字幕中指示代名詞解析系統：結合多模態視覺感知技術
Demonstrative Pronoun Resolution System in Subtitles: A Multimodal Visual Perception Approach

組別：B-5　　專題成員：練亭蘭　　指導教授：李信杰 教授

研究動機與目的

現行的語音辨識技術及工具僅能擷取影片中的口語內容，無法保留畫面所承載的視覺資訊。當講者使用「這個」、「this」等指示代名詞 (demonstrative pronoun) 指涉畫面上的特定對象時，純語音逐字稿便會遺失這些指示代名詞所對應的實際內容，造成語意斷裂，導致讀者單看文字難以理解講者真正的意涵。

本專題旨在建構一套整合語音辨識與多模態視覺感知的指示代名詞解析系統。系統會自動解析逐字稿中的多語言指示代名詞，擷取其對應時刻的影片畫面，並運用多模態視覺感知技術進行分析，將畫面內容以補充說明的形式融合回原始逐字稿。針對回想型敘述，系統則進一步透過語義分析定位講者所回想的先前段落內容與時間點，將其加註於逐字稿中，並對該回想概念擷取對應畫面截圖以供使用者參考對照。

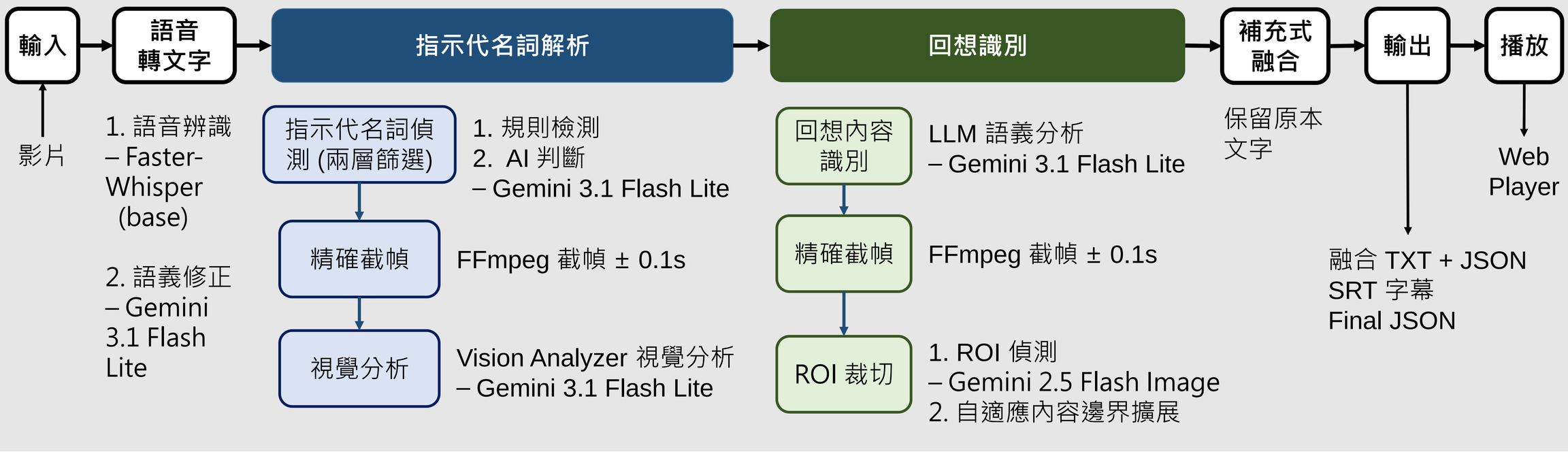
實際技術與環境

本系統以 Python 3.10 版本開發，整合多項開源框架與多模態模型。在語音辨識方面，採用 Faster-Whisper 轉錄，產生帶有時間戳記的逐字稿並自動進行語言偵測。多語言指示代名詞解析與回想偵測則以 Google Gemini 3.1 Flash Lite 進行視覺描述、決策判斷與語義分析。

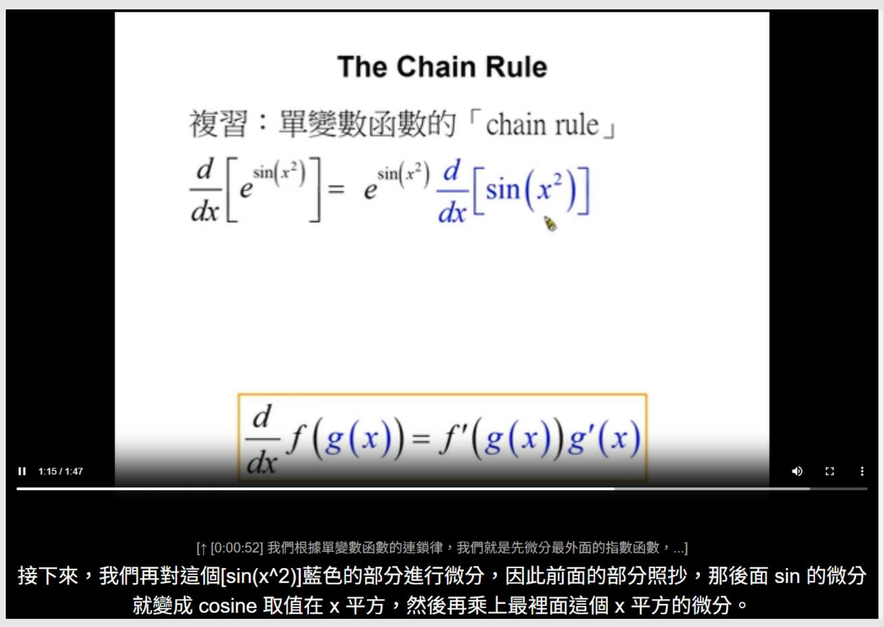
影片與影像處理方面，採用 FFmpeg 進行精確秒數截幀，可達正負零點一秒的時間定位精度，回想截圖的關注區域 (ROI) 邊界偵測採用 Gemini 2.5 Flash Image，並以 OpenCV 完成區域裁切。

開發與整合環境方面，系統透過 google-generativeai 套件介接 Gemini API，所有 API 金鑰均以環境變數方式管理，以兼顧安全性與部署彈性。

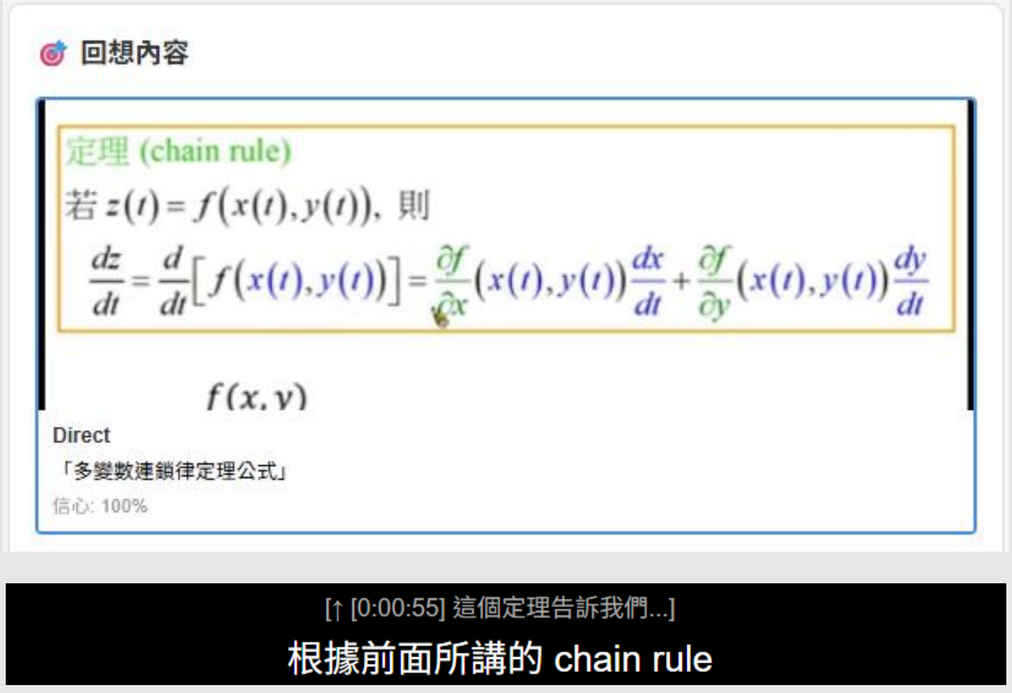
系統架構圖 Architecture



成果展示 Results



講者說「這個」→ 系統解析為畫面中的 sin(x^2)



講者說「根據前面所講的」
→ 系統自動定位講者回想的先前段落

1. 於字幕處標註時間 (↑ [0:00:55])
2. 於回想內容欄位附上可跳轉的畫面截圖、概念簡述與信心度

藍色框暗示使用者可以點擊跳轉影片回該概念(0:00:55)

結論與未來展望

本專題成功建構一套整合語音辨識與多模態視覺感知的字幕指示代名詞解析系統，有效解決傳統逐字稿僅保留口語內容、遺失畫面視覺資訊的問題。透過多語言指示代名詞解析、精確秒數截幀與多模態視覺分析的串接，系統能解析指示代名詞所指對象，並進一步透過回想內容識別，定位講者所回想的先前段落並提取對應畫面與區域裁切圖供參考。最終系統搭配網頁播放器呈現影片、字幕與回想內容，提升純文字的可理解程度。

在未來展望方面，本系統仍有多個面向可持續優化。於視覺處理層面，可進一步提升描述指示代名詞的精確度與回想畫面截圖的區域裁切精度。此外，系統亦可朝向指標偵測（如滑鼠或雷射筆位置）、更多輸出格式支援，以及處理效能的優化等方向延伸，以提升整體的實用性與適用範圍。

專題連結

歡迎掃描 QR Code 觀看系統 Demo 影片 。

Github：