In modern tourism and cultural education, users often capture photos of buildings or monuments but lack immediate contextual information.

This application addresses that problem by allowing users to upload an image of a landmark, after which the AI automatically identifies it and provides a bilingual (English – Chinese) travel guide introduction.

Practical value:

Assists tourists and students in learning about historical and cultural sites.

Can serve as an AI-powered tour guide assistant in travel apps.

Demonstrates how multimodal LLMs can interpret images beyond simple object recognition.

III. Implementation Process

A. Environment Setup

Installed Ollama on macOS.

Downloaded the LLaVA model via:

ollama pull llava

Verified the service was running at http://localhost:11434 through:

ollama serve

B. Web Application Development

Implemented a lightweight HTML + JavaScript front-end interface.

The interface allows users to upload an image and send it as a base64 string to Ollama's REST API.

Streaming responses are parsed and rendered dynamically in a styled output box.

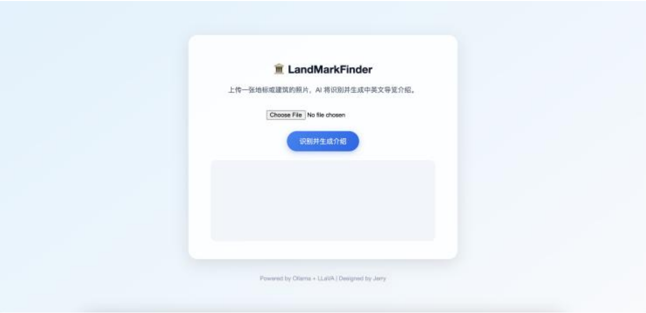
C. User workflow:

Open http://localhost:8000 via local HTTP server (python3 -m http.server 8000).

Upload an image of a landmark.

Wait for the AI to identify and describe it.

View the bilingual explanation generated by the model.



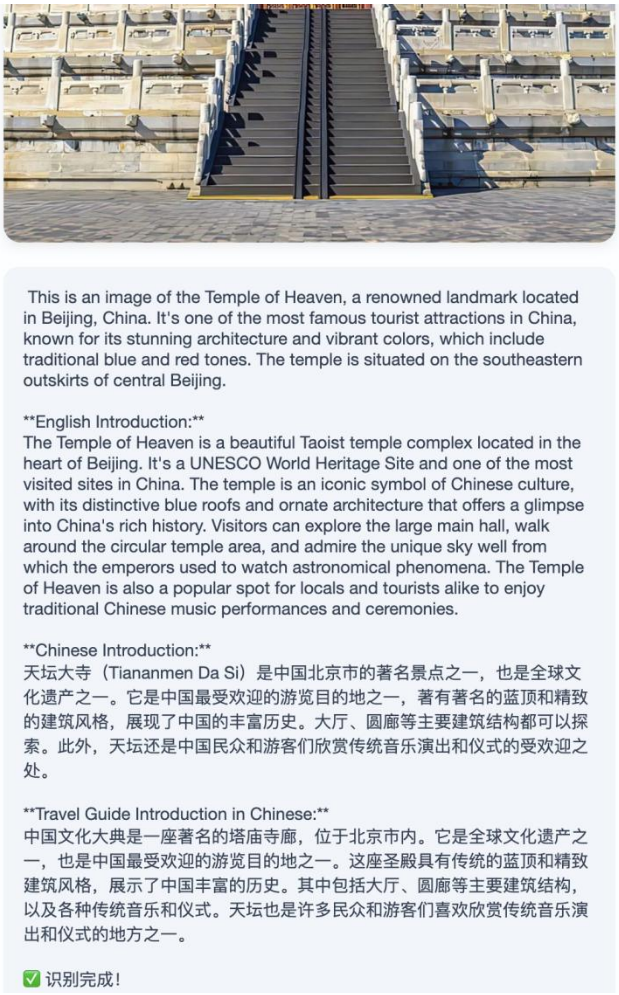
🏛 LandMarkFinder

上传一张地标或建筑的照片，AI 将识别并生成中英文导览介绍。

IV. Effect of the Web Application

When tested with real-world landmark images, the system performed well in identifying and describing them. For example, when an image of the Temple of Heaven (天坛) in Beijing was uploaded, the output included detailed English and Chinese introductions about its history, architecture, and cultural background.

The model generated coherent, contextually relevant paragraphs that demonstrated both visual understanding and linguistic reasoning.



This is an image of the Temple of Heaven, a renowned landmark located in Beijing, China. It's one of the most famous tourist attractions in China, known for its stunning architecture and vibrant colors, which include traditional blue and red tones. The temple is situated on the southeastern outskirts of central Beijing.

English Introduction:
The Temple of Heaven is a beautiful Taoist temple complex located in the heart of Beijing. It's a UNESCO World Heritage Site and one of the most visited sites in China. The temple is an iconic symbol of Chinese culture, with its distinctive blue roofs and ornate architecture that offers a glimpse into China's rich history. Visitors can explore the large main hall, walk around the circular temple area, and admire the unique sky well from which the emperors used to watch astronomical phenomena. The Temple of Heaven is also a popular spot for locals and tourists alike to enjoy traditional Chinese music performances and ceremonies.

Chinese Introduction:
天坛大寺（Tiananmen Da Si）是中国北京市的著名景点之一，也是全球文化遗产之一。它是中国最受欢迎的游览目的地之一，著有著名的蓝顶和精致的建筑风格，展现了中国的丰富历史。大厅、圆廊等主要建筑结构都可以探索。此外，天坛还是中国民众和游客们欣赏传统音乐演出和仪式的受欢迎之处。

Travel Guide Introduction in Chinese:
中国文化大典是一座著名的塔庙寺廊，位于北京市内。它是全球文化遗产之一，也是中国最受欢迎的游览目的地之一。这座圣殿具有传统的蓝顶和精致建筑风格，展示了中国丰富的历史。其中包括大厅、圆廊等主要建筑结构，以及各种传统音乐和仪式。天坛也是许多民众和游客们喜欢欣赏传统音乐演出和仪式的地方之一。

✅ 识别完成！

V. Problems Encountered and Solutions

Problem	Description	Solution
Port Conflict (Address already in use)	When running ollama serve, the default port 11434 was already occupied.	Checked running processes with lsof -i :11434 and confirmed an existing Ollama service; restarted or reused it.
Streaming Output Parsing	The /api/generate endpoint returns multiple JSON lines (one per chunk), causing unformatted or broken output.	Implemented line-by-line JSON parsing in JavaScript with JSON.parse(line) inside a while (true) reader loop.
Cross-Origin Access	Occasionally, the browser blocked local requests to localhost:11434.	Ensured the web page was also served from localhost, avoiding CORS issues.
UI Design	The initial page was visually plain and unprofessional.	Redesigned the page using modern CSS (glassmorphism, gradient backgrounds, rounded buttons, centered layout).

VI. Results and Analysis

Accuracy: The model correctly identified landmarks such as the Temple of Heaven and the Eiffel Tower.

Language Generation: The generated texts were contextually accurate and grammatically correct, often including cultural insights.

Responsiveness: The streaming approach allowed partial results to appear in real-time, improving user experience.

Overall, the system demonstrates that LLaVA (via Ollama) can effectively handle vision-language tasks locally, offering a strong base for educational or tourism-oriented AI applications.

VII. Conclusion

The experiment successfully implemented a practical vision-language web system capable of understanding images and generating bilingual descriptive text.
By combining Ollama's local LLM serving and a lightweight HTML frontend, this project demonstrates a feasible architecture for deploying multimodal AI systems locally without dependence on cloud APIs.

Future improvements:

Support for multiple images and comparison analysis.

Integration of voice narration for accessibility.

Expansion to recognize artworks, historical figures, and museum exhibits.

VIII. References

Ollama Documentation: https://ollama.ai

LLaVA: Large Language and Vision Assistant

W3C Fetch API Specification