

2025年臺灣國際科學展覽會 優勝作品專輯

作品編號

190015

參展科別

電腦科學與資訊工程

作品名稱

基於LLM的互動式口述影像系統

就讀學校

國立花蓮高級中學

指導教師

趙義雄

作者姓名

楊知叡

沈祐睿

黃堉軒

關鍵詞

大語言模型、多模態辨識、口述影像

作者簡介



我是沈祐睿。對於資訊科學有相當的興趣，隨著近幾年LLM的蓬勃發展，希望能讓LLM為我們的生活帶來更多便利。

我是楊知叡。我認為用LLM做一項分析影片產生敘述的研究是一件非常有趣的事情。希望以後能對LLM有更進一步的認識和應用。

我是黃堉軒。近年來LLM成為了世界主流，希望可以透過這次研究學習更多LLM特殊的使用方式。

2025 年臺灣國際科學展覽會
研究報告

區別：中區

科別：電腦科學與資訊工程

作品名稱：基於 LLM 的互動式口述影像系統

關鍵詞：大語言模型、多模態辨識、口述影像

編號：

摘要

本研究旨在透過 LLM 將影片內容轉為口述影像，探討及比較不同影片處理方式、LLM 對於圖片及影片的敘述，串接成一套自動化的口述影像系統。口述影像原本是為視障者製作的，現在本研究調整系統，讓有需求的一般大眾也能有效、容易及快速的了解影片內容。

本研究中的口述影像系統具有以下特點：

- （一）利用 LLM 擅於處理視覺訊息及自然語言的優勢，將影片分為多張圖片，由 LLM 分別生成敘述後再整合為影片的整體敘述。
- （二）運用 LLM 會留存對話紀錄的特性，使用者可在系統中針對疑問與 LLM 進行問答。
- （三）串接不同的 LLM，尋找製作口述影像的最佳組合。

藉由 LLM 將影片轉換為口述影像，實現互動式的口述影像服務。除了可以滿足視障者的觀影需求，更重要的是當一般民眾沒有時間觀看影片時，也能透過口述影像系統了解影片內容。

Abstract

This study aims to convert video content into video descriptions using Large Language Models (LLMs) and to explore various approaches to video processing and LLM-generated image and video descriptions, creating an automated video description system. Traditionally, video descriptions are designed for visually impaired individuals; however, this study will adapt the system for broader public use, making it easier and more efficient for anyone to understand video content. The video description system developed in this study utilizes the strengths of LLMs in processing visual information and natural language by segmenting videos into multiple frames, generating individual descriptions for each frame, and then consolidating them into a cohesive narration of the entire video. Additionally, by taking advantage of LLMs' ability to retain conversation history, users can ask follow-up questions and interact with the system for deeper clarification and detail. To identify the best solution for generating video descriptions, this study also explores the integration of different LLMs. By converting video content into video descriptions through LLMs, this study offers an interactive video description service that not only fulfills the needs of visually impaired audiences but also enables general users with limited time to quickly understand the content and details of videos.

壹、 研究動機

網路上各式各樣的電影、短影音、影片都是人們平常不可或缺的一部分，但是有些影片雖有聽覺成分，有些畫面卻是需要擁有視覺才可體會，而這無疑是視障者的一大不方便。視障者在生活中有各種需求，如需要給予聽覺與觸覺的提示外，觀賞影片也需要他人透過口述來體驗影片情節和內容。

但是一部長的影片，製作口述影像極可能需要 2 至 3 個月之久，我們隨即想到現代善於處理自然語言的 LLM、圖像理解技術已經興起並發展成熟，於是決定利用這些技術，整合並製作出一套自動產生口述影像的系統，除了讓視障者可以使用功能完善口述影像，也可以藉此減輕明眼人負擔。

聯合國提出的 SDGs（Sustainable Development Goals，永續發展目標）當中的第 10 項——消除不平等的目標 10.2 當中，即有提到融合身心有障礙者等較弱勢族群的社會經濟地位。因此本研究希望透過一套基於 LLM 的口述影像服務，來輔助視障者的觀影權利。

貳、 研究目的

- 一、 透過適當工具將原影片畫面擷取出多個適當畫面。
- 二、 透過適當評分方式評測 LLM 產生的結果。
- 三、 透過實驗找出適當工具獲取圖片四項要素。
- 四、 透過實驗尋找生成圖片敘述效果最好的多模態 LLM。
- 五、 透過實驗找出生成整合性描述效果最好的 LLM，以及生成整體敘述最好的方式。
- 六、 透過實驗將各文字敘述交由 LLM 生成整合性描述，並測試互動問答的效果。
- 七、 利用膠捲圖片進行影片整體敘述的產生以及問答任務。
- 八、 探討膠捲數量對於問答效果的影響。
- 九、 透過影片直接做問答，探討影片問答效果與口述影像系統問答效果是否可相比擬。
- 十、 將以上功能整合成自動化程式，實現自動化互動式口述影像服務。

參、 研究過程或方法

一、 研究設備與器材：

（一）、硬體：

1. 筆記型電腦 ASUS 12th Gen Intel(R) Core(TM) i5-1235U 1.30 GHz 8GB RAM。

2. 桌上型電腦 GIGABYTE 12th Gen Intel(R) Core(TM) i7-12700K 3.60 GHz
128GB RAM。

(二)、軟體：

1. Anaconda
2. VScode-Python (Python 3.12)
3. Pyscenedetect v0.6.4
4. Poe-API-wrapper v1.6.2
5. OpenCV
6. Llama3-taide-lx-8b-chat-alpha1

二、研究架構圖：

本實驗的研究架構圖如圖 2-1 所示。本研究設計的一連串的實驗，首先透過適當工具擷取影片畫面，再從畫面中提取四項要素。在透過實驗選定 LLM 代替人為評分後，將以不同 LLM 產生單張圖片敘述，接著以膠捲圖片、串接單張圖片敘述、串接四項要素+單張圖片敘述等三種方式產生影片整體敘述，並從評分中得知何 LLM 的效果最佳，或是以甚麼方式產生整體敘述最佳。

在整體敘述產生完畢後，本研究會進行問答實驗，透過不同方式進行問答，探討何種方式對於問答的效果最好，並進一步做問答相關的實驗，如膠捲數量是否影響、與影片問答的正確率與時間比較等等。本研究透過多個實驗，找出效果最佳的方式或 LLM，旨在實現最佳的口述影像服務。

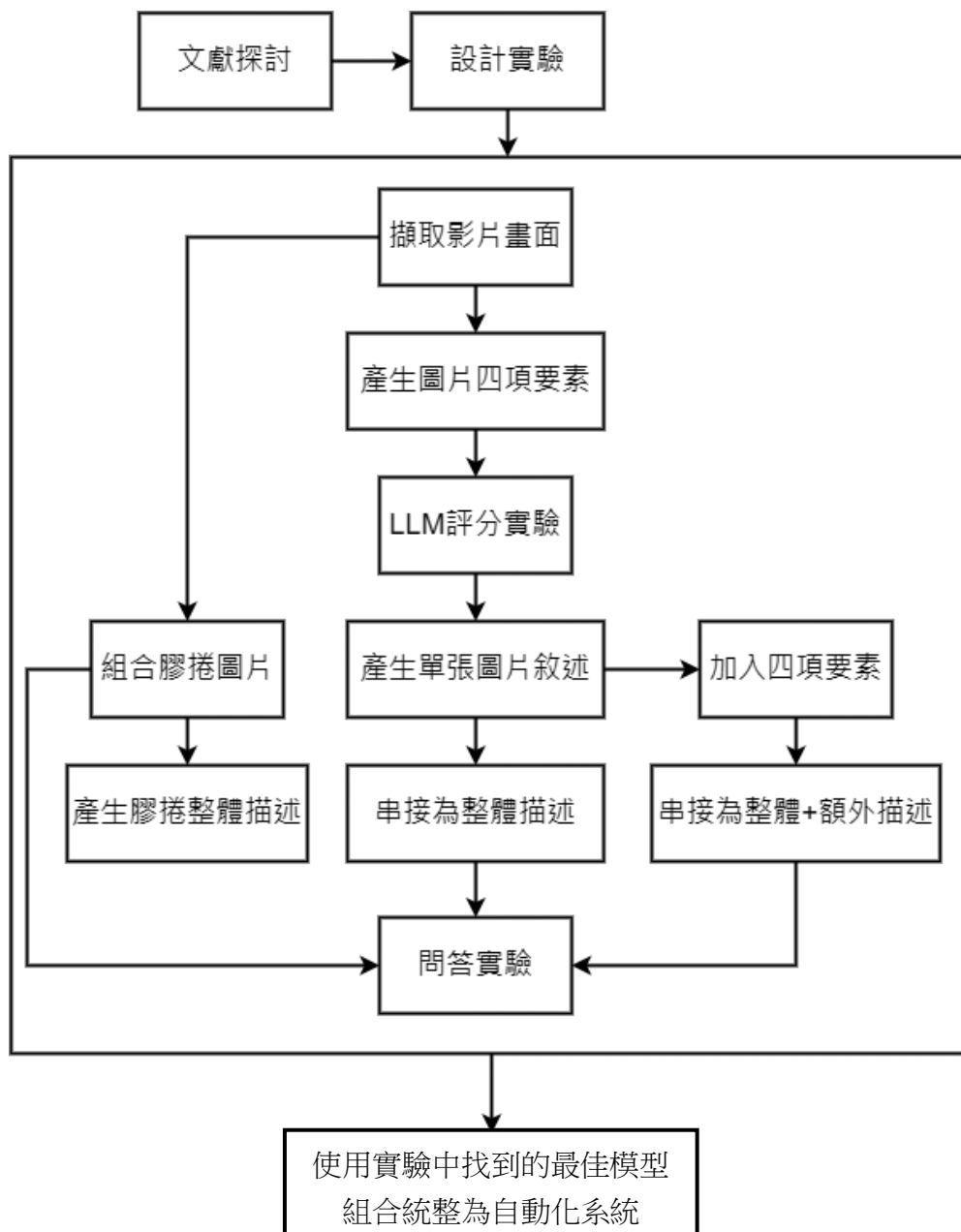


圖 2-1、研究架構圖

三、文獻探討：

(一)、口述影像、視障者的需求：

口述影像⁽³⁾是指不影響原影片聲音的情況下，對影片內容出現的視覺成分做出的敘述，通常需要明眼人來協助，讓視障者透過語言的方式理解影片當中的內容。

視障者是指視覺方面有障礙的人士，而視障者對於看影片的需求，根據淡江大學文學院趙雅麗教授所說⁽²⁾，視障者常常需要明眼人提供關於視覺上的訊息，而觀賞電影、影片等等也是一種敘述的過程，但是口述影像當中

除了「告知」，也要提供一套不同的論述手法，讓視障者有再看一遍的興趣。

有視障人士表示^[1]，因為觀賞電影時明眼人可以討論，但視障者卻無法融入，因此口述影像製作除了需要掌握影片當中的內容和事實，也要經過故事性的處理，讓視障者可以完整的了解故事的內容並感同身受。

另外，美國 Smith-Kettlewell 影片描述研究和開發中心（VDRDC）推出 YouDescribe 網站^[4]，透過有興趣的使用者紀錄一些 YouTube 影片的描述，YouDescribe 在幕後透過剪輯、分配這些敘述，並發布至網站上，讓視障者也有機會獲得觀賞影片的權利。YouDescribe 的 Q&A 當中也有提到，如果有些影片大多數是視覺元素，視障者就無法觀賞，這對於視障者來說是不公平的。

因此，本研究主要研究一套產生口述影像的自動化系統，讓視障者透過口述影像理解影片內容，從而保障視障者觀賞影片的權利。

（二）、擷取影片畫面：

明眼人透過觀看影片提供敘述的方式製作口述影像，現今技術可以讓 LLM 辨識圖片並產生圖片敘述，讓 LLM 直接透過影片獲取資訊的研究也逐漸成熟。本研究將會將影片分割成許多畫面，再將各畫面敘述連接成影片整體敘述的方式進行。本研究使用 Pyscenedetect 作為裁切影格的工具。

Pyscenedetect 目前版本為 0.6.4，提供了五種偵測方式^[5]。

1. **AdaptiveDetector**：計算相鄰畫面的滾動平均值，若超過自適應閾值則會觸發剪切。
2. **ContentDetector**：比較相鄰畫面之間的 HSV 差異，若超過設定的閾值則會觸發剪切。自適應檢測可視為內容偵測的加強版本。
3. **ThresholdDetector**：計算畫面平均像素亮度，若在設定的閾值範圍之外則觸發剪切。
4. **HistogramDetector**：計算相鄰幀的 YUV 顏色直方圖。若其直方圖差值超過設定的臨界值，則觸發剪切。
5. **HashDetector**：計算相鄰兩幀的哈希值（hash value），若相鄰幀之間哈希值的漢明距離（Hamming Distance）數值超過設定的臨界值，則觸發剪切。

由於 HashDetector 計算較為簡單，又可以確保相鄰畫面之間不會有太高重複性，因此本研究會選用 HashDetector 來執行裁切影格的任務。

（三）、LLM 結果評估方式：

LLM 生成的結果通常為較長的文字，因此很難評估其結果究竟是好是壞，因此本研究探討了幾種不同的評分標準：

1. BLEU 分數^[6]：比較選文與參考譯文的 n-gram 重合度，重合度愈高，代表翻譯的愈精確。BLEU 較適用於翻譯或比對兩文本相似度相關的任務。
2. Perplexity 分數^[7]：又稱困惑度，可表示一個文字後面加上合理的文字可能數量，Perplexity 分數數值越低，代表模型愈接近預測模型。Perplexity 分數較適用於評估一個語言模型對於自然語言的習慣程度，有時會在計算時做其他處理，以方便計算。
3. Tempcompass 評分方法^[8]：此資料集透過將影片依照「動作、速度、方向、屬性變化、事件先後」進行分類，相較於過去(2024/03/17 以前)的資料集，更能夠評估一個多模態模型對影片的了解程度。

本研究經過探討不同評分方式之後，由於 BLEU 和 perplexity 在判斷相似詞時可能會有分數較低的問題，並且適用的任務與本研究較無相關，因此本研究認為此兩種評分方式不適合。TempCompass 評分方法是針對模型對於視覺訊息的了解程度，因此本研究參考 Tempcompass 評分方法後，最終訂定一套評分標準，詳細內容將在實驗設計部分加以描述。

（四）、多模態語言模型：

在擷取影片當中畫面後，製作口述影像即需要一套將各個畫面內容轉為文字敘述的功能，而多模態語言模型即能達成此功能。本研究利用網站中提供的 LLM 資源，實現自動化系統使用了以下模型：

1. Claude 系列^[9]：

Claude 是由 Anthropic 推出的強大 AI 模型，並且也是很安全、值得信賴的模型，因此企業可使用 Claude 建構安全的應用程式。Claude 3 系列依規模大小可分為 Opus、Sonnet、Haiku，Opus 善於處理複雜的任務，Haiku 則擅長即時有效率的對答，Sonnet 則兼顧知識量和效率。未來 Anthropic 將推出 Claude 3.5 系列，模型能力將會更加強大。

本研究主要使用 Claude-3-Sonnet 以及 Haiku 為實驗對象，包含辨識圖片、敘述生成、互動問答等相關任務。

2. GPT4o 系列：

GPT4o^[10]是一種 OpenAI 推行的強大多模態 LLM，除了精於多國

語言、文本推理以及更進一步的含情緒聊天功能，處理多模態任務（如語音辨識、圖像處理等）表現也優於許多先進的 LLM 如 GPT-4、Claude-3-Opus，也是少數能處理影片的 LLMs 之一。其執行效率更遠勝於其他 LLM。最近推出的 GPT4o-mini^[11] 提供了費用降低，但是仍可以高效、廣泛且方便的處理各種任務，兩者皆是 OpenAI 目前最先進的 LLMs。

因此，本研究會使用 GPT4o 和 GPT4o-mini，除了處理生成圖片敘述相關任務，也會將其用於產生膠捲圖片敘述以及評分相關事宜。

3. Gemini 1.5 系列^[12]：

Gemini 1.5 Pro 和 Flash 是兩款由 Google 推出的多模態 LLM，擁有可以處理大量文字、圖片、程式碼的能力，並可以處理至多長達 1 小時的影片，兩者皆屬於少數可以處理影片的 LLMs。未來 Gemini 將會推出 Advanced 以及 Live，提供使用者更方便的使用 Gemini 模型。

本研究會利用 Poe API wrapper 串接 Gemini-1.5-Pro 和 Flash 做一般處理圖片和串接文字敘述等的任務，另外以 Google 的 Vertex AI studio 來使用 Gemini 1.5 Pro 和 Flash 多模態辨識功能，並測試其對於影片的理解能力以及執行時間，也會用於處理圖片敘述生成、產生影片整體敘述、評分相關事項。

（五）、大型語言模型：

大型語言模型（Large Language Model，以下簡稱 LLM）是基於大量資料預訓練出來的深度學習模型。本研究運用的 LLM 皆屬於 GPT（Generative Pre-trained Transformer，基於變換器的生成式預訓練模型），為一種 Transformer 模型，擁有自注意力機制，處理效率高，並且經過大量資料進行預訓練，擅長自然語言處理。

本研究將利用 LLM 的特性進行口述影像的整合。本研究會使用到的 LLM 比較表如表 3-1 所示。

表 3-1、LLM 比較表^[13]

LLM	開發者	Token 數量	價格 (USD)	品質	速度
Claude-3-Haiku	Anthropic	200K	0.5M	2	3
Claude-3-Sonnet	Anthropic	200K	6M	2	2
GPT4o	OpenAI	128K	7.5M	3	2
Gemini-1.5-Pro	Google	1M	5.25M	3	2
Gemini-1.5-Flash	Google	1M	0.53M	2	3

GPT-3.5-turbo	OpenAI	16K	0.75M	1	2
TAIDE	臺灣 國研院	8192	免費	2	依硬體 而定

1. 網路 LLM 資源：

由 Quora 平台開發的 Poe（全名為 Platform for Open Exploration）是一種搭載多個語言模型的平台，在 Poe 當中有提供 ChatGPT、Claude Sonnet、Claude Haiku、Gemini 1.5 Pro、GPT4o 等模型服務，考量有些官方網站的網站有問答限制，本研究會利用 Poe 當中的一些服務作替代，並利用 Poe API wrapper 進行自動串接工作。

2. TAIDE^[14]：

TAIDE 基於 meta 發行的語言模型 llama3-8B，另外多加了繁體中文的 token 訓練，因此 TAIDE 善於處理不同領域的繁體中文相關任務，如寫信、寫文章、自動摘要等等。TAIDE 對於台灣的文化、語言相關知識也相當了解，並具有多輪問答的能力。

由於本研究預計實現的口述影像服務是以繁體中文為主，本研究將利用 TAIDE 做影片整體敘述的串接以及互動問答工作，應用 TAIDE 的繁體中文能力以及台灣文化背景製作優良的口述影像。

（六）、串接語言模型：

本研究會使用不同的 LLM 進行實驗，但是本研究考量硬體設備不足，無法順暢執行 LLM，因此需要一些其他方式呼叫。以下為串接 LLM 的工具。

1. API key^{[15][16]}：

API（應用程式介面）是指一個軟體的仲介，可以讓兩軟體之間互動，API key 則是負責驗證和授權應用程式。API key 通常由一串字元或數字組成，並且在兩端只有唯一一組 API key。利用不同系統的 API key 可以連接到語言模型、應用程式等等，API key 的擁有者也可以透過 API key 來監管應用程式的使用紀錄、呼叫次數等等。

2. Poe-API-wrapper^{[17][18]}：

Poe 有提供免費的 API 接口，本研究將會透過此接口連接 Poe 當中的 LLM 服務，包含 Claude-3-Haiku、Claude-3-Sonnet、GPT-3.5-turbo（ChatGPT）、Gemini-1.5-Pro、GPT4o 等服務。

可架設於本地端的語言模型正在進步中，但是由於直接架在本地端的語言模型目前仍有硬體設備記憶體不足、處理速度慢、處理任務限制等問題，因此本研究將使用 Poe API wrapper 進行自動串接實驗，便可

以較為方便的使用較先進的 LLM 服務。

3. Google Vertex AI studio [19]：

Vertex AI studio 為一個基於瀏覽器的整合式開發環境。Vertex AI studio 的主要目的即為讓使用者探索生成模型的原型。可以輕鬆的測試不同的提示字 (Prompt)，並且可引入 Google Drive 當中的資料，支援圖片以及影片資料處理。

本研究會以 Vertex AI studio 使用 Gemini-1.5-Pro 及 Flash 來測試基於影片的問答以及執行時間的探討。

4. TAIDE 的串接方式：

本研究以 Ollama、Langchain-community 模組串接 llama3-taide-lx-8b-chat-alpha1 模型。

Langchain-community 為一個 python 套件，Langchain 自身已有一些基本的功能，而 Langchain-community 則提供第三方程式碼的整合，讓 Langchain 功能更加強大。Ollama 則是一個開源軟體，使用者可以 Ollama 當中輕鬆探索不同語言模型，亦可以 Langchain 串接 Ollama 使用，本研究即是如此串接 TAIDE 模型。

四、實驗設計：

本研究的實驗使用的 LLM 皆為自動串接，並且實驗內容環環相扣，一個實驗的結果可能會應用在不同的實驗上。實驗的工作、使用工具、評分、以及實驗細項如表 3-2 所示。

表 3-2、實驗評估基準點及使用工具和實驗說明

實驗編號	評估基準點	實驗工具	實驗說明
1	人為辨識	Pyscenedetect HashDetector	為求得相鄰畫面重複性不高的結果，本實驗以人為辨識以及 Pyscenedetect 中的 HashDetector 辨識進行比較，比較辨識出來影格數量相似度，並調整臨界值，比較最佳的臨界值數值。
2-1	人為評估	多模態 LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o GPT4o-mini Gemini-1.5-Pro	本研究參考 TempCompass 評分方法後，從圖片中提取四項要素： 1. 圖片主要物體（人物） 2. 主要物體（人物）的動作或狀態 3. 背景資訊

		Gemini-1.5-Flash	4. 圖片細節（包含畫面中文字資訊、圖片次要物體等較不顯眼成分） 透過人為和 LLM 辨識單張圖片四項要素，以人為方式比較兩者差異，找出結果與人為辨識最相近的 LLM，則可以以該 LLM 作為辨識四項要素的工具。
2-2	人為評估	多模態 LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash	本實驗透過人為和 LLM 辨識膠捲圖片四項要素，以人為方式比較兩者差異，找出結果與人為辨識最相近的 LLM，則可以以該 LLM 作為辨識四項要素的工具。
3	採用實驗 2-1 結果作為基準點，再以人為觀點評估	LLM 評分： （採用實驗 2-1 結果做基準點） Claude-3-Haiku Claude-3-Sonnet GPT4o Gemini-1.5-Pro GPT-3.5-turbo TAIDE	本實驗將提取出的四項要素分配權重： 1. 圖片主要物體（人物）：30% 2. 主要物體（人物）的動作或狀態：30% 3. 背景資訊：20% 4. 圖片細節（包含畫面中文字資訊、圖片次要物體等較不顯眼成分）：20% 透過含權重的四項要素以及圖片敘述，以人為和 LLM 計算加權平均分數（以下稱為準確分數），再以人為比較兩者評分差異，找出結果與人為評分最相近的 LLM，則可以以該 LLM 作為評分工具。
4	採用實驗 2-1 結果作為基準點，採用實驗 3	多模態 LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o	本實驗透過不同多模態 LLM 產生單張圖片敘述，以實驗 3-1 選擇的 LLM 計算其準確分數（在此要搭配實驗 2 產生的四項要素），找出

	效果最佳的 LLM 評估	GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash	準確分數最高的 LLM，則可以以該 LLM 執行生成單張圖片敘述的任務。
5-1	採用實驗 2-2 結果作為基準點，採用實驗 3 效果最佳的 LLM 評估	多模態 LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash	本實驗透過不同多模態 LLM 辨識以 opencv 模組橫向拼接成的膠捲圖片，產生該整體敘述，以膠捲圖片直接辨識四項要素，以實驗 3-1 選擇的 LLM 計算其準確分數，找出準確分數最高的 LLM，則可以以該 LLM 執行透過膠捲圖片生成影片整體敘述的任務。
5-2	採用實驗 2-2 結果作為基準點，採用實驗 3 效果最佳的 LLM 評估	LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash GPT-3.5-turbo TAIDE	本實驗透過實驗 4 結果產生的單張圖片描述，以不同 LLM 串接成影片整體敘述，利用實驗 5-1 辨識出的膠捲圖片四項要素，以實驗 3-1 選擇的 LLM 計算準確分數，找出準確分數最高的 LLM，則可以以該 LLM 執行串接單張圖片描述生成影片整體敘述的任務。
5-3	採用實驗 2-2 結果作為基準點，採用實驗 3 效果最佳的 LLM 評估	LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash GPT-3.5-turbo TAIDE	本實驗透過實驗 4 結果產生的單張圖片描述以及實驗 2 產生的圖片四項要素合成含額外資訊的單張圖片描述，再以不同 LLM 串接成影片整體敘述，利用實驗 5-1 辨識出的膠捲圖片四項要素，以實驗 3-1 選擇的 LLM 計算準確分數，找出準確分數最高的 LLM，另外此實驗將和實驗 5-1、5-2 做比較，找出產生影片整體敘述的最佳方式。
6	人為評分	LLM： Claude-3-Haiku Claude-3-Sonnet GPT4o	本實驗將與 LLM 產生的整體敘述做問答，並計算問答正確率，找出最適合問答任務的 LLM，以及最適合的問答方式。

		GPT4o-mini Gemini-1.5-Pro Gemini-1.5-Flash GPT-3.5-turbo TAIDE	
7	人為評分	實驗 6 決定膠捲圖片問答 LLM	本研究將控制膠捲圖片的數量，並進行該片段的問答，以實驗 3-2 選擇的 LLM 計算問答分數，找出最佳的膠捲數量。
8	人為評分	Gemini-1.5-Pro Gemini-1.5-Flash	本實驗透過影片直接問答，紀錄執行時間和問答正確率。

肆、 實驗結果

一、 本研究實驗影片來源如表 4-1 所示：

表 4-1、影片編號、來源及性質整理

編號	影片名稱	影片來源	影片性質
1	The Most AMAZING Vanilla Cake Recipe	Youdescribe	烘焙類影片
2	Bailey Fires Qadri – Grey’s Anatomy		劇情影集
3	Cale Makar Nails Amazing Empty-Netter Goal From Own End		運動類影片
4	Seoul’s streets rumble with military hardware as South Korea stages rare parade		新聞類影片
5	Shelter dogs LOVE sand!		動物類影片
6	校園建築物外觀拍攝	自行錄製	景觀類影片
7	0729 凱米颱風災情		寫實類影片

二、 經由實驗結果比較人為與 Pyscenedetect 的擷取畫面效果：

- (一)、 說明：透過人為方式辨識影片的影格數量，與 Pyscenedetect HashDetector 做比對，判斷不同臨界值時和人為的誤差多少，從而選擇最適合臨界值。
- (二)、 評分方式：紀錄人為和 Pyscenedetect 擷取影格數量，並做比較。結果如圖 4-1 所示。詳細實驗數據詳見附錄 1。

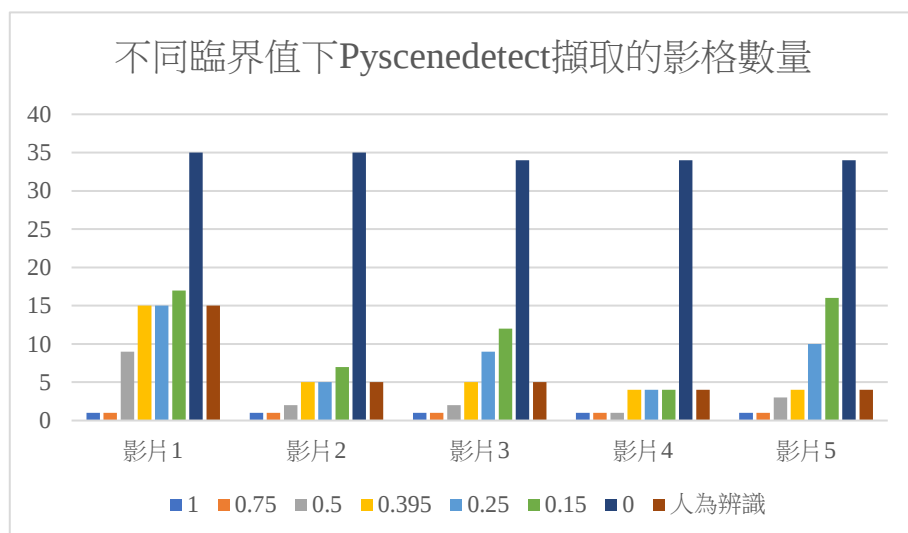


圖 4-2、不同臨界值下 Pyscenedetect 擷取的影格數量與人為比較

由結果可知，臨界值設為 0.395 的效果最接近人為辨識。因此本研究在使用 HashDetector 時，即會以此臨界值為主做後續的實驗。此實驗也間接印證 Pyscenedetect HashDetector 的預設臨界值為何是 0.395。

三、經由實驗結果比較人為與不同 LLM 辨識四項要素效果：

- (一)、說明：以人為與 LLM 辨識四項要素，再以人為辨識四項要素為基準點，以人為方式判斷 LLM 辨識四項要素符合程度。
- (二)、評分方式：透過人為比較相似度，將四項細節當中的內容給予 0~2 分，以 LLM 的任一個項目為例，

出現人為敘述有提及並且描述清晰得 2 分，出現人為辨識有提及的元素但不清楚者得 1 分，未出現人為辨識提及的內容的得 0 分。最後將所有部分分數平均成為一項要素的分數，將四項要素的分數作加權平均，成為整體分數。範例如圖 4-3-1 所示，人為計算結果如圖 4-3-2 所示。

提示字：

請用繁體中文說明這張圖片的主要物件、其動作或狀態、圖片的背景及圖片的其他細節，並用

物件：

動作或狀態：

背景：

細節：

的方式呈現，若一項有多個內容請以逗號方式隔開。

若非主要物件可歸類在細節。

除了上述四項，請不要回答其他內容。

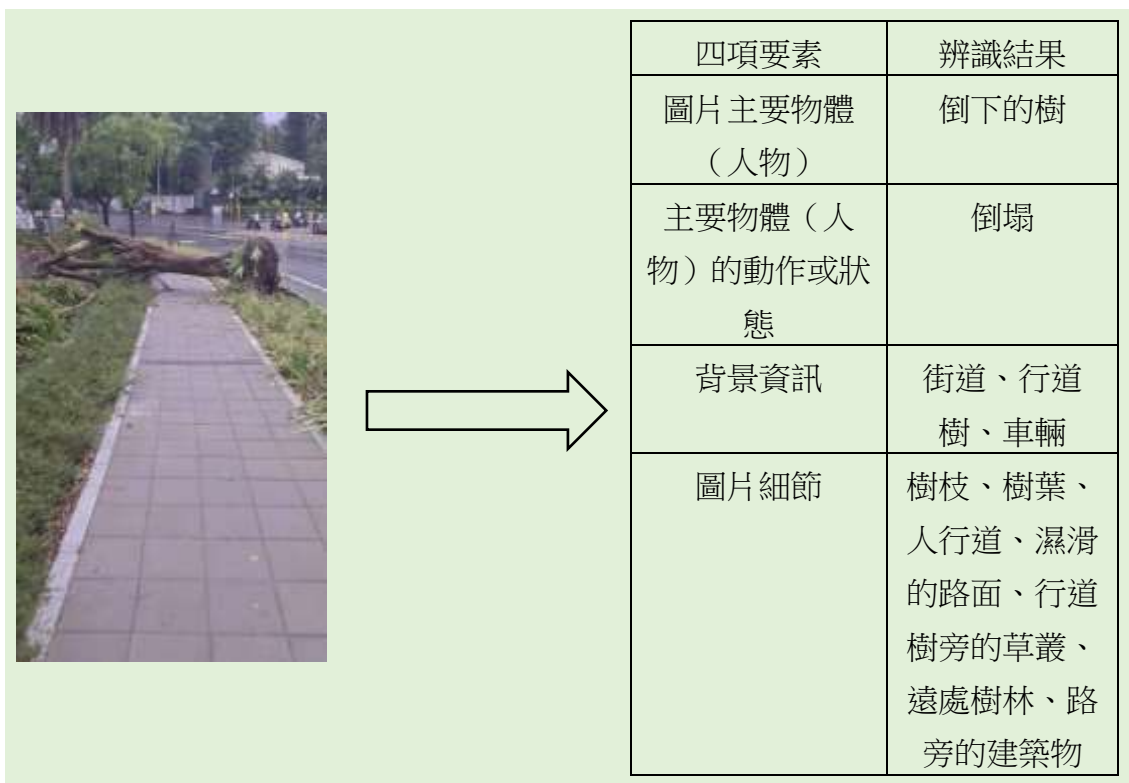


圖 4-3-1、產生單張圖片四項要素範例

完整範例以及詳細實驗數據詳見附錄 2。本實驗每個影片將挑選三張圖片做四項要素分數計算，圖 4-3-2 結果為三張圖片的平均值。

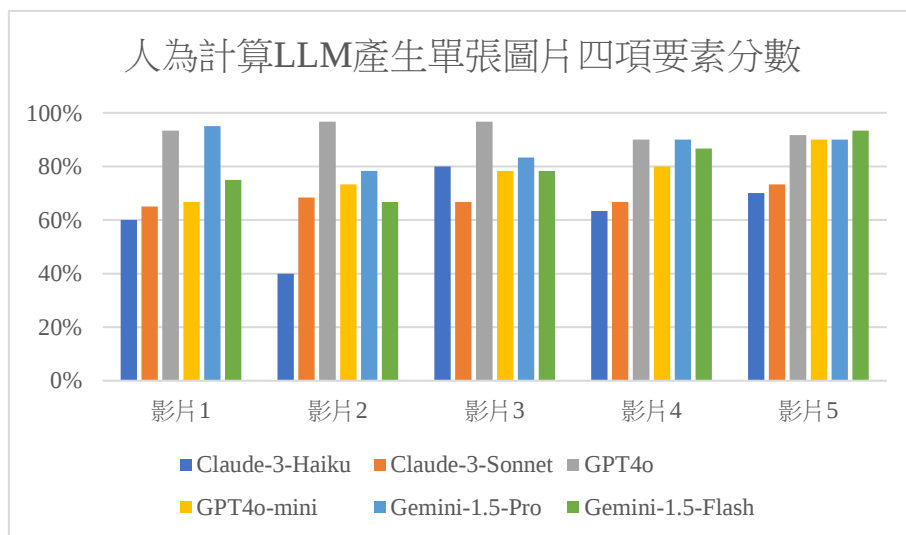


圖 4-3-2、人為計算 LLM 產生單張圖片四項要素分數

由結果可知，GPT4o 對於生成圖片的四項要素效果是最好的，本研究將會使用 GPT4o 進行單張圖片四項要素的生成。

若以膠捲圖片產生四項要素，範例如圖 4-3-3 所示：

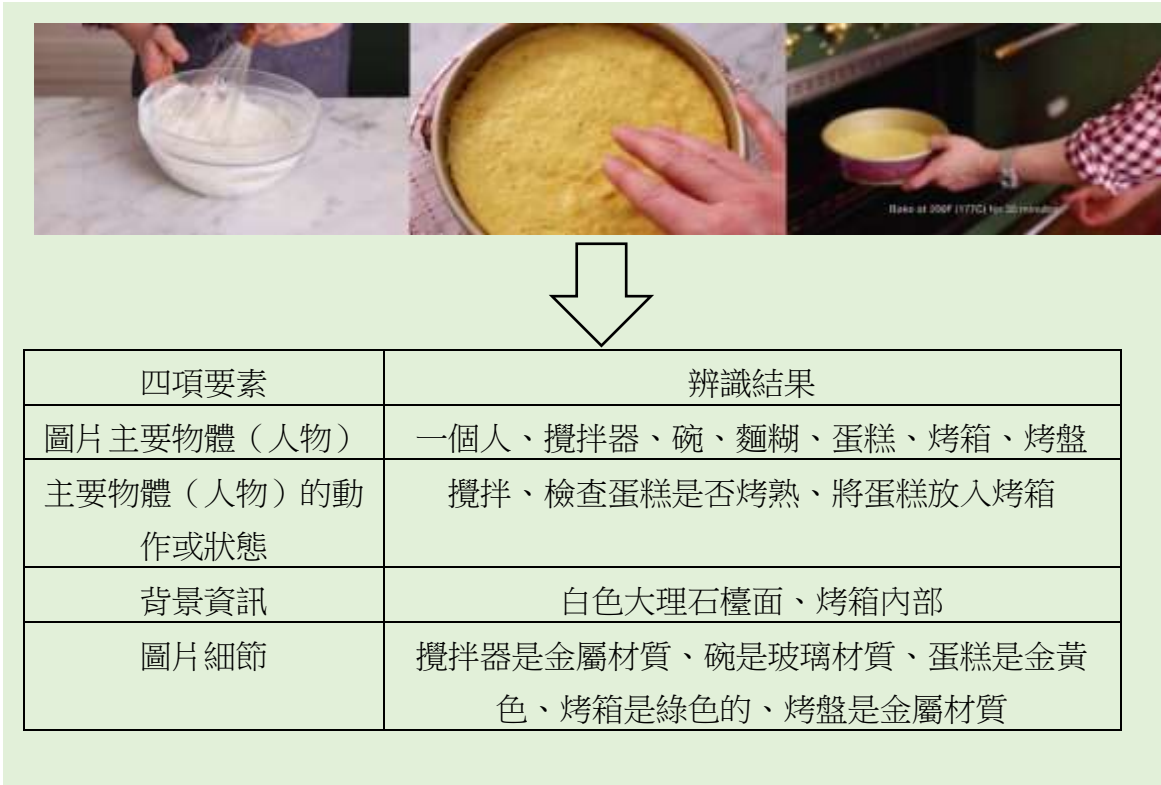


圖 4-3-3、產生單張圖片四項要素範例
完整範例以及詳細實驗數據詳見附錄 3。

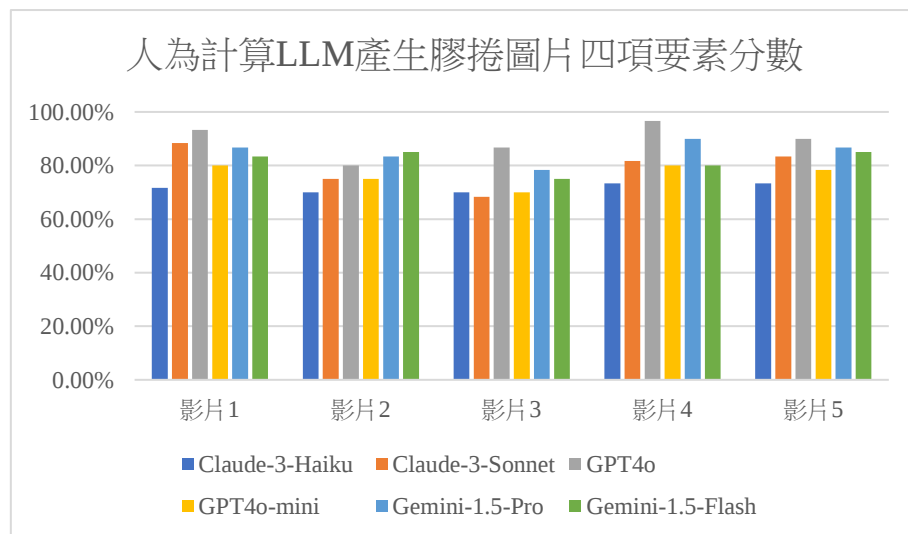


圖 4-3-4、人為計算 LLM 產生膠捲圖片四項要素分數
由結果可知，GPT4o 對於生成圖片的四項要素效果是最好的，本研究將會使用 GPT4o 進行膠捲圖片四項要素的生成。

四、經由實驗結果比較人為與不同 LLM 評分效果：

- (一)、說明：以實驗 2-1 效果最佳的 LLM 產生四項要素作為基準點，在以人為和 LLM 評斷一個圖片敘述的分數，最後以人為評分為基準，計算

LLM 評分與人為評分的誤差絕對值。

本研究計算誤差可分為一般誤差平均以及絕對值誤差平均，絕對值誤差平均代表每次分數皆與人為相減並取絕對值，計算平均所得結果；一般誤差平均則為每部影片 LLM 平均分數與人為平均分數相減並取絕對值，計算各影片平均值結果。

(二)、 評分方式：比較 LLM 評分與人為評分的差值，比較以人為為基準，LLM 評分的誤差值平均。

(三)、 準確分數計算方式：四項要素中一項要素若完全符合且描述清楚，得 2 分；若有提到但描述不明顯者，得 1 分；若未提及或敘述不清楚至無法辨認者，得 0 分。若一項當中有多個內容，則分別評分並做平均。將四項要素的分數皆算出來以後，再做加權平均，得到的分數即為準確分數。範例如圖 4-4-1 所示。

提示字：

請依照下面的基準點：

(GPT4o 辨識出四項要素內容)

依照以下權重：物件 30%、動作或狀態 30%、背景 20%、細節 20%，將以下敘述依照基準點給予 0~2 分，若有一項基準點當中有多個內容，請分別評分並平均後，再做四個基準點的加權平均。

請用「基準點：平均分數」的方式列點呈現每個基準點的分數，並在最後寫明最後加權平均分數，不要回答其他內容。

請用繁體中文回答。

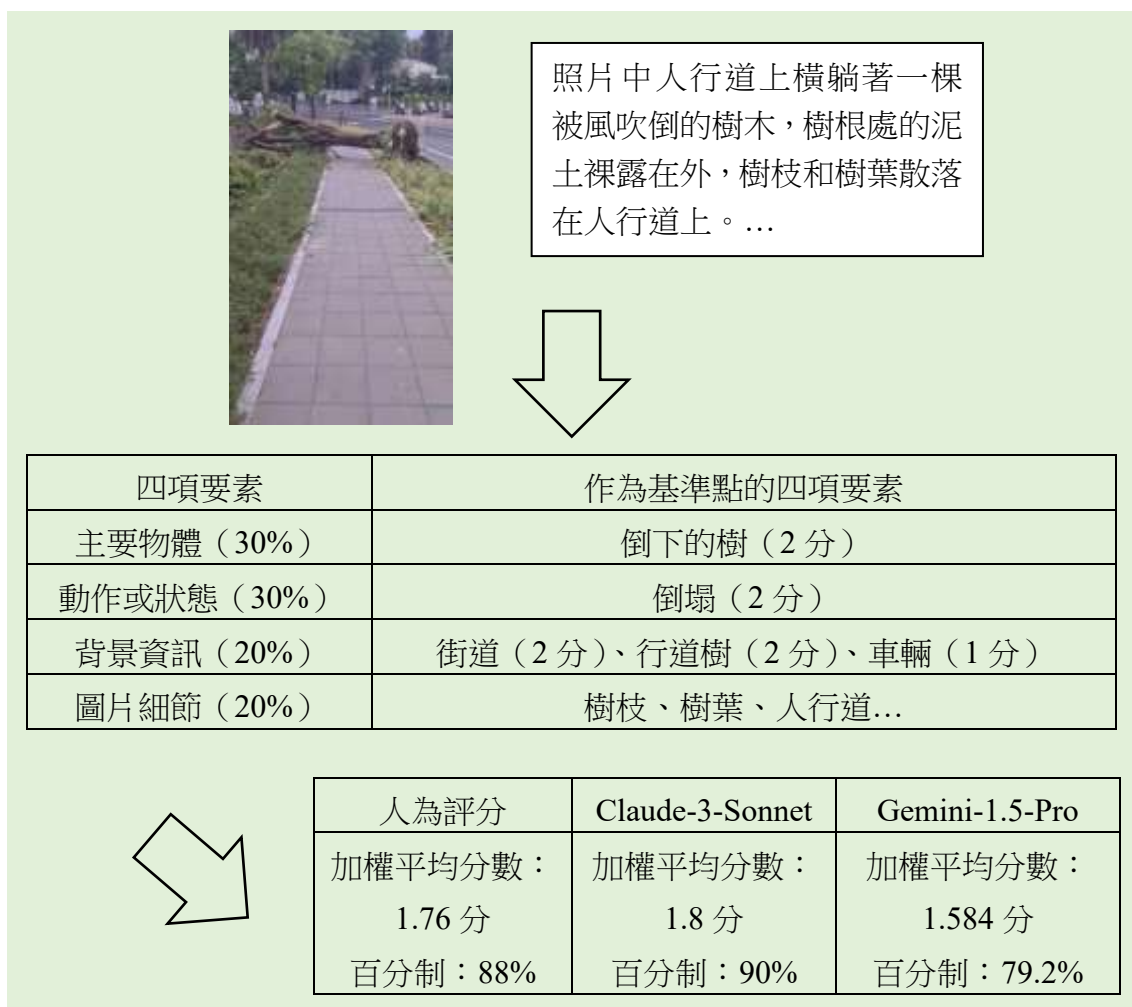


圖 4-4-1、準確分數計算範例

完整範例以及詳細實驗數據詳見附錄 4。本實驗一部影片使用三張不同圖片做評分，圖 4-4-2 中每部影片的結果為三張圖片分數平均後的結果。

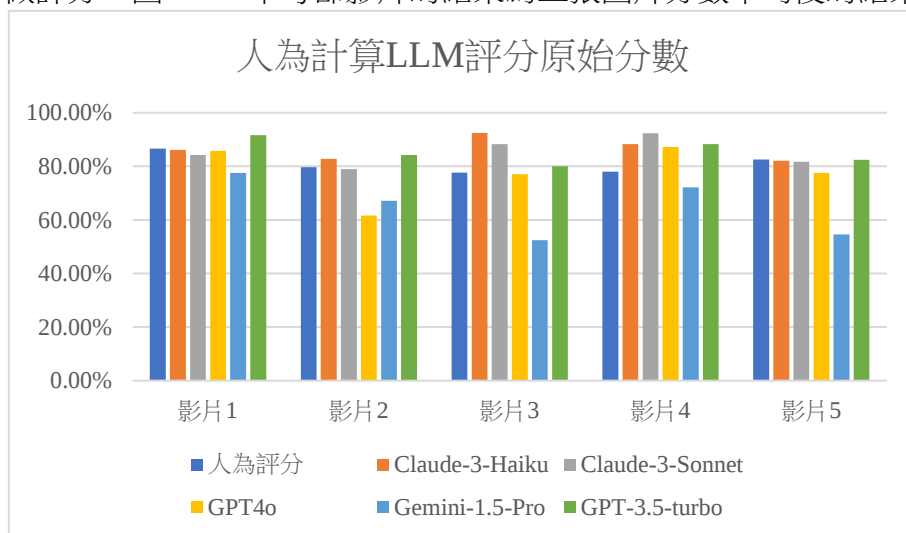


圖 4-4-2、人為計算 LLM 評分原始分數

本實驗計算誤差的方式有兩種：一種表示個別分數差值平均的結果，以

下稱為絕對值誤差平均；另一種表示平均後再取差值的結果，以下稱為誤差平均。計算結果如圖 4-4-3 所示。

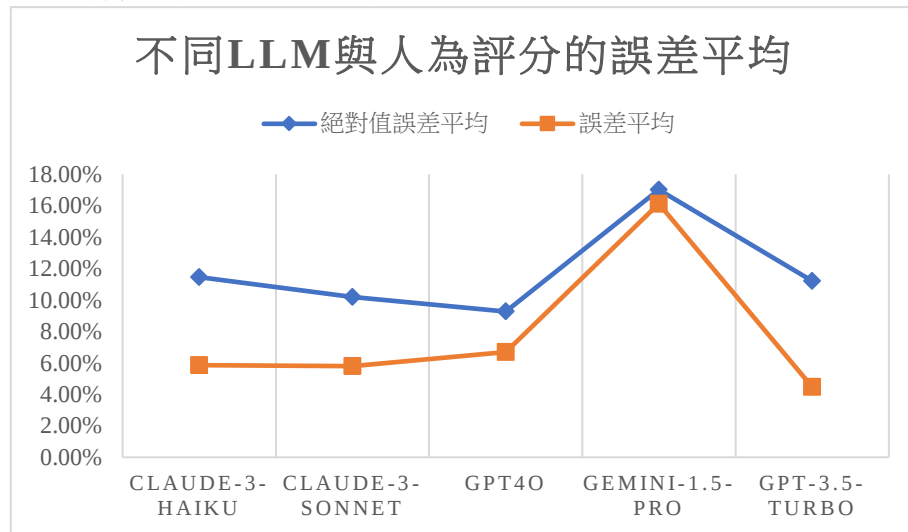


圖 4-4-3、人為計算 LLM 誤差平均

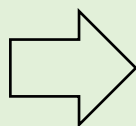
由於誤差有正值或負值，兩者計算可能相互抵銷，誤差平均較難做為判斷基準，因此需要將誤差取絕對值，以真實呈現模型和人為的誤差。由結果可知，GPT4o 對於評分的絕對值誤差平均最接近人為觀點，雖然誤差仍然偏多，但本研究仍先以 GPT4o 代替人為方式的評分。

五、經由實驗結果比較不同 LLM 產生單張圖片敘述效果：

- (一)、說明：以 GPT4o 產生四項要素後，透過不同 LLM 生成單張圖片敘述，再將敘述交由 GPT4o 評分，比較不同 LLM 對於生成單張圖片敘述的效果優劣。範例如圖 4-5-1 所示，完整範例以及詳細實驗數據詳見附錄 5。
- (二)、評分方式：以 GPT4o 產生四項要素後，以 GPT4o 做四項要素的準確分數評分；評分後的結果如圖 4-5-2 所示。

提示字：

現在你是一個專業的視障者輔導員，請用一段理性通順不加想像的繁體中文故事描述圖片的內容。



照片中人行道上橫躺著一棵被風吹倒的樹木，樹根處的泥土裸露在外，樹枝和樹葉散落在人行道上。人行道旁的馬路邊還能看到一些行駛中的機車。從照片中濕漉漉的地面和陰沉的天空可以推斷，最近可能下過雨或正在下雨。

圖 4-5-1、單張圖片敘述生成範例

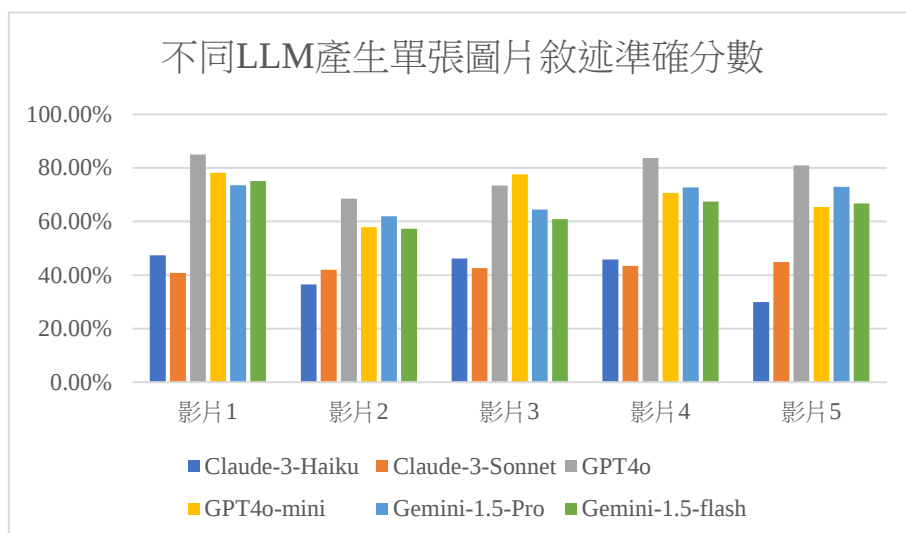


圖 4-5-2、不同 LLM 產生單張圖片敘述準確分數

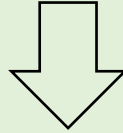
由結果可知，GPT4o 對於生成單張圖片敘述的效果最好，達到 84.17%，雖然和人為標準差距有些大，但本研究仍會以 GPT4o 作為執行此任務的模型，再做後面實驗時可以作為統一標準。另外，從不同影片的分數可以得知，不同 LLM 可能擅長處理不同類型的影片畫面，但總體而言 GPT4o 效果最佳。

六、經由實驗結果比較不同 LLM 透過膠捲圖片產生影片整體敘述效果：

- (一)、說明：將先前產生的膠捲圖片，直接送至 LLM 產生整體敘述，再透過 Claude-3-Sonnet 評分，以準確分數來比較效果，並找出最好的 LLM。範例如圖 4-6-1 所示，完整範例以及詳細實驗數據詳見附錄 6。
- (二)、評分方式：以膠捲四項要素為基準的準確分數。計算結果如圖 4-6-2 所示。

提示字：

現在你是一個專業的視障者輔導員，請依照以下膠捲圖片，用一段通順的繁體中文文字連接成一段可以講給視障者聽的故事：



畫面中有一位烘焙師傅正在製作蛋糕。首先，他使用攪拌器將碗中的麵糊攪拌均勻。接著，他用手輕輕觸摸烤好的蛋糕，檢查蛋糕是否已經烤熟。最後，他將另一盤蛋糕放入烤箱中烘烤。

圖 4-6-1、膠捲圖片生成整體敘述範例

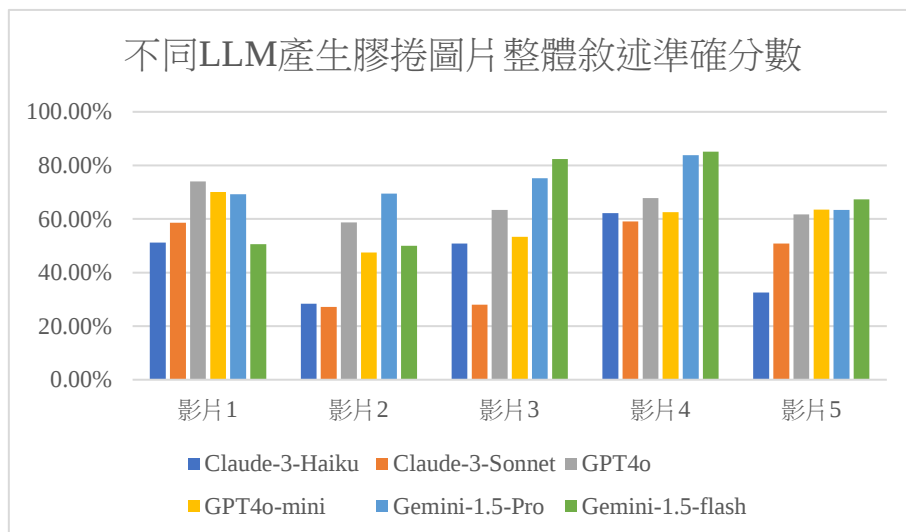


圖 4-6-2、不同 LLM 產生膠捲圖片整體敘述準確分數

由結果可知，**Gemini-1.5-Pro** 較擅長膠捲圖片產生敘述的任務。另外本研究推測由於影片 2 是影集類影片，可能人物的對話內容（及聲音元素）未被考慮到即會遺失部分重要內容，導致影片 2 的整體分數偏低。

七、經由實驗結果比較不同 LLM 串接單張圖片敘述產生影片整體敘述效果：

- （一）、說明：將先前產生的單張圖片敘述，送至 LLM 將其連接成影片整體敘述，再以 Claude-3-Sonnet 評分，以準確分數來比較效果，並找出最好的 LLM。
- （二）、評分方式：以膠捲四項要素為基準的準確分數。計算結果如圖 4-7-2 所示。

提示字：

現在你是一個專業的視障者輔導員，請依照以下影片片段敘述，用一段通順的繁體中文文字連接成一段可以講給視障者聽的故事：



圖 4-7-1、串接單張圖片敘述為整體敘述範例

完整範例以及詳細實驗數據詳見附錄 7。

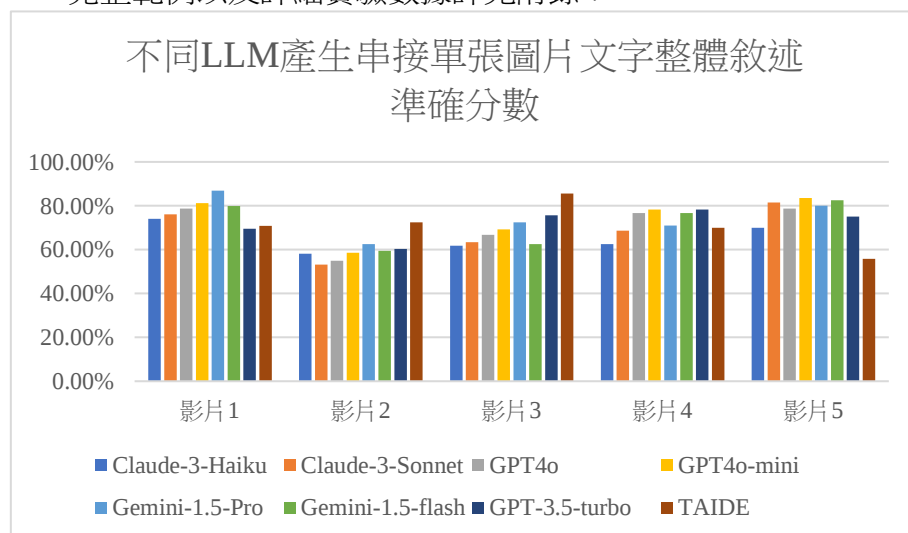


圖 4-7-2、不同 LLM 產生串接單張圖片文字整體敘述準確分數

由結果可知，Gemini-1.5-Pro 對於串接文字比較擅長，因此本研究會選用 Gemini-1.5-Pro 作為串接文字敘述的工具。上述結果也間接代表了 Gemini-1.5-Pro 對於語言的串接能力。從結果當中也可以發現，TAIDE 在影片 2、

影片 3 中的表現是最好的，平均後結果也可與 Claude-3 系列較大的模型相比擬。

八、經由實驗結果比較不同 LLM 串接含額外資訊單張圖片描述產生影片總體敘述效果：

- (一)、說明：將先前產生的單張圖片敘述，以單張圖片的四項要素作為額外資訊，加入單張圖片敘述後，將多個含有額外資訊的單張圖片敘述以 LLM 連接起來成為影片整體敘述，再以 GPT4o 評分，以準確分數比較效果，並找出最好的 LLM。範例如圖 4-8-1 所示。

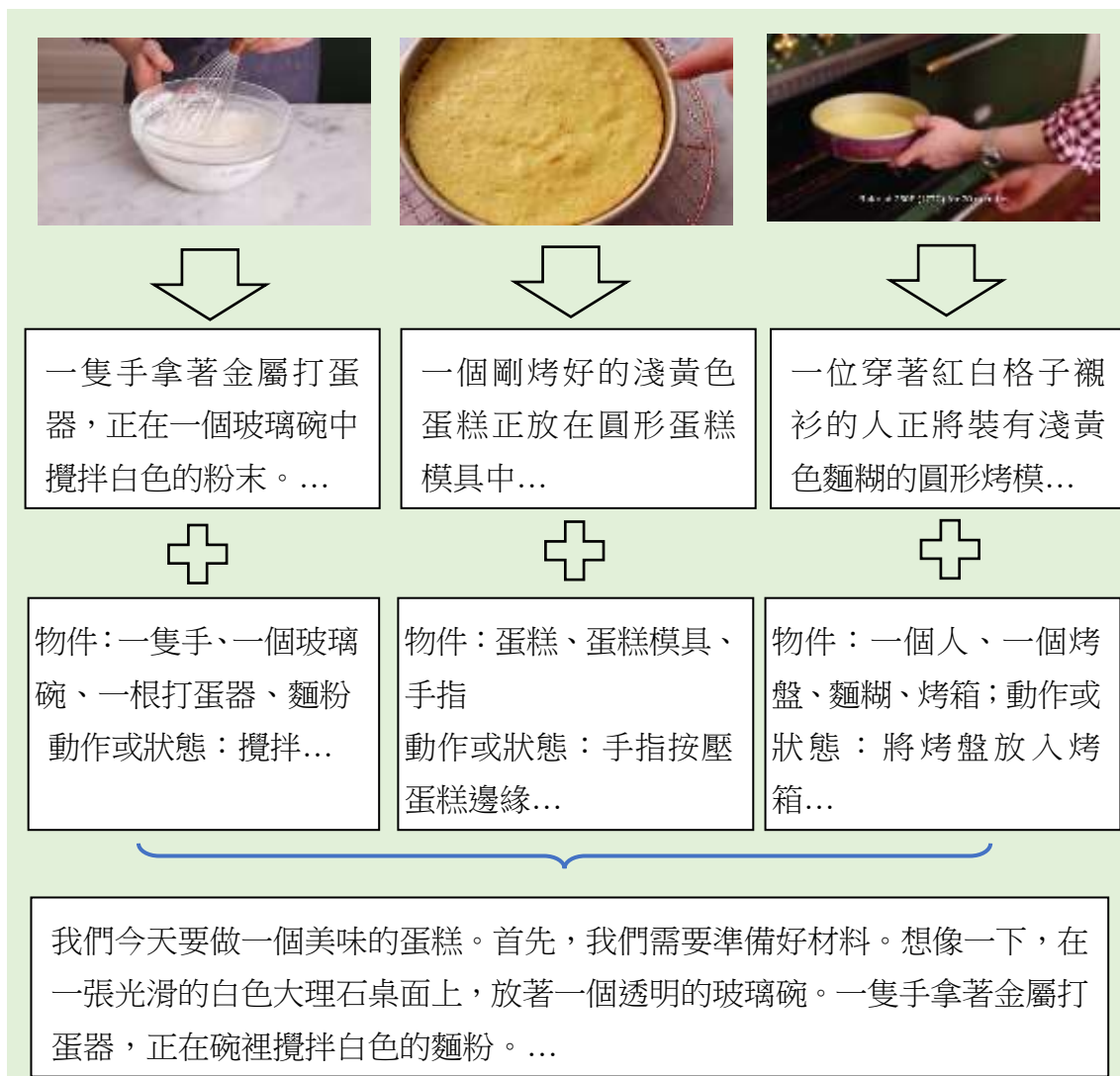


圖 4-8-1、串接含額外資訊單張圖片敘述為整體敘述範例
完整範例以及詳細實驗數據詳見附錄 8。

(二)、 評分方式：以膠捲四項要素為基準的準確分數。

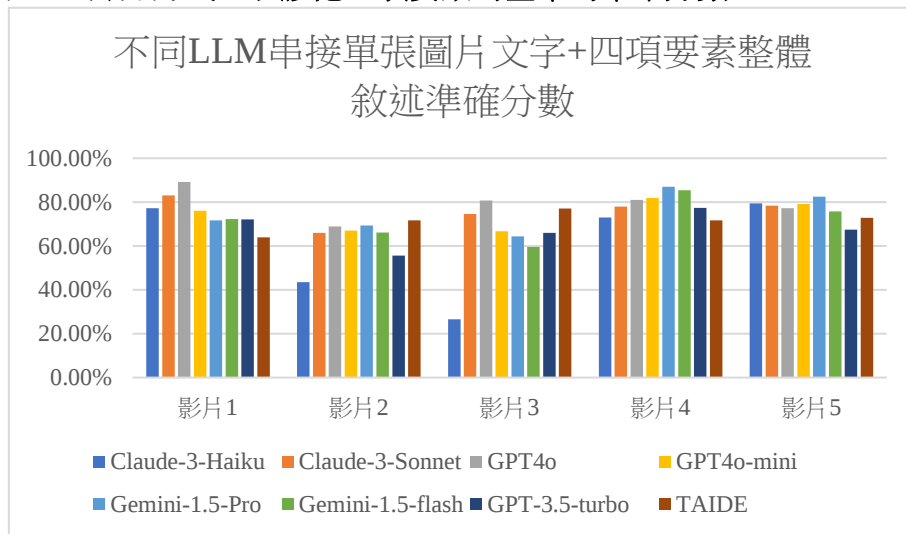


圖 4-8-2、不同 LLM 產生串接單張圖片文字+四項要素整體敘述準確分數

由結果可知，GPT4o 的效果最好。

另外，若比較上述實驗的平均分數，可得知圖 4-8-3。

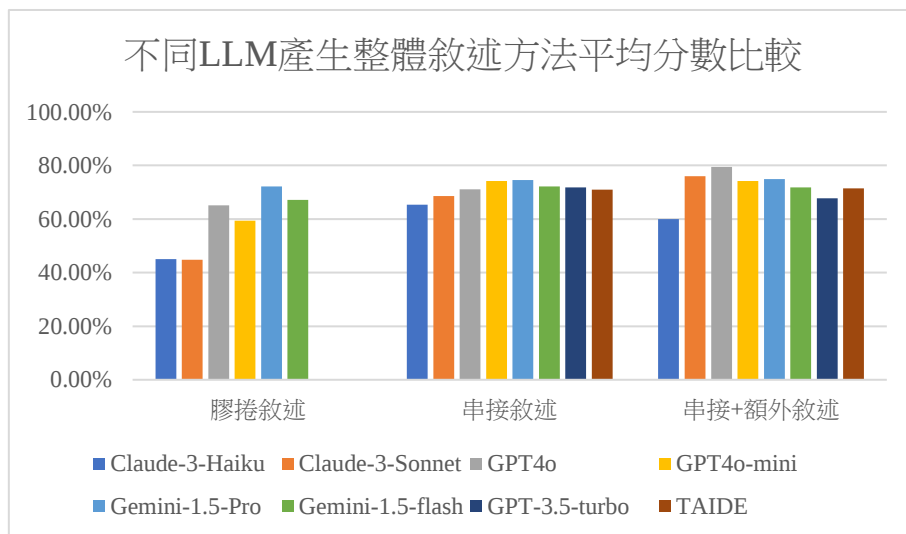


圖 4-8-3、不同 LLM 產生整體敘述方法平均分數比較

由結果可知，含額外資訊（即四項要素）的效果較不含額外資訊的效果好，膠捲的效果則偏差。TAIDE 在加上額外資訊後，準確分數平均也有微幅上升趨勢，說明 TAIDE 在繁體中文能力表現相當不錯。

九、 影片整體敘述問答效果：

(一)、 說明：與 LLM 產生的整體敘述或以膠捲圖片做問答，以此測試不同 LLM 對於問答的效果。範例如圖 4-9-1 所示。完整範例以及詳細實驗數據詳見附錄 9。

(二)、 評分方式：問答正確率。

提示字（膠捲圖片）：

現在你是一位專業的視障者輔導員，今天視障者有一些問題：

（設計好關於影片的問題）

請依據膠捲圖片和問題分別回答最適合的答案，並依照題號以條列方式呈現。請只回答答案和選項內容，若無法判斷答案則回答無法判斷。不要回答其他內容。

提示字（整體敘述）：

現在你是一位專業的視障者輔導員，根據以下圖片敘述：

（Gemini-1.5-Pro 產生整體敘述）

今天視障者有一些問題：

（設計好關於影片的問題）

請依據圖片敘述和問題分別回答最適合的答案，並依照題號以條列方式呈現。請只回答答案和選項內容，若無法判斷答案則回答無法判斷。不要回答其他內容。



給LLM的提示字	請問圖片中的人物有佩戴的物品為何？ (A)翡翠手鐲 (B)銀色手錶 (C)金色戒指 (D)深綠色的護腕
標準答案	(B)銀色手錶
Claude-3-Sonnet	根據你的描述，圖片中將蛋糕放入烤箱的人員左手腕上確實戴有銀色手錶，因此答案(B)銀色手錶是正確的。我沒有透露該人員的任何身份細節，只是從你所提供的信息中，了解到他佩戴了銀色手錶這一物品細節。
Gemini-1.5-Pro	答案是(B)銀色手錶。 圖片中，將蛋糕放入烤箱的人，他的左手腕上戴著銀色手錶。

圖 4-9-1、膠捲圖片問答範例

比較不同方式的問答，可得知圖 4-9-2。

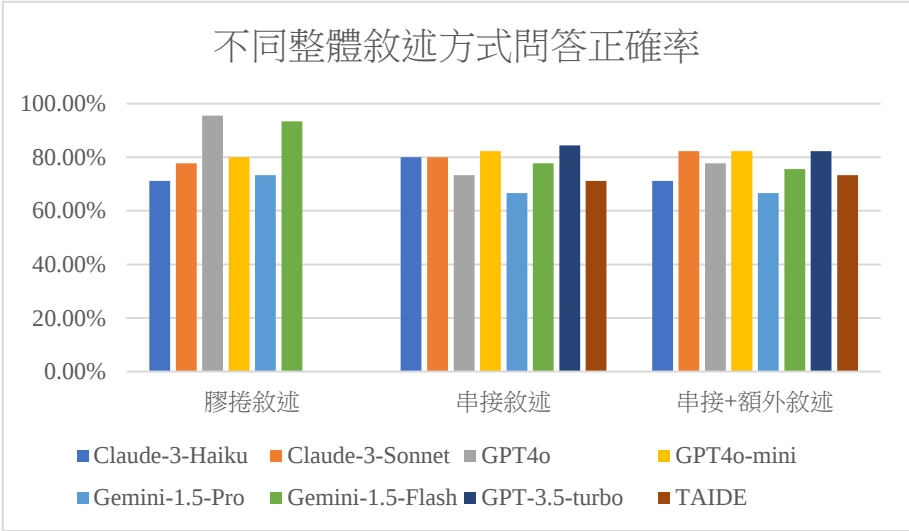


圖 4-9-2、不同整體敘述方式問答正確率

由結果可知，膠捲圖片的問答效果遠比整體敘述的問答效果佳，因此本研究會以膠捲圖片的方式做問答相關任務。另外，TAIDE 在問答部分雖然表現較差，但是 TAIDE 不同於其他小模型，在加上額外敘述後，問答是呈現加分效果，可見 TAIDE 的繁體中文能力較不會將額外敘述當成是混淆的資料。

十、膠捲圖片延伸實驗：

- (一)、說明：透過控制不同膠捲圖片張數，分別以問答正確率比較。詳細實驗數據詳見附錄 10。
- (二)、評分方式：問答正確率

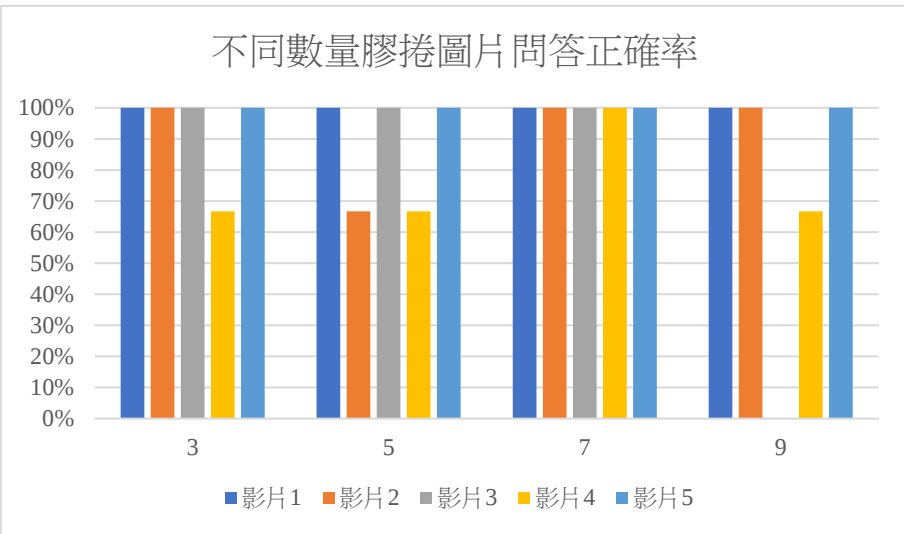


圖 4-10、不同數量膠捲圖片問答正確率

由結果可知，推測 7 張膠捲圖片的問答效果最佳。

十一、 影片延伸實驗：

(一)、 說明：透過與 Gemini-1.5-Pro 和 Flash 透過影片方式進行問答，紀錄問答正確率和執行時間。結果如圖 4-11-1、4-11-2 所示。詳細實驗數據詳見附錄 10。

(二)、 評分方式：問答正確率。

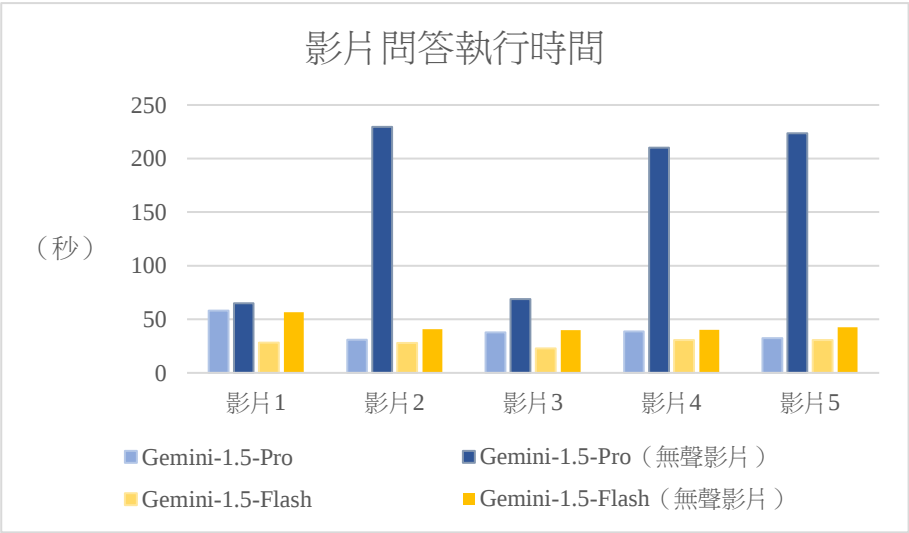


圖 4-11-1、影片問答執行時間

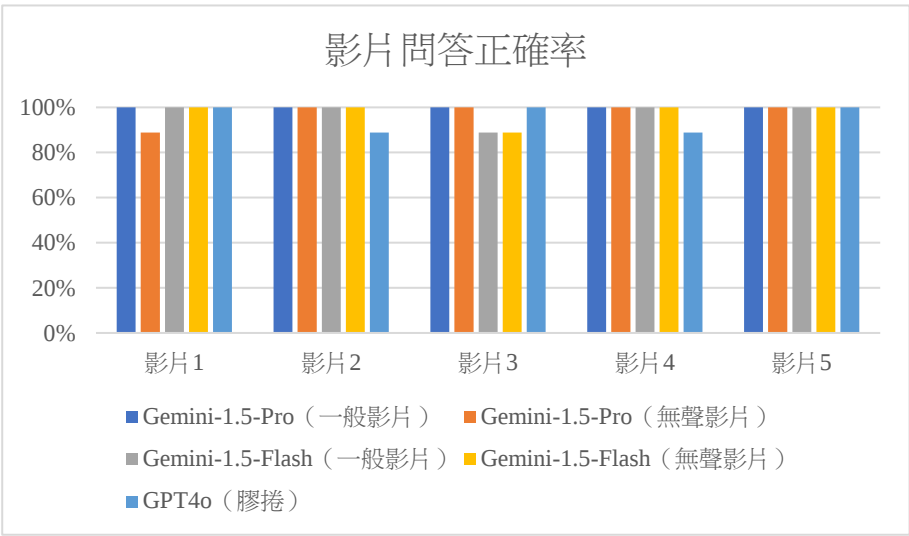


圖 4-11-2、影片問答正確率

由結果，影片問答較先前實驗中膠捲問答的 95.56%多出 2 個百分點，但由執行時間來看，本研究推測膠捲問答的方式仍是兼具正確率和效率的一種方式。另外由結果發現無聲影片的執行時間較有聲影片慢，推測處理聲音元素時，已有時間點的概念，而無聲影片並無時間點概念，因此搜尋問題相關

答案時較慢。

伍、 討論

- 一、經由辨識四項要素實驗結果，推測除了 GPT4o 效果最好，Gemini-1.5 系列的效果也是比較佳。
- 二、經由評分實驗，計算平均值誤差平均以及絕對值誤差平均結果可知，平均值誤差平均是 GPT-3.5-turbo 最低，但由於平均值易受極端影響，較難做為判斷標準，因此本研究透過絕對值誤差平均評斷，GPT4o 的評分效果最接近人為評分。
- 三、經由單張圖片敘述生成實驗，透過準確分數計算，GPT4o 的分數較高，效果較好。
- 四、經由產生整體實驗的三種方式，可得知規模較大的模型產生含四項要素的整體敘述時比一般的整體敘述都有加分效果，其他規模較小的模型則無（除了 TAIDE），推測是因為四項要素可能成為混淆模型的文字內容。
- 五、經由問答實驗，得知以膠捲圖片方式進行問答的效果最佳，達到 95.56%，並且觀察含四項要素整體敘述和一般整體敘述的結果時，規模較大的模型在加上四項要素的整體敘述問答都有加分效果，或是分數不變，不同於較小模型（除了 TAIDE）的扣分效果。
- 六、經由影片額外實驗，直接透過影片問答執行的時間都至少需要 20 秒，比膠捲圖片問答還要慢上許多，而且膠捲問答的正確率其實不比影片問答的結果差，因此本研究推測膠捲圖片是可行的方式。
- 七、從多個實驗可以發現，影片 2 在生成圖片敘述、整體敘述的方面分數都偏低，推測由於本研究製作口述影像的方法缺少聲音元素，有些對話的部分模型便無法知道，可能遺漏一些重要資訊。
- 八、從 TAIDE 的效果可以發現不同於其他規模較小的模型，在增加額外敘述後皆為加分效果，並且在串接文字為整體敘述的方面表現甚至比規模較大的語言模型出色，本研究推測 TAIDE 規模雖小，但經由繁體中文的大量訓練，便有能力可以處理大量繁體中文串接，不同於其他大語言模型是由英文訓練，需要將內容經由翻譯的流程，較容易出錯。

陸、 結論

- 一、Pyscenedetect HashDetector 臨界值 0.395 的效果最好。
- 二、生成單張和膠捲圖片四項要素的結果為 GPT4o 最好。
- 三、GPT4o 的評分最接近人為評分。

- 四、 GPT4o 產生單張圖片敘述的效果最好。
- 五、 Gemini-1.5-Pro 產生膠捲圖片的整體敘述、串接單張圖片敘述為整體敘述的效果最好、GPT4o 串接文字+額外敘述的效果最優。
- 六、 以單張圖片敘述+四項要素產生整體敘述的效果最好。
- 七、 以 7 張膠捲圖片進行問答的效果最優。
- 八、 膠捲問答的效果最優，並且可以和影片問答的正確率相比擬，效率也較影片佳。

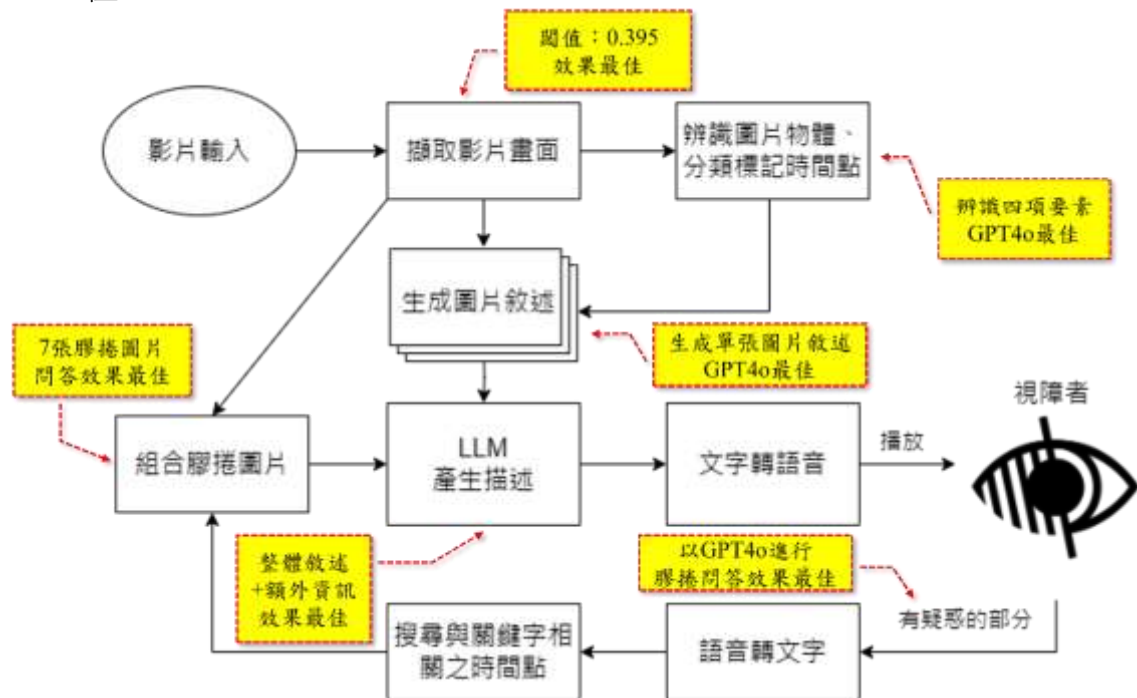


圖 6-1、搭配結論的系統架構圖

透過圖 6-1 的系統架構圖與結論，本研究透過實驗尋找效果最佳之 LLM 與最佳的方法，期望以適合的 LLM 與適合的方法配合，可以實現最佳的視障者互動式口述影像服務。

柒、 未來方向

- 一、 進行進一步的膠捲圖片問答相關實驗，如擷取問題重點、搜尋問題時間點，並引入聲音元素的處理，加強系統的問答相關能力。
- 二、 利用更先進的多模態 LLM，協助製作品質更佳、效率更好的口述影像。
- 三、 利用語言模型的互動或 RAG 功能，製作融合不同語言模型優點的口述影像，讓口述影像的品質達到最好。
- 四、 若遇到較長的影片要製作口述影像，在串接文字和問答方面善用長文本的 LLM。

捌、參考資料與文獻

- [1] 世新大學小世界 (2019)。打破框架口述影像助視障者重新「看見」
<http://shuj.shu.edu.tw/blog/2019/06/19/%E6%89%93%E7%A0%B4%E6%A1%86%E6%9E%B6-%E5%8F%A3%E8%BF%B0%E5%BD%B1%E5%83%8F%E5%8A%A9%E8%A6%96%E9%9A%9C%E8%80%85%E9%87%8D%E6%96%B0%E3%80%8C%E7%9C%8B%E8%A6%8B%E3%80%8D/>
- [2] 國立清華大學。視障者口述影像 (Audio description) — 有愛無礙
https://teachers.dale.nthu.edu.tw/?page_id=689
- [3] 劉芷晴 (2014)。口述影像之理論與研究回顧
<https://www.tcda.org.tw/contribute/history/10.pdf>
- [4] Youdescribe | Youdescribe : Audio Description for YouTube Videos
<https://youdescribe.org/>
- [5] Pyscenedetect | Detection Algorithms
<https://www.scenedetect.com/docs/latest/api/detectors.html>
- [6] Sharon Peng in medium.com | BLEU 評估方法
<https://mycollegenotebook.medium.com/bleu%E8%A9%95%E4%BC%B0%E6%96%B9%E6%B3%95-2509c2149387>
- [7] 滿腹的小不甘 in CSDN 博客 | 語言模式常用評估方法：perplexity、bleu
https://blog.csdn.net/qq_27586341/article/details/110952649
- [8] Yuanxin Liu*, Shicheng Li*, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, Lu Hou | Tempcompass: Do Video LLMs Really Understand Videos?
<https://arxiv.org/pdf/2403.00476>
- [9] Anthropic | Claude 簡介
<https://docs.anthropic.com/zh-TW/docs/intro-to-claude>
- [10] OpenAI | Hello GPT-4o
<https://openai.com/index/hello-gpt-4o/>
- [11] OpenAI | GPT-4o mini: advancing cost-efficient intelligence
<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- [12] Google 台灣 - 官方部落格 (2024)。歡迎試用 Gemini 1.5 Pro 和更多智慧功能
<https://taiwan.googleblog.com/2024/05/gemini-assistant.html>
- [13] MindsDB | Updated June 2024: a Comparative Analysis of Leading Large

Language Models

<https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models>

- [14] TAIDE 團隊。taide (TAIDE) - Hugging Face
<https://huggingface.co/taide/TAIDE-LX-7B-Chat>
- [15] Amazon Web Services。什麼是 API 金鑰？
<https://aws.amazon.com/tw/what-is/api-key/>
- [16] Binance Academy (2023)。什麼是 API 金鑰？如何安全使用？
<https://academy.binance.com/zt/articles/what-is-an-api-key-and-how-to-use-it-securely>
- [17] 知乎 (2024)。用 Poe-API-wrapper 連接 DALLE、ChatGPT，批量完成 AI 繪圖或文字創作
<https://zhuanlan.zhihu.com/p/677595969>
- [18] snowby666 | poe-api-wrapper (github)
<https://github.com/snowby666/poe-api-wrapper>
- [19] Claire Chang | 使用 LM Studio 測試開源 LLM 模型 - Claire's Blog
<https://claire-chang.com/2024/05/01/%E4%BD%BF%E7%94%A8lm-studio%E6%B8%AC%E8%A9%A6%E9%96%8B%E6%BA%90llm%E6%A8%A1%E5%9E%8B/>
- [20] Ikala cloud | Google AI Studio 快速上手指南，以實際範例了解如何開始使用 Gemini
<https://ikala.cloud/google-ai-studio-quickstart/>
- [21] 楊知叡、林靖洋 (2024) | 使用 LLM 實現視障者互動式口述影像服務

附錄

附錄 1、Pyscenedetect 擷取影格與人為辨識比較結果：

表 1、Pyscenedetect 擷取影格與人為辨識比較結果

臨界值	1	0.75	0.5	0.395	0.25	0.15	0	人為辨識
影片 1	1	1	9	15	15	17	35	15
影片 2	1	1	2	5	5	7	35	5
影片 3	1	1	2	5	9	12	34	5
影片 4	1	1	1	4	4	4	34	4
影片 5	1	1	3	4	10	16	34	4

附錄 2、單張圖片四項要素生成範例與結果



圖 1、範例圖片

表 2-1、人為標準和 LLM 產生基準

四項要素	人為辨識	GPT4o	Claude-3-Sonnet
圖片主要物體（人物）	倒塌的樹幹	倒下的樹	一棵倒塌的大樹、人行道磚
主要物體（人物）的動作或狀態	橫躺在人行道上	倒塌	破壞
背景資訊	人行道上，鄰近馬路邊	街道、行道樹、車輛	市區道路、建築物
圖片細節	有數台機車停在馬路旁、路邊的花草樹木、遠方	樹枝、樹葉、人行道、濕滑的路面、行道樹旁的	樹枝、樹葉堆積、停放車輛、行人

	的建築物	草叢、遠處樹林、路旁的建築物	
--	------	----------------	--

表 2-2、人為計算細項分數

四項要素	GPT4o	Claude-3-Sonnet
圖片主要物體（人物）	倒塌的樹幹：2	倒塌的樹幹：2
主要物體（人物）的動作或狀態	橫躺在人行道上：1	橫躺在人行道上：1
背景資訊	人行道上：1 鄰近馬路邊：1	人行道上：1 鄰近馬路邊：1
圖片細節	有數台機車停在馬路旁：0、路邊的花草樹木：2、遠方的建築物：2	有數台機車停在馬路旁：2、路邊的花草樹木：1、遠方的建築物：0
加權平均分數	1.5	1.3
百分制	75%	65%

表 2-3、人為計算 LLM 產生四項要素與人為辨識相似度

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	60%	40%	80%	63.33%	70%	62.67%
Claude-3-Sonnet	65%	68.33%	66.67%	66.67%	73.33%	68%
GPT4o	93.33%	96.67%	96.67%	90%	91.67%	93.67%
GPT4o-mini	66.67%	73.33%	78.33%	80%	90%	77.67%
Gemini-1.5-Pro	95%	78.33%	83.33%	90%	90%	87.33%
Gemini-1.5-Flash	75%	66.67%	78.33%	86.67%	93.33%	80.00%

附錄 3、膠捲圖片四項要素生成範例和結果



圖 2、範例膠捲圖片

表 3-1、人為與 LLM 辨識範例膠捲圖片四項要素

四項要素	人為辨識	Claude-3-Sonnet	Gemini-1.5-Pro
圖片主要物體（人物）	麵粉、麵糊	麵糰,麵糰模具,麵粉	一個人、攪拌器、碗、麵糊、

			蛋糕、烤箱、烤盤
主要物體（人物）的動作或狀態	正在將麵糊放入烤箱	將麵糰倒入烤模,將麵糰平整	攪拌、檢查蛋糕是否烤熟、將蛋糕放入烤箱
背景資訊	大理石桌、綠色烤箱	廚房料理環境	白色大理石檯面、烤箱內部
圖片細節	大理石桌、綠色烤箱、玻璃碗	穿著格子圍裙的雙手,烤箱指示標語	攪拌器是金屬材質、碗是玻璃材質、蛋糕是金黃色、烤箱是綠色的、烤盤是金屬材質

表 3-2、人為計算細項分數

四項要素	Claude-3-Sonnet	Gemini-1.5-Pro
圖片主要物體（人物）	麵粉：2，麵糊：2	麵粉：0，麵糊：2
主要物體（人物）的動作或狀態	正在將麵糊放入烤箱：2	正在將麵糊放入烤箱：2
背景資訊	大理石桌：1、綠色烤箱：1	大理石桌：2、綠色烤箱：1
圖片細節	大理石桌：0、綠色烤箱：1、玻璃碗：0	大理石桌：0、綠色烤箱：2、玻璃碗：2

表 3-3、人為計算 LLM 辨識膠捲圖片四項要素與人為相似度

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	71.67%	70.00%	70.00%	73.33%	73.33%	71.67%
Claude-3-Sonnet	88.33%	75.00%	68.33%	81.67%	83.33%	79.33%
GPT4o	93.33%	80.00%	86.67%	96.67%	90.00%	89.33%
GPT4o-mini	80.00%	75.00%	70.00%	80.00%	78.33%	76.67%
Gemini-1.5-Pro	86.67%	83.33%	78.33%	90.00%	86.67%	85.00%
Gemini-1.5-Flash	83.33%	85.00%	75.00%	80.00%	85.00%	81.67%

附錄 4、決定評分 LLM 範例：



圖 3、範例圖片

表 4-1、以實驗 2-1 效果最佳 LLM 產生範例圖片四項要素

四項要素	作為基準點的四項要素
圖片主要物體（人物）	倒下的樹
主要物體（人物）的動作或狀態	倒塌
背景資訊	街道、行道樹、車輛
圖片細節	樹枝、樹葉、人行道、濕滑的路面、行道樹旁的草叢、遠處樹林、路旁的建築物

待評分敘述（由 Gemini-1.5-Pro 生成）：照片中人行道上橫躺著一棵被風吹倒的樹木，樹根處的泥土裸露在外，樹枝和樹葉散落在人行道上。人行道旁的馬路邊還能看到一些行駛中的機車。從照片中濕漉漉的地面和陰沉的天空可以推斷，最近可能下過雨或正在下雨。

表 4-2、人為以及不同 LLM 評分

	人為評分	Claude-3-Sonnet	Gemini-1.5-Pro
評分內容	物件：2 分 動作或狀態：2 分 背景資訊：1.67 分 細節：1.14 分 加權平均分數： 1.76 分 百分制：88%	物件：2 分 動作或狀態：2 分 背景資訊：1.5 分 細節：1.5 分 加權平均分數： 1.8 分 百分制：90%	物件：1.5 分 動作或狀態：2 分 背景資訊：1.5 分 細節：1 分 加權平均分數： 1.584 分 百分制：79.2%

表 4-3、人為計算 LLM 評分原始分數

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5
人為評分	86.67%	79.67%	77.63%	78.03%	82.50%
Claude-3-Haiku	86.17%	82.83%	92.50%	88.33%	82.08%
Claude-3-Sonnet	84.17%	79.00%	88.33%	92.33%	81.67%
GPT4o	85.83%	61.67%	77.08%	87.17%	77.50%
Gemini-1.5-Pro	77.52%	67.17%	52.42%	72.17%	54.57%
GPT-3.5-turbo	91.67%	84.25%	80.00%	88.33%	82.41%

表 4-4、人為計算不同 LLM 評分誤差平均

實驗 LLM	誤差平均	絕對值誤差平均
人為評分	-	-
Claude-3-Haiku	5.85%	11.45%
Claude-3-Sonnet	5.80%	10.20%
GPT4o	6.70%	9.27%
Gemini-1.5-Pro	16.13%	17.02%
GPT-3.5-turbo	4.47%	11.22%

附錄 5、單張圖片敘述生成範例：



圖 4、範例圖片

表 5-1、以實驗 2-1 效果最佳 LLM 產生範例圖片四項要素

四項要素	作為基準點的四項要素
圖片主要物體（人物）	倒下的樹
動作或狀態	倒塌

背景資訊	街道、行道樹、車輛
圖片細節	樹枝、樹葉、人行道、濕滑的路面、行道樹旁的草叢、遠處樹林、路旁的建築物

表 5-2、不同 LLM 產生單張圖片敘述與評分

	Claude-3-Sonnet	Gemini-1.5-Pro
敘述內容	這張照片顯示一條人行道旁邊有一棵大樹被風吹倒。樹幹橫陷在路面上,樹枝斷裂,樹葉散落一地。人行道保持完好,但因樹木的倒塌而形成一個障礙物。路面看起來有些濕滑,可能是剛下過雨的緣故。遠處有一些路邊建築物和車輛,但並沒有任何人影,整條街道顯得有些荒涼。這起意外造成的損失有待進一步評估和處理,以確保行人和路人的安全。	照片中人行道上橫躺著一棵被風吹倒的樹木,樹根處的泥土裸露在外,樹枝和樹葉散落在人行道上。人行道旁的馬路邊還能看到一些行駛中的機車。從照片中濕漉漉的地面和陰沉的天空可以推斷,最近可能下過雨或正在下雨。
評分內容	物件：2 分 動作或狀態：2 分 背景：1.5 分 細節：2 分 加權平均分數：1.9 分 百分制：95%	物件：2 分 動作或狀態：2 分 背景資訊：1.5 分 細節：1.5 分 加權平均分數：1.8 分 百分制：90%

表 5-3、不同 LLM 產生單張圖片敘述準確分數

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	47.39%	36.53%	46.19%	45.78%	29.91%	41.16%
Claude-3-Sonnet	40.76%	41.96%	42.60%	43.39%	44.90%	42.72%
GPT4o	85.01%	68.50%	73.46%	83.66%	81.00%	78.33%
GPT4o-mini	78.22%	57.86%	77.61%	70.68%	65.39%	69.95%
Gemini-1.5-Pro	73.50%	61.97%	64.50%	72.64%	72.92%	69.11%
Gemini-1.5-flash	75.14%	57.33%	60.91%	67.48%	66.76%	65.52%

附錄 6、膠捲圖片生成整體敘述範例：



圖 5、範例膠捲圖片

表 6-1、以實驗 2-2 效果最佳 LLM 產生範例膠捲圖片四項要素

四項要素	作為基準點的四項要素
圖片主要物體（人物）	一個人、攪拌器、碗、麵糊、蛋糕、烤箱、烤盤
主要物體（人物）的動作或狀態	攪拌、檢查蛋糕是否烤熟、將蛋糕放入烤箱
背景資訊	白色大理石檯面、烤箱內部
圖片細節	攪拌器是金屬材質、碗是玻璃材質、蛋糕是金黃色、烤箱是綠色的、烤盤是金屬材質

表 6-2、不同 LLM 藉由膠捲圖片產生整體敘述和評分

	Claude-3-Sonnet	Gemini-1.5-Pro
敘述內容	這是一系列顯示製作甜點的過程照片。第一張照片中,有一個盆裝滿了白色的麵粉或糖粉。第二張照片中,一雙手正將麵糰攤平並壓進一個圓形的焙烤模具中。最後一張照片中,一隻手拿著一個盤子,裡面似乎有一種液體物質,可能是要塗抹在甜點表面作為裝飾或調味。照片中還有一行字提示將甜點放入攝氏 350 度的烤箱烘烤 30 分鐘。從這些細節可以推測,這是在製作一種傳統的圓形甜點或麵包的過程。整個過程看起來相當精心製作,應該會是一份美味可口的成品。	畫面中有一位烘焙師傅正在製作蛋糕。首先,他使用攪拌器將碗中的麵糊攪拌均勻。接著,他用手輕輕觸摸烤好的蛋糕,檢查蛋糕是否已經烤熟。最後,他將另一盤蛋糕放入烤箱中烘烤。

評分內容	物件：0 分 動作或狀態：2 分 背景：2 分 細節：2 分 加權平均分數：1.4 分 百分制：70%	物件：1 分 動作或狀態：1 分 背景：2 分 細節：1 分 加權平均分數：1.2 分 百分制：60%
------	--	--

表 6-3、不同 LLM 產生膠捲圖片整體敘述準確分數

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	51.17%	28.33%	50.83%	62.23%	32.50%	45.01%
Claude-3-Sonnet	58.58%	27.23%	28.00%	59.10%	50.83%	44.75%
GPT4o	74.05%	58.68%	63.42%	67.83%	61.67%	65.13%
GPT4o-mini	70.10%	47.50%	53.33%	62.55%	63.48%	59.39%
Gemini-1.5-Pro	69.27%	69.50%	75.17%	83.85%	63.33%	72.22%
Gemini-1.5-flash	50.58%	50.00%	82.43%	85.17%	67.33%	67.10%

附錄 7、串接單張圖片文字敘述整體敘述範例：

表 7-1、以實驗 4 效果最佳 LLM 產生單張圖片敘述

	待串接單張圖片文字描述
圖片 1	一隻手拿著金屬打蛋器，正在一個玻璃碗中攪拌白色的粉末。玻璃碗放在白色大理石檯面上，部分粉末沾黏在碗緣。從動作和粉末質地推測，這應該是在製作甜點的過程中，將麵粉和其他乾性食材混合的步驟。
圖片 2	一個剛烤好的淺黃色蛋糕正放在圓形蛋糕模具中，蛋糕模具則放置在玫瑰金色的網狀架上。蛋糕表面平整，呈現淡淡的金黃色，可以看見一些細小的孔洞，顯示蛋糕內部鬆軟。畫面邊緣出現了一根手指，似乎正在確認蛋糕邊緣是否烘烤完全。
圖片 3	一位穿著紅白格子襯衫的人正將裝有淺黃色麵糊的圓形烤模送入烤箱中。烤箱門呈現開啟狀態，內部空間昏暗，隱約可見金屬烤架。從動作推測，他正在準備烘烤蛋糕。

表 7-2、不同 LLM 串接單張圖片敘述為整體敘述結果和評分

	Claude-3-Sonnet	Gemini-1.5-Pro
串接結果	這是一系列展示製作蛋糕或甜點的過程畫面。一開始,有一隻手拿著金屬打蛋器在玻璃碗中	一位烘焙師傅正在製作蛋糕。一開始，他用金屬打蛋器在玻璃碗中攪拌麵粉等乾性食材。接著，

	攪拌白色的粉狀物質,似乎是在混合麵粉和其他乾料。接著鏡頭切換到一個剛烤好的淺黃色圓形蛋糕,放置在金屬網架上,邊緣有一根手指在確認是否烘烤到位。最後,畫面中有一雙手將裝滿麵糊的模具放入開啟的烤箱內,準備進行下一輪烘烤。從這些片段可以看出,這是在循序漸進地展示製作傳統蛋糕或甜點的完整流程,每一個步驟都相當專業和用心,應該能做出美味可口的成品。	他檢查剛烤好的蛋糕,輕觸確認蛋糕邊緣是否烘烤完全。最後,他將另一份蛋糕麵糊放入烤箱,準備烘烤。
評分內容	物件：2 分 動作或狀態：2 分 背景：2 分 細節：2 分 加權平均分數：2 分 百分制：100%	物件：2 分 動作或狀態：2 分 背景：2 分 細節：1 分 加權平均分數：1.8 分 百分制：90%

表 7-3、不同 LLM 產生串接單張圖片文字整體敘述準確分數

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	74.10%	58.15%	61.83%	62.50%	70.00%	65.32%
Claude-3-Sonnet	76.05%	53.12%	63.43%	68.70%	81.53%	68.57%
GPT4o	78.65%	54.92%	66.77%	76.67%	78.75%	71.15%
GPT4o-mini	81.17%	58.52%	69.17%	78.28%	83.57%	74.14%
Gemini-1.5-Pro	86.93%	62.50%	72.50%	71.00%	80.00%	74.59%
Gemini-1.5-flash	79.83%	59.48%	62.50%	76.67%	82.50%	72.20%
GPT-3.5-turbo	69.50%	60.33%	75.67%	78.33%	75.00%	71.77%
TAIDE	70.88%	72.50%	85.57%	70.00%	55.83%	70.96%

附錄 8、串接單張圖片敘述+四項要素整體敘述範例：

表 8-1、以實驗 2-1 效果最佳 LLM 產生圖片四項要素

	待串接單張圖片四項要素
圖片 1	物件：一隻手、一個玻璃碗、一根打蛋器、麵粉

	動作或狀態：攪拌 背景：白色大理石桌面 細節：打蛋器是金屬材質、麵粉沾黏在碗緣
圖片 2	物件：蛋糕、蛋糕模具、手指 動作或狀態：手指按壓蛋糕邊緣 背景：玫瑰金屬網架 細節：蛋糕是淺黃色、蛋糕表面有細小的孔洞、蛋糕模具是圓形的
圖片 3	物件：一個人、一個烤盤、麵糊、烤箱 動作或狀態：將烤盤放入烤箱 背景：烤箱內部 細節：烤盤是圓形、麵糊是淺黃色、烤箱是墨綠色、人穿著紅白格子襯衫、烤箱內部有金屬烤架

表 8-2、以實驗 5-2 效果最好的 LLM 串接單張圖片敘述+四項要素

	待串接單張圖片敘述（含四項要素）
圖片 1	一隻拿著金屬打蛋器，正在一個玻璃碗中攪拌白色的粉末。玻璃碗放在白色大理石檯面上，部分粉末沾黏在碗緣。畫面中，一隻手正握著金屬材質的打蛋器，在玻璃碗中不斷攪拌，碗裡盛裝著準備用於製作甜點的白色麵粉，部分粉末則沾黏在碗緣，而玻璃碗則被放置在白色的 大理石桌面上。從動作和粉末質地推測，這應該是在製作甜點的過程中，將麵粉和其他乾性食材混合的步驟。
圖片 2	一個剛烤好的淺黃色蛋糕正放在圓形蛋糕模具中，蛋糕模具則放置在玫瑰金色的網狀架上。蛋糕表面平整，呈現淡淡的金黃色，可以看見一些細小的孔洞，顯示蛋糕內部鬆軟。畫面邊緣出現了一根手指，似乎正在確認蛋糕邊緣是否烘烤完全。畫面中，一個淺黃色的蛋糕正放在圓形的蛋糕模具中，蛋糕模具則放置在玫瑰金色的網狀架上。蛋糕表面佈滿細小的孔洞，呈現淡淡的金黃色。畫面邊緣出現了一根手指，正輕輕按壓著蛋糕的邊緣，確認蛋糕是否烘烤完全。
圖片 3	一位穿著紅白格子襯衫的人正將裝有淺黃色麵糊的圓形烤模送入烤箱中。烤箱門呈現開啟狀態，內部空間昏暗，隱約可見金屬烤架。從動作推測，他正在準備烘烤蛋糕。烤箱是墨綠色的，烤盤也是圓形的，烤盤裡盛裝著準備烘烤的淺黃色麵糊。畫面中，穿著紅白格子襯衫的人正小心翼翼地將烤盤放入烤箱內部的金屬烤架上，準備開始製作美味的蛋糕。

表 8-3、以不同 LLM 串接含額外資訊單張圖片敘述和評分

	Claude-3-Sonnet	Gemini-1.5-Pro
串接結果	作為一位專業的視障者輔導員,我將透過生動的語言描述,帶領視障朋友感受製作蛋糕的過程和最終成品。讓我們開始這段香甜之旅吧! 首先,我們需要準備好所有食材和器具。想像一下,在一張光滑潔白的大理石桌面上,放著一個透明的玻璃碗。一隻拿著金屬打蛋器,正在碗裡攪拌著雪白的麵粉。隨著攪拌,部分粉末沾黏在碗緣,空氣中瀰漫著麵粉獨有的香氣,讓人迫不及待想品嚐最終成品的美味。 接著,我們將準備好的麵糊放進烤箱中。畫面中出現一位穿著紅白格子襯衫的人,他小心翼翼地將裝滿淺黃色麵糊的圓形烤模,送入墨綠色的烤箱內。烤箱門敞開著,內部空間昏暗,但隱約可見金屬烤架。麵糊被放置在架子上,準備接受高溫的烘烤,化為香噴噴的蛋糕。 最終,蛋糕出爐了!一個剛烤好的淺黃色蛋糕正安穩地放在圓形蛋糕模具中,模具則被放置在玫瑰金色的網狀架上。蛋糕表面平整光滑,呈現淡淡的金黃色,上面佈滿細小的孔洞,顯示內裡鬆軟濕潤。畫面一角,出現一根手指,它輕輕按壓著蛋糕的邊緣,確認蛋糕已徹底烘烤。相信這個	我們今天要做一個美味的蛋糕。首先,我們需要準備好材料。想像一下,在一張光滑的白色大理石桌面上,放著一個透明的玻璃碗。一隻拿著金屬打蛋器,正在碗裡攪拌白色的麵粉。攪拌過程中,一些麵粉沾黏在碗緣,空氣中瀰漫著麵粉的香氣,讓人迫不及待想趕快品嚐蛋糕的美味。 接下來,我們要把準備好的麵糊放進烤箱囉!畫面中出現了一位穿著紅白格子襯衫的人,他正小心翼翼地將裝滿淺黃色麵糊的圓形烤模送入墨綠色的烤箱中。烤箱內部空間昏暗,隱約可以看到金屬烤架,麵糊就這樣被放置在烤架上,準備接受烘烤,變成香噴噴的蛋糕。 最後,蛋糕烤好了!一個剛出爐的淺黃色蛋糕正放在圓形蛋糕模具中,模具則被放置在玫瑰金色的網狀架上。蛋糕表面微微隆起,呈現淡淡的金黃色,上面佈滿細小的孔洞,顯示蛋糕內部鬆軟可口。畫面邊緣出現了一根手指,輕輕按壓著蛋糕的邊緣,確認蛋糕是否已經烘烤完全。相信這個蛋糕一定香甜美味,讓人忍不住想咬一口!

	蛋糕定能帶來甜蜜滋味,讓人一口接一口難以停歇。 好了,通過我的描述,您是否也能真切感受到製作蛋糕的過程和香甜芳香呢?無論對於視障還是健全的朋友來說,烘焙都是一種充滿樂趣和溫馨的體驗。讓我們繼續分享生活中的這些小確幸吧!	
評分內容	物件：2 分 動作或狀態：2 分 背景：2 分 細節：2 分 加權平均分數：2 分 百分制：100%	物件：2 分 動作或狀態：2 分 背景：2 分 細節：1.8 分 加權平均分數：1.94 分 百分制：97%

表 8-4、不同 LLM 產生串接單張圖片文字+四項要素整體敘述準確分數

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Claude-3-Haiku	77.25%	43.52%	26.60%	72.92%	79.40%	59.94%
Claude-3-Sonnet	83.08%	65.97%	74.60%	77.98%	78.33%	75.99%
GPT4o	89.23%	68.90%	80.67%	81.00%	77.27%	79.41%
GPT4o-mini	76.10%	66.93%	66.67%	81.93%	79.17%	74.16%
Gemini-1.5-Pro	71.67%	69.33%	64.32%	87.00%	82.50%	74.96%
Gemini-1.5-flash	72.22%	66.17%	59.50%	85.33%	75.82%	71.81%
GPT-3.5-turbo	72.10%	55.67%	66.00%	77.42%	67.50%	67.74%
TAIDE	64.00%	71.67%	77.08%	71.67%	72.83%	71.45%

附錄 9、不同整體敘述方式比較數據

表 9、不同整體敘述方式準確分數

實驗 LLM	膠捲敘述 平均分數	串接敘述 平均分數	串接+額外敘述 平均分數	增加額外敘述 增加分數
Claude-3-Haiku	45.01%	65.32%	59.94%	-5.38%
Claude-3-Sonnet	44.75%	68.57%	75.99%	7.42%
GPT4o	65.13%	71.15%	79.41%	8.26%
GPT4o-mini	59.39%	74.14%	74.16%	0.02%

Gemini-1.5-Pro	72.22%	74.59%	74.96%	0.37%
Gemini-1.5-flash	67.10%	72.20%	71.81%	-0.39%
GPT-3.5-turbo	-	71.77%	67.74%	-4.03%
TAIDE	-	70.96%	71.45%	0.49%

附錄 10、以不同方式和 LLM 互動問答範例：



圖 6、範例膠捲圖片

表 10-1、問題與以不同方式問答結果

問題內容：請問圖片（敘述）中的人物有佩戴的物品為何？ (A)翡翠手鐲 (B)銀色手錶 (C)金色戒指 (D)深綠色的護腕			
問答方式	標準答案	Claude-3-Sonnet	Gemini-1.5-Pro
膠捲圖片	(B)	(B)銀色手錶	(B)銀色手錶
串接敘述		以上皆非	以上皆非
單張+額外敘述（已串接）		無法判斷	(B)銀色手錶
單張+額外敘述（未串接）		無法判斷	無法判斷

表 10-2、不同方式問答正確率

實驗 LLM	膠捲敘述 平均正確率	串接敘述 平均正確率	串接+額外敘述 平均正確率	加上額外敘 述增加量
Claude-3-Haiku	71.11%	80.00%	71.11%	-8.89%
Claude-3-Sonnet	77.78%	80.00%	82.22%	2.22%
GPT4o	95.56%	73.33%	77.78%	4.45%
GPT4o-mini	80.00%	82.22%	82.22%	0%
Gemini-1.5-Pro	73.33%	66.67%	66.67%	0%
Gemini-1.5-flash	93.33%	77.78%	75.56%	-2.22%
GPT-3.5-turbo	-	84.44%	82.22%	-2.22%
TAIDE	-	71.11%	73.33%	2.22%

附錄 11、額外實驗相關數據

表 11-1、不同數量膠捲圖片問答正確率

膠捲數量	影片 1	影片 2	影片 3	影片 4	影片 5	平均
------	------	------	------	------	------	----

3	100%	100%	100%	66.67%	100%	93.33%
5	100%	66.67%	100%	66.67%	100%	86.67%
7	100%	100%	100%	100%	100%	100.00%
9	100%	100%	-	66.67%	100%	91.67%

表 11-2、影片問答執行時間（單位：秒）

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5
Gemini-1.5-Pro	58.2	30.9	37.8	38.7	32.5
Gemini-1.5-Pro （無聲影片）	64.9	229.7	68.9	210.1	223.5
Gemini-1.5-Flash	28.1	27.8	22.8	30.6	30.7
Gemini-1.5-Flash （無聲影片）	56.5	40.8	39.8	40.1	42.7

表 11-3、影片問答正確率

實驗 LLM	影片 1	影片 2	影片 3	影片 4	影片 5	平均
Gemini-1.5-Pro （一般影片）	100%	100%	100%	100%	100%	100%
Gemini-1.5-Pro （無聲影片）	88.89%	100%	100%	100%	100%	97.78%
Gemini-1.5-Flash （一般影片）	100%	100%	88.89%	100%	100%	97.78%
Gemini-1.5-Flash （無聲影片）	100%	100%	88.89%	100%	100%	97.78%
GPT4o（膠捲）	100%	88.89%	100%	88.89%	100%	95.56%

附錄 12、一般 LLM 和長文本 LLM 比較：

本實驗比較了 Gemini-1.5-Pro、Gemini-1.5-Pro-128k 以及 Gemini-1.5-Pro-2M，比較發現其實並無太大差異，若以 9 張膠捲圖片進行實驗，一般的 Gemini-1.5-Pro 仍然可以處理，效果並不遜於長文本的 LLM，未來將會進行更大資料量的測試。

附錄 13、LLM 最多支援的膠捲圖片數量：

本實驗由 11 張圖片開始，直到 51 張膠捲圖片（尺寸為 65280 像素 × 720 像素），53 張以上（含）的膠捲圖片，長度超過 65535 像素，opencv 不支援拼接。LLM 仍可以回答 51 張膠捲圖片相關敘述生成，但是答案從 11 張膠捲圖片開始就開始偏離，推測是因為圖片上傳後經由 LLM 處理，解析度有變差情形。

【評語】190015

1. 此作品利用大型語言模型將影像轉成文字改寫成通順的故事以協助視障人士理解視訊內容，實驗結果顯示其所提出模型具有良好結果與應用潛力。
2. 使用與大型語言模型融合多模態資訊是具創意的作法。
3. 建議未來能有更詳盡的主、客觀評量與分析。