

# 中華民國第 65 屆中小學科學展覽會

## 作品說明書

---

高級中等學校組 電腦與資訊學科

第三名

052516

基於多項人工智慧之多模態失語症溝通系統探究

學校名稱： 國立花蓮高級中學

作者：  高二 林炫宇  高二 莊家瑋	指導老師：  趙義雄
---------------------------------	------------------

關鍵詞： 失語症、多模態、自然語言處理

## 摘要

本研究開發了一個基於多項人工智慧技術的多模態失語症溝通系統，旨在提供失語症患者能提升表達自我和參與社交的能力的工具。系統以 iOS App 作為主體，接收環境影像、語音、唇語、手勢及情緒等多模態資訊後，透過大型語言模型產生完整敘述，再以語音合成播放。研究創新開發 AphasiaSim-LLM 生成高度擬真的模擬失語症語料，並以量化評估取代主觀評分，證實 Gemini 2.5 Flash 於語句還原表現最佳；此外亦建構了輕量級手勢辨識模型，並以 ORB 演算法優化關鍵幀提取；透過異步處理、FFmpeg 影片取幀及輕量 Flux 文生圖模型等優化策略，最終完成能有效輔助失語症患者進行流暢溝通之系統。

## 壹、研究動機

家人曾提及，對於某位知名好萊塢影星因罹患失語症而不得不息影一事感到惋惜。在對失語症有了更多了解後，我們不禁好奇近年來快速發展的人工智慧模型，是否能為這些病患提供協助，減輕他們的生活負擔？

失語症患者因中風、腦部創傷等原因，喪失用言語表達思想的能力，無法順暢與他人溝通。這種溝通困難對照顧者和患者都是巨大的挑戰。醫護人員在照顧失語症患者時，通常需借助肢體語言、圖片或繪圖等方式輔助，但這些方法仍有局限性，難以準確理解患者的想法。

隨著人工智慧、大型語言模型（LLM）和 Generative Pre-trained Transformer（GPT）的快速發展，解決失語症患者的溝通問題出現了新契機。我們的目標是開發一款 iOS App，結合環境影像分析、患者語音分析、唇語辨識、手勢動作識別和情緒識別的多模態大語言模型架構，協助醫護人員更精確地理解失語症患者的需求，從而提升溝通的流暢度與效率。

## 貳、研究目的

- 一、開發 AphasiaSim-LLM 生成高度擬真的模擬失語症語料。
- 二、建構量化評估系統，取代傳統主觀評分方式。
- 三、分析不同多模態大型語言模型於失語症語句還原的效果。
- 四、輕量手勢辨識模型的建構與優化，透過肢體補充語意表達效果。
- 五、分析多種幀間差異演算法，設計適用於本系統之關鍵幀提取機制。

六、整合 Chain-of-Thought 推理策略與多代理模型架構，強化系統的理解能力，並建立多種對話支援模式。

七、結合 LLM、Flux 及 CLIP 模型，探討圖卡即時生成與繪圖建議系統的最佳化策略。

八、結合 LLM 與圖像修改模型，開發自動個性化圖卡生成系統。

九、研究微調開源視覺 LLM，實現即時、低成本的專家模型 AphasiaLite-LLM 應用在系統中。

十、整合語音、肢體語言、唇語、對話場景、對話膠卷圖、圖卡與手繪白板等多模態資訊，建構失語症流暢溝通系統。

## 參、文獻回顧

### 一、失語症

失語症（Aphasia）是一種由大腦損傷引起的語言障礙，影響個體的言語表達、理解、閱讀和書寫能力。大多數情況下，這種損傷發生在大腦左半球的語言處理區域，原因包括中風、頭部創傷、腦腫瘤等。根據語言表現特徵，失語症大致可分為流利性失語症（Fluent Aphasia）與非流利性失語症（Non-Fluent Aphasia）。

（一）流利性失語症（Fluent Aphasia）：患者能夠流利地說話，但語言內容可能缺乏意義，詞語錯誤頻繁，甚至可能包含無關或自創的詞彙。此外，患者的語言理解能力通常受損，導致無法準確理解對話內容。

（二）非流利性失語症（Non-Fluent Aphasia）：患者在語言表達上遇到困難，說話緩慢、句子簡單且結構不完整，通常僅能使用少量詞語來傳達訊息。然而，他們的語言理解能力大多功能正常，因此能夠理解他人的話語。

其中，失語症若按照是否講話流利（Fluent）、是否能理解他人言語（Comprehend）、是否能重複一遍別人說的話（Repeats），還可細分為以下幾種 [1]。如下圖所示：

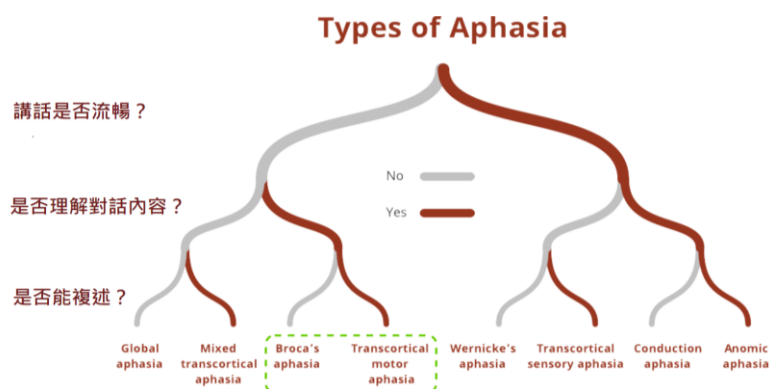


圖 3-1：失語症種類(修改自 National Aphasia Association [1])

而我們作品中所設計的系統，旨在將患者不流暢的話語，透過多模態架構，使其轉換成順暢的語句。因此，系統能夠幫助的對象僅限於非流暢型的失語症，也就是布洛卡失語症（Broca's aphasia）以及跨皮質運動型失語症（Transcortical Motor Aphasia）。

## 二、失語症治療與溝通輔助

### （一）輔助失語症的方式

1. 視覺輔助：患者雖難以閱讀與書寫，但通常能辨識並使用圖片。視覺輔助能營造有效溝通情境，當患者詞彙受限時，圖片可替代語言，減輕表達壓力。研究顯示，使用與患者相關的圖片能提升學習效果與對話流暢度。
2. 繪畫輔助：失語症患者常能透過繪畫來表達想法，即便難以找到合適的詞彙。相比語言或文字，繪畫對失語症患者而言更易理解，因而成為一種有效的溝通工具。
3. 手勢輔助：手勢無需設備或訓練，隨時可用，為簡便高效的溝通方式，雙方可透過手勢改善交流。常見手勢如：拇指上下示意同意或反對、點頭或搖頭表示肯否、手指指向方向，或模仿吃飯、喝水、打電話等動作表達需求。
4. 裝置輔助：對於溝通技巧不足的失語症患者，替代溝通（Augmentative and Alternative Communication, AAC）系統和語音產生裝置（SGD）是有效的輔助工具，幫助患者更順暢地進行溝通。

本研究主要目標是以手機 App 作為輔助設備，透過結合包含以上視覺（圖卡）、繪畫、手勢動作等輔助要素，嘗試解決失語症患者溝通問題。

### （二）對於失語症的自然語言處理

在 Manir et al.（2024）的論文中，使用 BERT 模型，幫助失語症患者進行文本補全與語意理解[2]。他們的研究在數據處理部分採用 Aphasia Bank 數據集。把資料集文本進行以下操作後用於訓練模型：去除不必要字符（刪除多餘符號及標點符號）、移除填充詞（如"uh"）、標準化空格（合併空格）、處理重複詞與錯誤詞（刪除多餘的結巴語氣）。

該研究邀請專家基於 Contextual Appropriateness、Grammatical Accuracy、Fluency 作為標準進行評分。然而，專家評分標準差較大，最大者達 0.98，顯示模型的泛化性能不佳，同時 Fleiss' Kappa 分數僅 0.32（該數值代表不同人之間評估分數的一致性，介於 0~1 之間），表示不同專家對模型輸出的看法具有較大差異，此分數僅略高於以隨機評分的狀況。這項研究

結果明確指出採用人為評估還原語句的方法並不可靠。本研究採用量化方法來評估語句還原的效果，突破以往依賴人工評估的方式，提升評分結果的客觀性與參考價值。評估指標包括 BERTScore、餘弦相似度（Cosine Similarity）以及編輯距離（Edit Distance），詳見肆、研究方法之二、模擬語句還原效果評估方法。

### （三）目前市面上的失語症輔助 App

傳統的失語症輔助工具大多透過圖像化的內容，讓使用者選擇圖像庫中的元素表達想法。Lingraphica 公司所推出的 SmallTalk Apps 點擊圖片即可播放常用詞句的語音，Proloquo2G 則由 AssistiveWare 公司開發，可將所選取的符號和文字生成一串自然語音並播放。然而，這種系統常會因為個人差異需要設計不同的資料庫，同時對話場景也受到侷限，圖像資料庫中的元素可能無法滿足於任何狀況。

因此，我們想解決傳統資料庫的侷限性，透過從對話的鏡頭畫面中擷取新物件，來實現個性化圖卡的自動生成。同時結合多模態的輸入來實作出智慧且創新的失語症輔助 App。

## 三、多模態大語言模型、Chain-of-Thought（思維鏈提示）與代理（Agent）

大型語言模型（Large Language Models, LLM）專為自然語言處理（Natural Language Processing, NLP）任務設計，通常採用 Transformer 架構，並透過大規模文本數據進行訓練。LLM 能夠執行文本生成、翻譯、對話等多種任務。多模態大型語言模型（Multimodal Large Language Models, MLLM）進一步擴展了 LLM 的應用範疇，能夠同時處理與理解文本、圖像、音訊與影片等多種模態資料。

Chain-of-Thought（CoT）是一種提升 LLM 推理能力的技術，透過引導模型逐步推理，將複雜問題拆解為一系列較為簡單的步驟[5]。這使模型能夠更清晰地理解問題邏輯，進而提升在複雜任務中的表現。

智能代理人（Agent）則是一種具備自主推理與行動能力的人工智慧模組，能根據任務需求主動決策與調度資源。結合大型語言模型後，Agent 更具語言理解與任務規劃能力，能執行多步驟任務、進行知識檢索，並具備語境感知等高階功能。

本研究運用 MLLM 整合關鍵畫面中所偵測出的手勢、表情、唇語、語音轉文字結果、對話場景及對話膠卷圖等多模態資訊，生成多個可能符合失語症患者意圖的語意表達。同時，透過思維鏈（Chain-of-Thought）技術，提升 MLLM 在語句生成方面的準確性與表達力。

#### 四、臉部表情識別

在 Azrien et al. (2024) 研究中，透過 Xception 架構來改進臉部表情辨識，測試準確率相比以往研究有顯著提升，最終達到 94.34% [6]。因此後續會使用 Xception 架構的模型實現表情辨識，本研究中主要用於檢測患者的情緒表達並修正還原語句以符合當下的對話情境。

#### 五、唇語識別

在 Ma et al. (2021) 的研究中，研究者改進端到端 (E2E) 視覺語音唇語識別 (AVSR)，提出基於 Conformer 的架構，能直接從音頻波形與圖像像素提取特徵，並融合 CTC/Attention 解碼機制，提高識別準確度[7]。該模型已成功應用於中文唇語資料集。後續研究將沿用該模型架構進行中文唇語識別。

#### 六、幀間差異算法（應用在關鍵幀提取）

表 3-1：多種幀間差異演算法（表格來源：作者自行整理）

方法名稱	核心計算方式簡述	運算基礎
Dense Optical Flow (Farneback)	將區域影像建模為二次多項式，透過相鄰影格多項式係數變化估算每像素的位移向量	多項式擬合 + 金字塔多層級光流估算
Frame Differencing	計算相鄰影格的像素差異，取絕對值後求平均作為運動分數	像素絕對差值平均
Block Matching	將影像分割為區塊，搜尋相鄰影格中最相似的區塊，計算絕對差值 (SAD)	區塊搜尋 + 最小絕對差距 (SAD)
Feature Tracking (Lucas-Kanade)	利用光流假設建立區域內的線性聯立方程，求解特徵點的位移向量	最小平方方法求解光流方程
Histogram Difference	計算每張圖的色彩直方圖後，用卡方距離評估兩影格間整體色彩分布的變化	直方圖 + 卡方距離
ORB 特徵比對	使用 FAST 偵測特徵點與 BRIEF 建立描述子，並計算 ORB 描述子的平均匹配距離	漢明距離比對二進位描述子
CLIP 嵌入特徵差異	使用 CLIP 將影像轉換為語意向量，計算相鄰影像間的餘弦相似度差異	Transformer 編碼 + 餘弦相似度

在實驗四中，我們將分析多種幀間差異演算法，權衡準確率與運算效率之取捨，藉此設計適用於即時語意判斷之優化關鍵幀提取機制。

## 七、LoRA (Lora rank adaptation)

在大型語言模型 (LLM) 的微調領域，LoRA (Low-Rank Adaptation) 是一種廣泛應用且高效的參數微調技術。它將預訓練模型的權重固定，僅在特定層中引入可訓練的矩陣，以顯著減少需要更新的參數數量。這種方法不僅降低了微調所需的計算資源，還能在保持模型性能的同時，實現更快的訓練速度和更低的記憶體需求。

在 LoRA 的基礎上，QLoRA (Quantized LoRA) 進一步優化了微調過程。QLoRA 透過將預訓練模型進行 4 位元量化，顯著減少了模型的記憶體佔用。此外，它引入了 4 位元的 NormalFloat (NF4) 數據類型和雙重量化技術，進一步提升了微調的效率。

## 八、實驗資料集

### (一) Visual Genome

Visual Genome 是由史丹佛大學於 2016 年發布的大型圖像語義理解資料集[9]。該資料集包含 108,077 張圖像，標註內容涵蓋語義細節。這些標註提供物體之間的空間關係與語義關係，例如「車-停在-街道上」，讓機器能學習影像中的互動模式。該資料集將作為生成模擬失語症語句時的對話場景資料。

### (二) HaGRID (HAnd Gesture Recognition Image Dataset)

HaGRID 是一個大規模的手勢識別數據集，專為人機交互設計[10]。該數據集包含 554,800 張影像，涵蓋 18 種帶有特定語義的手勢，可用於靜態手勢以及動態手勢的識別。該資料集將被用於訓練手勢分類器的模態。

### (三) 失語症資料集

Aphasia Bank 是一個共享的多媒體資料庫[12]。其中包含大約 180 名失語症患者和 140 名非失語症對照者，錄製具有統一任務腳本的話語影片。影片中的語言為 CHAT 格式，可以執行多種語言分析。該資料集將被用於關鍵幀算法實驗與最終系統臨床驗證。

## 肆、研究方法

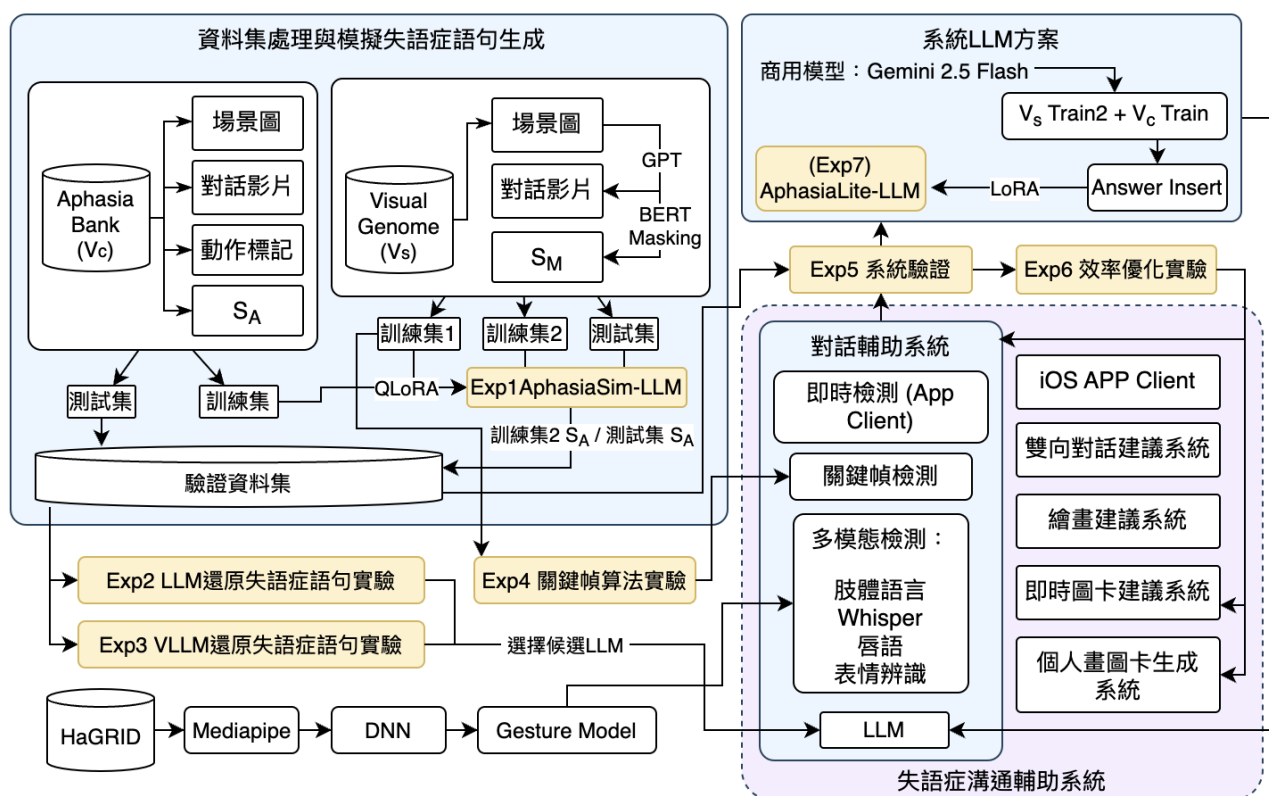


圖 4-1：系統與研究架構圖（圖片來源：作者繪製）

我們嘗試許多創新的實驗與系統架構設計，透過從這些過程之中得到的結果，我們成功地實現出一個輔助失語症患者流暢溝通的系統，以下是針對各項細節的說明。

### 一、失語症模擬資料生成

由於 AphasiaBank 失語症語料庫中的中文資料量有限，且僅包含失語症患者的語句標註，缺乏對這些語句所對應之真實語意的標註，亦即缺乏作為基準事實（Ground Truth）的對照資料，因而無法直接應用於我們所需之以「對話場景－失語症語句－真實語意」為結構的實驗設計。為了解決此一資料不足問題，我們參考了 Manir et al. (2024) 及 Misra et al. (2022) 對失語症者語言特徵的描述，設計一套結合低精度微調技術（QLoRA）的四階段失語症語句模擬方法。

為了微調出一個能生成失語症語句的大型語言模型（AphasiaSim-LLM），我們需建立一組大規模的語句對，內容包括流暢語句及其對應之模擬失語症語句。方便起見，我們將在後續內容中使用以下簡寫，如表 4-1。



表 4-1：本研究中不同類型的語句資料簡寫

中文	英文	簡寫
AphasiaBank 的臨床失語症語句資料	Clinic Aphasic Sentence Samples	S <sub>C</sub>
AphasiaSim-LLM 生成的模擬失語症語句	AphasiaSim Simulated Aphasic Sentence	S <sub>A</sub>
S <sub>C</sub> 人為標註還原為流暢語句的結果或從模擬失語症語句之演算法產生的流暢語句	Fluent Sentence	S <sub>F</sub>
失語症語句經過系統還原為流暢語句的結果	Restored Sentence	S <sub>R</sub>

### (一) 步驟一：標註 Aphasia Bank 失語症語句

本研究首先根據 AphasiaBank 資料庫中的錄影資料，人工標註出每句失語症語句所欲表達的流暢語句版本，形成臨床語料對。

### (二) 步驟二：產生模擬失語症語句，擴增樣本數

接著，我們設計演算法以擴增模擬資料，進一步用於語言模型微調。根據 Manir et al. (2024) 以及 Misra et al. (2022) 兩篇文獻，失語症語者常見語言表現包含：重複字詞、使用錯誤詞彙、以及自創詞。Manir et al. (2024) 論文中採用遮蔽 (masking) 技術處理這類語句，以下為其處理方式之整理 (見表 4-2)：

表 4-2：失語症說話語句預處理 (整理自 Manir et al., 2024)

Original Dataset	Pre-processed Dataset
&-uh well ‡ Cinderella was a poor child in &-uh [/] in &-uh &-uh +...	[MASK] well Cinderella was a poor child in [MASK] in [MASK] [MASK]
and &-um the mice (.) &-um give her the key.	and [MASK] the mice [MASK] give her the key

本研究則進行逆向操作並加以改良，設計模擬失語症語句之演算法，如圖 4-2 所示，我們以 GPT-4o-mini 模型為 Visual Genome 資料集中 (train 集) 之日常場景生成對話句 (即 S<sub>F</sub>)，假設這些句子即為失語症者原欲表達之語意。接著透過 jieba 函式，將句子分割成一連串具有獨立意義的中文字詞序列，並將其轉化成 token 形式，輸入 bert-base-chinese 預訓練模型，取得最後一層的注意力分數 (Attention)。該分數代表某個字詞對整句話語意的影響程度。將注意力分數序列 (A) 取負值，通過 Softmax 函數，生成一個概率分布 (P)。此用意是讓對句意影響較大的字詞有較小的機率替換掉，確保最後經過操作後的句子中保有足夠多有意義的字詞訊息，讓 LLM 能夠推測原始句子。根據概率分布 (P) 選擇部分比例的字詞進行重複操作，接著同樣以該概率分布選擇字詞進行挖空與替換錯誤詞。如表 4-3 所示，此演算法可產出具重複詞、錯詞與遮蔽現象的模擬失語症語句。

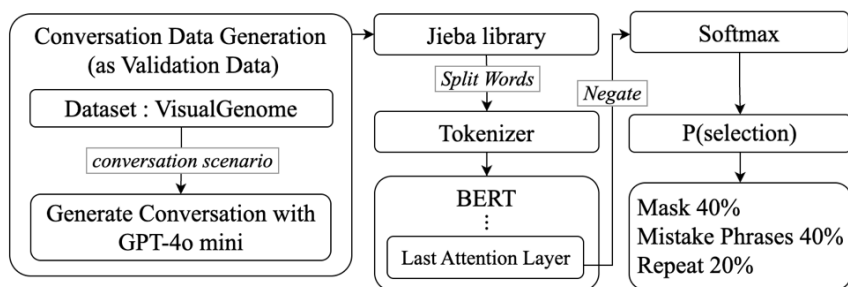


圖 4-2：  
失語症語句模擬之演算法流程（圖片來源：作者繪製）

表 4-3：模擬失語症語句範例（表格來源：作者自行製作）

流暢語句	模擬失語症語句
當然記得，那時候我們經常熬夜做報告，真是辛苦啊	[MASK][MASK]，[MASK]咱們經常[MASK]作做做報告[MASK]哎啊[MASK]！

### (三) 步驟三：微調 LLM，訓練模擬失語症語句的 AphasiaSim-LLM

以步驟一與步驟二所得之  $S_C/S_F$  對與  $S_A/S_F$  對，以 QLoRA 微調開源小參數語言模型，使其具備將流暢語句轉換為模擬失語症語句的能力。該模型於本研究中稱為 AphasiaSim-LLM。


### (四) 步驟四：訓練 Naïve Bayes，驗證 AphasiaSim-LLM 微調效果（實驗一）

為驗證模擬失語症語句的品質，我們參考 van Vaals et al. (2024) 所提出的方法[14]，透過訓練 Naïve Bayes 分類器來比較真實失語症語句（ $S_C$ ）與模擬語句（ $S_A$ ）之間的相似程度。若模擬語句愈接近真實語句，則分類模型的辨識準確率應愈低，表示兩者難以區分。

然而，原文僅提及使用 Naïve Bayes 進行分類，並未明確說明所採用的具體模型類型。因此，實驗一針對三種常見的 Naïve Bayes 模型進行測試，包括：BernoulliNB、MultinomialNB 與 GaussianNB。

最後以 GPT-4o-mini 為 Visual Genome 資料集中的日常場景生成  $S_F$ 。表 4-5 是我們希望透過實驗得出的 AphasiaSim-LLM 產生的失語症資料集，包含對話場景、 $S_F$  與其對應的  $S_A$ 。

表 4-5：資料集範例（表格來源：作者自行製作）

流暢語句（ $S_F$ ）	對話場景
這個房間看起來有點亂，需要好好整理一下。 是啊，我的書桌上堆滿了東西，都沒地方放了。	
模擬失語症語句（ $S_A$ ）	
這 房 間 看 起 來 看 來 沒 有 整 理 是 啊 ， 我 的 上 都 堆 了 東 西 ， 都 了 沒 有 放	

## 二、模擬語句還原效果評估方法

本研究旨在協助失語症患者，將其所表達的語意轉換為流暢且完整的語句，以促進其與他人的有效溝通。為了評估 LLM 將失語症語句 ( $S_A$ ) 還原為標準語意語句 ( $S_F$ ) 的能力，我們設計了一套驗證流程。過去相關研究大多仰賴人工評分的方式來判斷模型的還原效果，然而此方法不僅耗時費力，且如文獻所示，專家評分的一致性偏低，其結果往往與隨機評分相差無幾，顯示其可信度與參考價值相對有限。

為突破此一限制，本研究創新性地導入自動化且量化的評估機制，透過三種語意相似度量指標，客觀比較 LLM 產出的還原語句 ( $S_R$ ) 與對應的流暢語句 ( $S_F$ ) 之間的相似性。據我們所知，過去尚未有研究在失語症語句還原任務中採用此類自動量化方法，顯示本研究在方法論上的前瞻性與貢獻。

### (一) BERTScore

BERTScore 是基於 BERT 語言模型的文本相似性評估方法。與傳統的 BLEU 或 ROUGE 方法不同，它不依賴於詞彙匹配，而是通過詞嵌入的相似性來評估句子之間的關聯。BERTScore 計算原始句子與模擬語句還原結果之間的詞向量匹配分數，主要指標包括 Precision、Recall、F1-score。

### (二) Cosine Similarity

BERT 的 [CLS] token 在最後一層的隱藏狀態可作為整個句子的語意表示。也就是透過 BERT 將整個句子轉換為高維度向量，以 [CLS] 向量作為整體句子的語義代表。以此向量作為計算原始句子與模擬語句還原結果餘弦距離的依據，評估兩個句子之間的差異。

### (三) Edit Distance (編輯距離)

Edit Distance (Levenshtein Distance) 是一種基於字串編輯操作的相似度衡量方法，用於計算將一個字串轉換為另一個字串所需的最小編輯步驟數。Edit Distance 允許三種基本操作：插入 (Insertion)、刪除 (Deletion)、替換 (Substitution) 透過動態規劃，可以高效計算 Edit Distance，時間複雜度為  $O(mn)$ ，其中  $m$  和  $n$  分別為原始語句和模擬語句還原結果的長度。

## 三、LLM 還原模擬失語症語句 ( $S_A$ ) 的實驗 (實驗二與實驗三)

此步驟中我們設計了兩個實驗 (實驗二和實驗三) 來驗證不同 LLM 將  $S_A$  還原為  $S_R$  時的效果。實驗二中我們僅提供  $S_A$  並由 LLM 生成  $S_R$ 。而實驗三中，我們不只提供  $S_A$ ，同時提供 Visual Genome 中的對話場景，讓 Vision LLM 參考者兩種資料，生成  $S_R$ 。最後我們將  $S_R$

與  $S_F$  透過 BERT Score、Cosine Similarity 以及 Edit Distance 的指標比較各模型的語意復原能力，找出在還原模擬語句時表現最佳的 LLM。在這個實驗中，我們分別驗證了商用 LLM 與開源小型 LLM，兩者在實際應用時各有優劣，商用 LLM 提供更快速且準確的回答，然而他們多不支援微調或者 RLHF 的強化學習，若使用開源小模型，則蒐集系統實際使用時的數據對模型進行微調，進一步提升其在還原失語症語句任務上的能力。同時，由於無需承擔 API 呼叫的費用，在大規模部署時將會是一個更低成本的選擇。

#### 四、訓練手勢分類模型

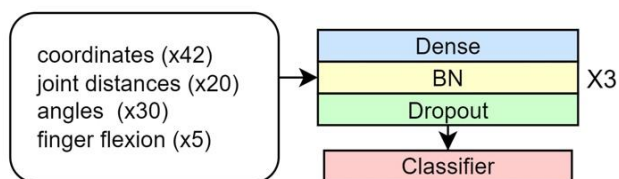


圖 4-3：手勢分類模型架構  
(圖片來源：作者繪製)

本研究所開發之手勢分類模型採用 HaGRID 資料集訓練，採用自行設計的深度神經網路 (Deep Neural Network, DNN) 架構，如圖 4-3。不同於現有研究多仰賴現成模型 (如 ResNet、MobileNet 等) 或開源專案進行遷移學習，我們強調模型設計的輕量與準確性，以符合系統即時計算的需求。特徵提取方面，透過 MediaPipe 取得 21 個手部關鍵點，並轉換為相對座標系以消除不同使用者與攝影角度的影響，然後計算關鍵點間的距離、關節角度以及手指彎曲度等結構性幾何特徵作為輸入。這些特徵再輸入至四層結構的深度神經網路中進行分類。

本模型支援辨識 HaGRID 定義的 15 種常見手勢，包括：call、dislike、fist、four、like、mute、ok、one、palm、peace、peace\_inverted、rock、stop、stop\_inverted 以及 three 等，具備高度實用性。經訓練後，模型整體準確率 (accuracy) 達 0.8613，精確率 (precision) 達 0.8872，F1 分數亦達到 0.8670，展現出優異的分類表現。

#### 五、關鍵幀提取技術比較實驗 (實驗四)

為了提升系統在處理影片時的性能，我們將在多模態辨識之前設計一個關鍵幀提取的步驟，提取出部分的影片幀，減少後續多模態系統的計算量。為了在關鍵幀提取技術上盡可能符合臨床情況，我們特別採用 Aphasia Bank 的失語症資料集作為實驗資料集來評估不同關鍵幀算法的效果。

Aphasia Bank 資料集包含了失語症患者與研究員的對話內容，其文檔標記 (.cha) 中包含  $S_C$  (Clinical Sentence Samples) 以及對話過程中失語症患者的肢體動作。然而這個資料集的肢體動作並不包含詳細的時間訊息，僅標記在患者說出某句話時做出什麼動作。如表 4-6 所示。

表 4-6：Aphasia Bank 資料集範例（整理自 Aphasia Bank [12]）

Aphasia Bank	
*INV（研究員）：	你 可 以 一 邊 看 圖，一 邊 說 給 我 聽 .37178_39963
*PAR（患者）：	&=left_index_finger_pointing_at_picture, 這 [/] 這 個 小 孩 兒 在 踢 [/] 踢 走 [: 足][*] 球 .39963_48605

根據時間戳記，將失語症患者說話的整段影片進行裁切，使單一影片片段僅包含失語症患者說的一句話，並且提取出所有開頭為 &=... 的動作敘述形成動作序列  $V_A$ 。我們將該影片片段每 0.2 秒取一幀，形成一個幀序列  $V_F$ ，由 Vision LLM 對  $V_F$  與  $V_A$  兩個序列進行交叉評分，最終透過峰值檢測精確定位出特殊肢體語言的確切幀  $V_{RA}$ （Reference Action Location）。如圖 4-4 所示。

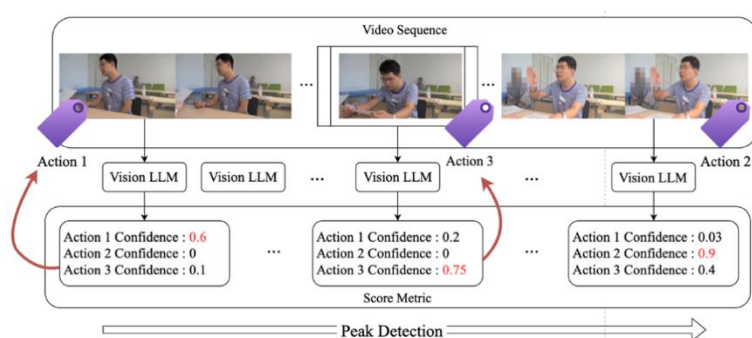


圖 4-4：  
動作確切幀  $V_{RA}$  計算方法  
（圖片來源：作者繪製）

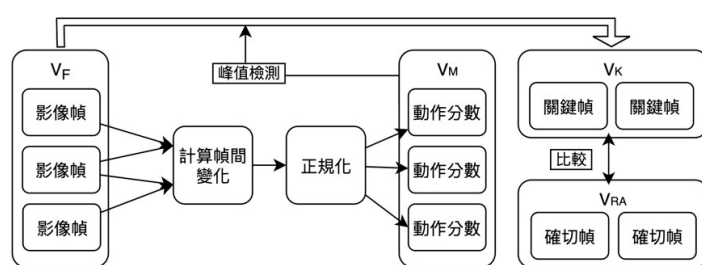


圖 4-5：  
關鍵幀  $V_K$  計算方式  
（圖片來源：作者繪製）

如圖 4-5 所示，下一個步驟中，以不同的關鍵幀算法計算出  $V_F$  各幀的分數，並正規化到 0~1 的區間，作為運動分數序列  $V_M$ 。最終以 0.15 作為閾值，從影片中提取出關鍵幀  $V_K$ 。透過比較不同關鍵幀算法的  $V_K$  與  $V_{RA}$  的相似性，允許  $V_K$  與  $V_{RA}$  之間正負一幀的誤差，計算  $V_K$  與  $V_{RA}$  的交集數量，再除以  $V_{RA}$  的長度作為準確度。使用以上方法，我們總共實驗了 Dense Optical Flow( Farneback )、Frame Differencing、Block Matching、Feature Tracking( Lucas-Kanade )、Histogram Difference、ORB（Oriented FAST and Rotated BRIEF）+ 暴力匹配 和 CLIP（Contrastive Language - Image Pre - Training）等八種方法。

## 六、系統設計與實現

### （一）系統目標

目前市面上多數失語症輔助系統採用「圖卡接龍」的方式，雖具有幫助，但理解語境能

力有限，且操作繁瑣、學習門檻高。我們的系統整合語音、圖像、手勢等多模態輸入，提升溝通流暢度，具體改進包括：

### 1. 語意還原與操作簡化：

傳統系統要求患者逐字選擇形容詞、助動詞、介詞等語法成分來組合句子，難以解決失語症患者常省略這些詞彙的問題。對此，我們的系統導入 LLM 聚焦於動詞與名詞等語意核心，並整合各個模組，讓使用者只需「說」、「選擇」或「畫」關鍵概念即可完成溝通，減少對複雜的圖卡分類與句型結構的使用。

### 2. 個人化圖卡自動生成：

依據 Aphasia Library 的建議，輔助圖卡應貼近患者日常生活 [8]。然而目前多數系統仍仰賴手動上傳圖片來建立圖卡，操作繁瑣。對此，我們系統可根據對話與環境，自動擷取新物件並生成個人化圖卡，貼近患者的日常，提升便利性與真實溝通效果。

### 3. 多維環境感知與語意選項生成：

多數現有輔助系統僅依賴單一模態（如圖像或文字），難以提供語句理解與語意推理的能力。我們整合語音、圖像、手勢等輸入方式，由 LLM 分析並整合，產出合理的語意選項來輔助。同時，也能根據說話對方（Counterpart, CPT）的對話語境，自動預測並推薦患者（Patient, PAT）可能欲表達的下一句話，進一步降低使用者表達負擔。

## （二）系統設計

如圖 4-6，本系統包含手機 App 與 Flask 伺服器協作運行，並包含四大重要功能。

### 1. 即時建議：即時圖卡建議系統與繪圖建議系統

伺服器根據對話場景與歷史，預測患者可能表達的內容，提供相關圖卡。患者可用手寫、手繪輔助表達，並藉由繪圖建議系統生成繪圖建議，已確保與失語症患者想表達的語意一致。

### 2. 多模態對話輔助（對話輔助系統）

iOS App 透過後鏡頭記錄對話場景、前鏡頭錄製患者發言，並整合圖卡、手繪、手寫資訊，上傳至伺服器進行多模態分析。系統解析語意並提供數個選項，讓患者選擇最貼近他意圖的表達，最終由 App 代為發聲。

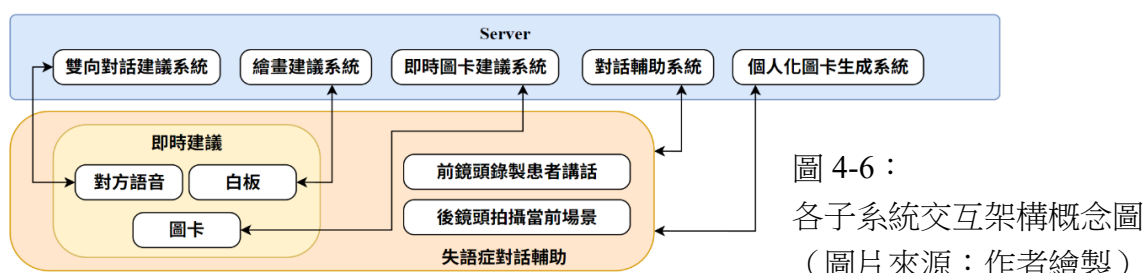
### 3. 雙向對話建議系統

伺服器會根據對方的說話語音與對話上下文關係，生成數個失語症患者可能的回覆選項，讓患者可以不用主動表達，直接選擇現成的語句進行回覆。



#### 4. 個人化圖卡自動生成

我們的系統也會從手機 App 過去上傳的鏡頭畫面中自動識別物件，透過 LLM 生成該圖卡的配合文字，並自動保存該圖卡，將個人化圖卡的自定義自動化。



#### (三) 系統效果驗證實驗（實驗五）

為了驗證本系統的實際成效與臨床應用潛力，我們分成兩部分進行，如圖 4-7。

**第一部分**延續先前 Visual Genome 資料集的處理方法，設計 15 組患者與正常人的對話，並標註表情與手勢。接著我們以自拍方式錄製前鏡頭影片，產出含有語音、動作與情境資訊的模擬驗證資料集（Validation Set of Simulate Samples, 簡稱  $V_s$ ）。相關內容如表 4-7 所示。評分方法與實驗二、二相同。

**第二部分**則針對系統的臨床應用效果。我們使用 Aphasia Bank，擷取患者的對話影片與文字標記，取得臨床語句樣本（ $S_c$ ）及肢體動作標註。由於該資料集中並未提供真實語意（ $S_F$ ），無法直接評估。因此我們以三位測試者分別觀察每段  $S_c$  對應的說話影片，人工撰寫對應的  $S_F$  作為系統評估。在驗證階段，我們將系統生成結果（ $S_R$ ）與三位測試者所寫的  $S_F$  計算相似度分數，取最高分作為該句的語意還原評分。此包含對話場景、 $S_c$ 、人為標註  $S_F$  及動作標示的臨床驗證集（Validation Set of Clinical Samples）在後續研究中簡稱為  $V_c$ 。

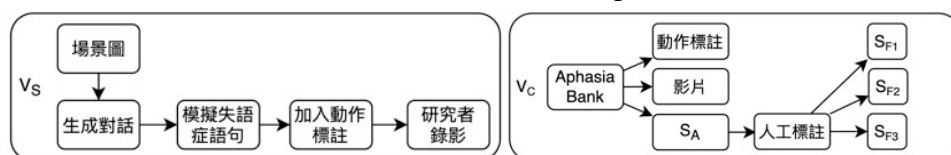



表 4-7：進行系統驗證範例（表格來源：作者自行製作）

$S_F$	$S_A$	肢體動作	說話錄影與場景
PAT：你好，我想點一杯珍珠奶茶。	你好 我想點 珍。	指向前上方	
CPT：當然可以！我們的甜度有無糖、三分糖、半糖、七分糖和全糖。	嗯 那 好了。	比出「3」，開心表情	
PAT：嗯那我還是選三分糖好了。	那 給 冰 冰吧	比出「不」的手勢	
CPT：好的，甜度三分糖，那冰塊的部分呢？			
PAT：那就給我去冰吧			

#### (四) 手機 App 運作詳細流程

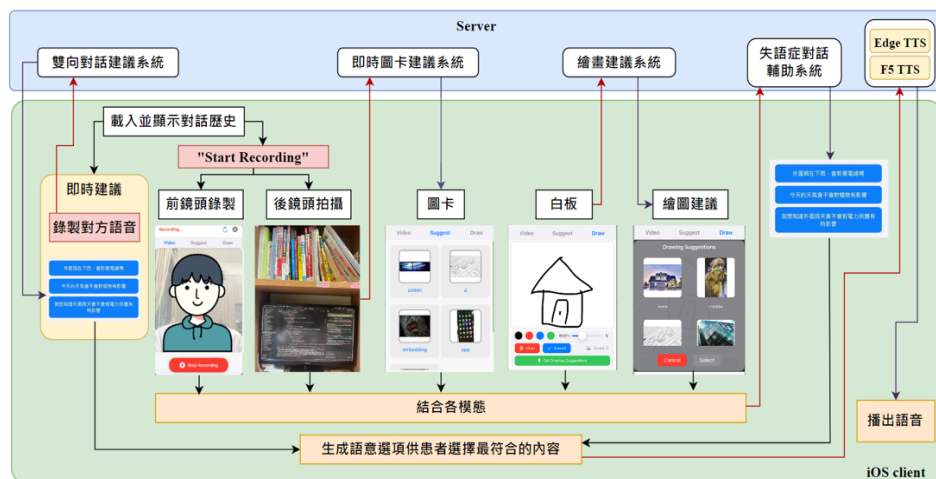


圖 4-8：  
手機 App 運作流程圖  
(圖片來源：作者繪製)

如圖 4-8，手機 App 提供兩種模式：

##### 1. 被動對話模式

App 自動錄製對方語音，上傳至伺服器的雙向對話建議系統。該子系統根據該錄音及對話歷史，生成患者可能想回應的語句，顯示於 App 介面中。患者從建議語句中選擇最符合其意圖的選項，然後伺服器運行 TTS（文字轉語音）模型，讓 App 朗讀，協助患者對話。

##### 2. 主動敘述模式

App 自動啟動後鏡頭拍攝當前對話場景，再切換至前鏡頭錄製患者的說話過程（這部份顯示於 Video 頁面）。與此同時，場景圖會上傳至伺服器的圖卡即時建議系統，該子系統分析場景與對話歷史，從患者可能想表達的名詞與動詞，生成對應的圖卡並回傳 App。

若患者自己無法順利表達詞彙，可以進入 Suggest 頁面，從圖卡中選擇物件或動作來輔助溝通。若無呼應的建議圖卡，患者還可在 Draw 頁面的白板手寫文字或繪製圖像表達意思。

對此，患者可選擇直接提交畫作，或使用繪畫建議系統辨識並提供選項，從而確認繪製內容的含意。最終，多模態資訊將整合並發送至伺服器，經由對話輔助系統分析，生成語意選項供患者選擇最符合其意圖的選項。最後，伺服器運行 TTS 模型，讓 App 朗讀該語句，協助患者完整表達想法。

##### 3. 模式自動切換

在人際對話中，多數交流採一問一答的模式。因此，當使用者選擇主動敘述模式，並由 App 朗讀患者選擇的語句後，系統將自動切換至被動對話模式，開始錄音並等待對方的回應。

當對方回覆結束後，App 會自動啟動雙向對話建議系統，根據對話內容生成患者可能想要回應的句子。患者可從建議的語意選項中選擇最符合其意圖的回應，作為下一句話。若建



議選項中沒有符合患者需求的語意，患者亦可手動切換回主動敘述模式，以自由表達自己的想法，確保溝通的流暢與靈活性。

### (五) 圖卡即時建議系統

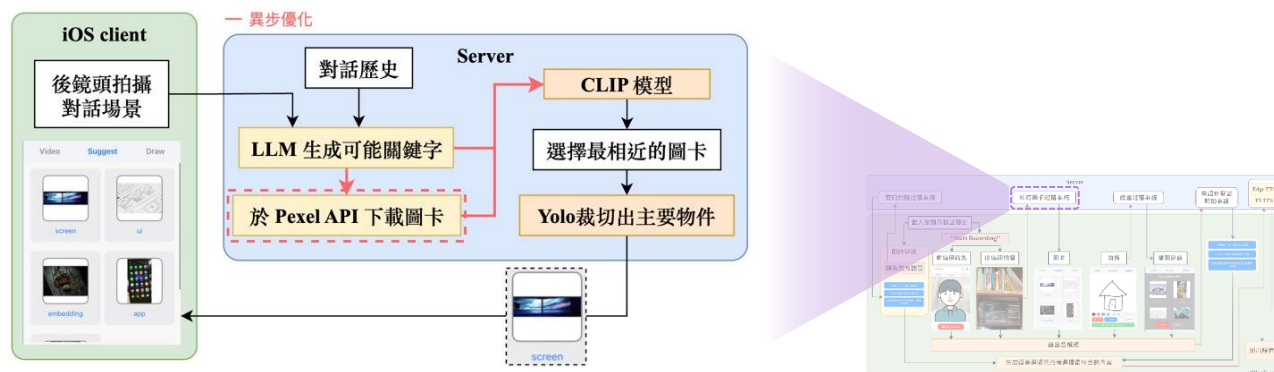


圖 4-9：圖卡即時建議系統（圖片來源：作者繪製）

本系統的主要作用是在患者開始講話時，預測可能會講到的元素或者動詞，如圖 4-9。本系統會將場景圖結合對話歷史，交由 Gemini 2.5 Flash 產生患者可能會說的關鍵字，針對這些關鍵字在 Pexel 圖片庫中找尋與關鍵字相符的圖片，對於每個關鍵字都會下載三張圖片，然後透過 CLIP 模型編碼對應的關鍵字與圖片，計算關鍵字編碼與其對應的三張圖片的編碼之間的餘弦距離。另外也會計算圖片的 Laplacian Variance 值，並透過以下公式：

$$\text{Sharpness Score} = \text{Min} (1.0, \ln(1 + \text{Laplacin Variance}))$$

$$\text{Image Suitability} = 0.2 \times \text{Sharpness score} + 0.8 \times \text{Cosine Similarity}$$

比較三張圖片的 Image Suitability，選出最高分的圖片，用 Yolo 辨識物件並裁切出來，搭配圖片做成圖卡。將圖卡回傳至手機 App 端供使用者選擇。本系統採用 Gemini 2.5 Flash，可在患者啟動主動對話模式時迅速產生可用圖卡，提升選擇效率，此設計以效率為核心考量。

### (六) 繪畫建議系統

本系統類似於圖卡即時建議系統的架構，然而後端我們接入輕量的 Flux 文生圖模型，讓系統可以更精準地按照使用者的繪畫產生圖卡，以符合使用者想表達的語意。如圖 4-10。

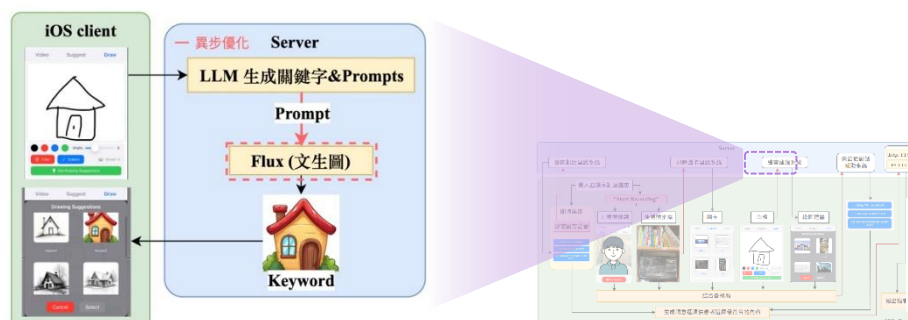


圖 4-10：繪畫建議系統（圖片來源：作者繪製）

## (七) 雙向對話建議系統

當 CPT（對方）發話時，iOS 端會錄製語音並傳送至伺服器，由 Whisper 轉換為文字，接著結合對話歷史，由 LLM 解析內容並生成患者可能的回應，顯示在手機 App 上供患者選擇。患者確認語句後，伺服器運行 TTS 模型將其轉換為語音，並由手機揚聲器播放。TTS 提供兩種選擇，一種是採用 F5TTS，語音合成時間較長，但準確度高；一種是採用 Edge TTS 的文字轉語音，透過 edge 瀏覽器合成語音，響應速度快，但聲音較機械。無論使用者選哪種方法，語音皆於伺服器端生成後發送至 App。TTS 模型可以在 App 的設定頁（Setting）中選擇。

## (八) 對話輔助系統

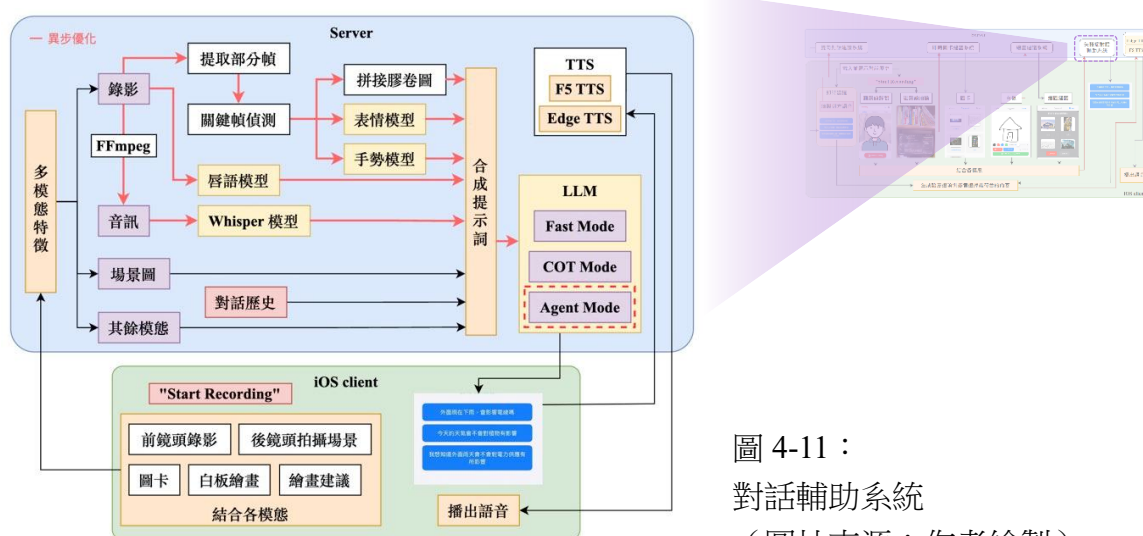


圖 4-11：  
對話輔助系統  
（圖片來源：作者繪製）

當患者說話結束時，iOS App 彙整使用者在主動敘述模式中使用的以下五個模態資訊：

1. 前鏡頭紀錄的患者說話過程錄影
2. 後鏡頭紀錄的對話場景圖
3. 患者在即時圖卡系統所生成的圖卡中所選擇的，與其語意意圖相關的圖卡
4. 患者在白板中的手寫或者手繪內容，直接提交者
5. 患者選擇透過繪畫建議系統的語意確認圖卡

將這些模態融合後上傳至伺服器，伺服器會使用 FFmpeg 提取患者說話錄影中的音訊，並透過 Whisper 模型進行語音轉文字。同時說話錄影也會透過 E2EAVSR 進行中文唇語識別。

說話錄影將根據 0.2 秒的採樣週期，提取影片中的部分幀，以減少後續計算量。接著以 ORB 進行關鍵幀提取（關鍵幀的效率與準確度分析請見實驗四的研究結果）。將關鍵幀以 Xception 表情識別模型及我們訓練的手勢識別模型，並形成一個與關鍵幀對應的特徵輸入序列。演算法會將相鄰且標籤相同的項目進行合併，並捨棄標籤相同但持續時間過短的項目，

最終形成一個具有時間順序的表情與手勢特徵序列。與此同時，關鍵幀也會被橫向拼接成一個膠卷圖，作為 VLLM 的其中一個輸入，提供 VLLM 更多的對話過程細節。

最後所有模態的特徵檢測結果將被統一嵌入到 Prompt（提示詞）中，交由 LLM 嘗試還原失語症語句，我們的系統提供三種 Prompt 與 LLM 模式，包含：快速響應模式（Fast Mode, 使用簡單的 Prompt 使 LLM 快速回答）、思維鏈模式（COT Mode, 使 LLM 深度分析各模態，以提升  $S_R$  的還原效果）、代理人模式（Agent Mode, 使用四個 Agent 代理，將語句還原任務進行拆解，由不同代理人分別負責一部分的工作，以提升  $S_R$  的還原效果）。

最終患者的手機上將會顯示 LLM 推測的數個可能語意，由使用者選擇最接近其想表達的語意後，運行 TTS 模型，由 App 唸出該語句。本系統中使用 Gemini 2.5 Flash 作為 LLM，我們也有附上前代系統中使用 Gemini 2.0 Flash 與 Grok2 的實驗數據，詳細實驗數據請見研究結果。

#### (九) 思維鏈思考流程

整體流程共分為七個步驟，依序引導模型從不同角度解析失語症患者的溝通意圖。**第一步**，從對話歷史中分析對方的最後一個提問類型，判斷其為開放式、是非題或簡短回答，有助於確立語句結構方向；**第二步**，檢視患者的語音內容，辨識關鍵詞語元素與大致語意；**第三步**，解析第一張場景圖像，了解對話背景與語境線索；**第四步**，觀察患者的行為膠卷，捕捉其手勢與視線等非語言表達；**第五步**，結合系統提供的情緒與手勢描述，判斷語氣與情感傾向；**第六步**，若有手繪圖像，進一步分析圖像中是否藏有與語意相關的圖形符號或情境提示；最後，**第七步**，透過患者所選的卡片內容進行語意確認，補足語言表達的不足之處。每一模塊聚焦特定資料來源，層層推理並整合資訊，最終生成五句通順且忠於患者原始意圖的繁體中文句子，同時嚴格限制不得添加未出現的內容，確保語句語法完整、語意真實。

#### (十) 代理人架構

我們提出的系統採用多代理框架，由四個專門的代理（Agent）組成，旨在全面分析和解讀患者想表達的語意，協同提升失語症患者對話輔助效率。**要素分析代理**從對話歷史、語音與圖卡中提取主題與關鍵詞，濃縮出不超過 100 字的語意分析，揭示患者意圖。**互動分析代理**則透過關鍵幀與場景圖解讀非語言行為，評估手勢與場景間的關聯。**關係分析代理**進一步整合情緒與手勢序列，判斷表達語氣並分析圖卡對情緒傳達的強化效果。**最終句子生成代理**整合前三者分析結果與語音、場景資訊，輸出五句忠實反映語境與意圖的最佳建議句，協助患者更準確表達。

### (十一) 個性化圖卡生成系統

系統透過 Yolo 檢測對話歷史的場景與影片中出現的物件。裁切後上傳至 Gemini 2.5 Flash 生成對應的關鍵字與修改建議，並透過 Gemini-2.0-Flash-exp-image-generation 依據修改建議修正圖片視角、大小並提升畫質。由於 Yolo 裁切物件不易控制其視角、畫質與長寬比例，而系統須確保圖卡為 1:1 比例，引入此功能可有效解決此問題，提升圖卡生成品質。如圖 4-12。

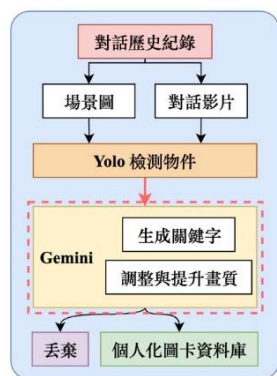


圖 4-12：個性化圖卡生成系統

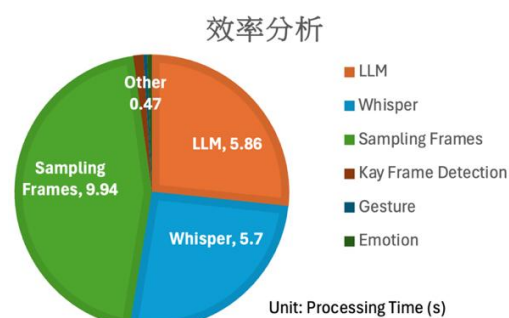


圖 4-13：初代系統效率分析

(圖片來源：作者繪製)

## 七、系統優化策略（實驗六）

由於本研究旨在幫助失語症患者流暢與他人溝通，系統的即時性至關重要。本章將重點討論如何提升系統的計算效率與性能，以實現最佳的即時反應和使用體驗。

### (一) 影片取幀優化

圖 4-13 顯示，對話輔助系統中，影片取幀（圖中 Sample Frames）的過程耗時最久。經過效率分析，我們發現主要原因是 OpenCV 函數在影片解碼時速度較慢。因此，我們將影片取幀的步驟改為使用 FFmpeg 搭配暫存資料夾，以提高取幀計算效率。

### (二) 異步優化

從圖 4-13 可以看出，除了影片取幀（Sampling Frames），Whisper 和 LLM 步驟的計算耗時也較長。這兩個步驟涉及呼叫 API，需要等待 API 的回覆，從而占用了較多處理時間。為了提升計算效率，我們在 FFmpeg、Whisper、Agent 與關鍵幀提取都引入了異步優化，如前文的圖 4-9 至 4-12 中以紅線標示處。

### (三) 輕量 Flux 文生圖與圖像資料庫優化

即時圖卡系統的功能是在按下「開始錄製」時，根據對話場景與對話歷史即時生成圖卡。最初設計透過 Pexels 圖像資料庫提供，但測試發現 Pexels API 回應時間極不穩定，且常常出現查無圖片或圖片解析度差的問題。即使加入異步處理，處理時間仍然偏長。因此，為提升

計算效率和圖像品質，我們結合個人圖像資料庫與輕量級 Flux 生成模型。這樣的優化方法使系統可直接從資料庫中取得已生成的圖卡，同時 Flux 模型加速系統獲得圖卡的時間，以減少因為擴散模型計算量大所帶來的延遲。

#### (四) 停用唇語識別模型

在實際測試中，由於 E2E AVSR 模型是基於央視主播的播報影片進行訓練，然而我們發現該模型在實際應用中泛化效果不佳，多數情況下識別出的唇語並不符合患者的敘述，反而可能誤導系統做出錯誤的語意猜測。然而自行訓練唇語模型對算力的要求極高，實現上較為困難，因此在最後的系統驗證階段我們暫時停用該模態，留待未來發展該功能。

#### (五) 伺服器資源分配與並行處理

由於伺服器同時運行五個主要系統，任務佇列順序可能導致手機 App 的額外響應時間。以下是由佇列順序引起的延遲情況：

第一種情況如圖 4-14-1 所示，當患者在手機 App 中選擇使用繪畫建議系統的功能時，伺服器需要完成該系統的計算。然而，如果在伺服器回應前，患者主動中斷錄影並提交影片來呼叫對話輔助系統，這時，由於伺服器的任務佇列中仍在運行繪畫建議系統，患者將需要等待額外的時間才能接收到對話輔助系統的語意選項。

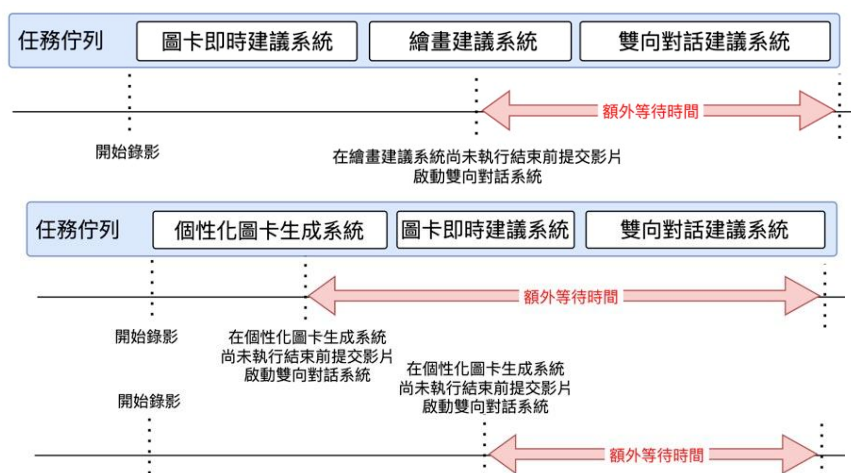


圖 4-14-1：  
伺服器任務佇列舉例一  
(圖片來源：作者繪製)

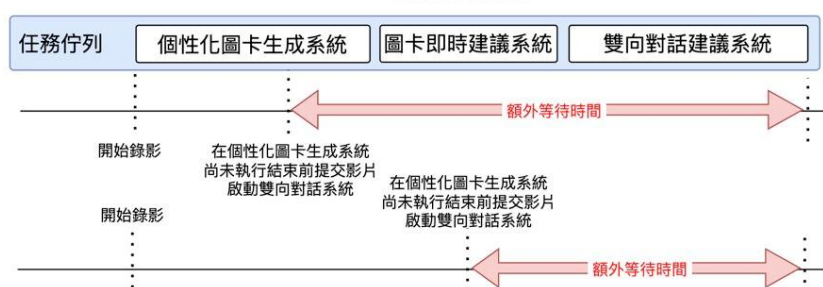


圖 4-14-2：  
伺服器任務佇列舉例二  
(圖片來源：作者繪製)

第二種情況如圖 4-14-2 所示。如果患者在即時圖卡系統尚未完成響應，或是在伺服器自動運行個人化圖卡生成系統時，主動中斷錄影並提交影片來呼叫對話輔助系統，也會因為任務佇列的順序問題，導致額外的計算等待時間。

儘管 Flask 伺服器支持多線程模式，但實測中常導致 API 呼叫錯誤，影響後續計算流程。因此，我們將五個子系統任務分配至三個佇列（如圖 5-18 所示）進行並行處理。被動與主動敘述模式共享一個佇列，因為不會同時發生；即時圖卡與繪畫建議系統共用 API 接口，因此



安排在同一佇列，以避免 API 錯誤；個人化圖卡生成系統因計算時間較長，單獨分配一個佇列，避免影響其他任務。

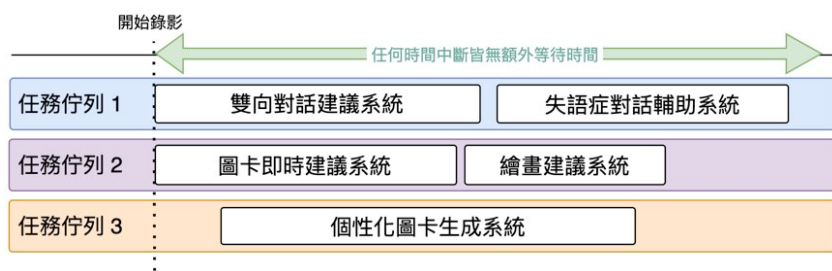


圖 4-13：  
伺服器任務佇列優化  
(圖片來源：作者繪製)

## 八、微調輕量開源模型（實驗七）

在實驗二、實驗三及系統效果驗證中（見研究結果第二點），我們特別關注輕量級開源模型 Qwen2.5 VL 3B。實驗顯示，其響應速度與 Gemini 2.5 Flash（Fast 模式）相近，F1-score 約低 0.07，Cosine Similarity 則約低 0.055。基於其輕量、開源與高速響應等優勢，我們認為該模型具備成為低成本部署選項的潛力。因此，我們以 Qwen2.5 VL 3B 為基礎，進行微調並打造出專為失語症溝通輔助設計的輕量模型 AphasiaLite-LLM，以提升其在實際應用中對失語症語句的還原能力。

與實驗二、二中僅使用  $S_A$  與  $S_F$  作為輸入輸出不同，在本應用場景中，輸入提示詞需結合系統實際運行時的多模態識別結果與對話情境；而輸出則需產生多個可能的  $S_F$  還原選項，供患者在手機應用中進行選擇。為此，我們採用類似資料集蒸餾(Dataset Distillation)的策略，以 Gemini 2.5 Flash 的輸入與輸出結果為基礎對 Qwen2.5 VL 3B 進行微調。為進一步強化微調成效，我們對 Gemini 模型的輸出進行如下處理：

- (一) 系統要求 LLM 產出五個可能的還原語句，並於手機 App 上提供給患者選擇。
- (二) 將這五個語句與  $S_F$  計算 F1-score，挑選出 F1-score 最低者，以 Ground Truth（即  $S_F$ ）替換之，確保每筆資料中至少包含一個正確答案。
- (三) 依照五個選項的 F1-score 高低進行排序，使資料具備語句品質的優劣順序。
- (四) 結合多模態資料、對話場景與膠卷圖等資訊合成微調用資料集。

經上述處理後，重組資料集以符合排序關係，並以此進行 LoRA 微調，期能顯著提升模型在真實場景中的應用表現。微調結果請見研究結果。

## 伍、研究結果

## 一、實驗一：微調 AphasiaSim-LLM 模擬失語症語句的效果

微調不同模型後，我們針對模擬語句進行分類實驗的結果，並與 **fluent vs aphasia** 對照組（ $S_F / S_C$ ）分類準確率做比較，如表 5-1 所示。

表 5-1：不同 AphasiaSim-LLM 的分類器訓練結果（表格來源：作者自行製作）

Accuracy (Score)	previous paper	llama3.2-3b	llama3.1-8b	qwen2.5-7b	qwen2.5-14b	Mistral v0.3-7b	fluent vs aphasia
BernoulliNB	0.83/0.81	0.6840	0.6558	0.6321	0.5892	0.6769	0.6965
GaussianNB	(experimental	0.5982	0.5756	0.5508	0.6038	0.9323	0.9117
MultinomialNB	/control)	0.6806	0.6512	0.6535	0.6738	0.7246	0.6978

從實驗一的結果可見，**qwen2.5-7b** 在多項評量中表現最佳，尤其在 **GaussianNB** 模型中，其 accuracy 降至 0.5508，相較 **fluent/aphasia** 對照組之 0.9117，下降達 36.8%，顯示模擬語句與真實語句之間的差異性低、相似性高。相對地，**van Vaals** 等人所提出的方法，在其所訓練之 **Naïve Bayes** 模型於實驗組（0.83）與對照組（0.81）之間差異有限，甚至出現模擬語句較易被識別的現象，進一步凸顯本研究所提出方法之效益與貢獻。

## 二、實驗二：純文字資訊下，LLM 還原失語症語句效果

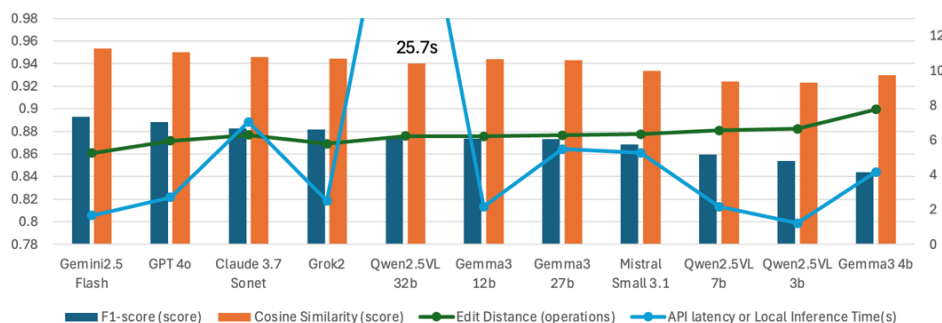


圖 5-1：  
LLM 還原模擬語句  
的結果評估

（圖片來源：作者繪製）

從圖 5-1 的實驗結果可知，各 LLM 都具備一定的語句還原能力。在五種指標中，F1 為 Precision 和 Recall 的平衡指標，在衡量模型效能時較為看重 F1 表現；Cosine\_Similarity 將文本轉換為向量後計算相似度，適合做語意判斷標準；Edit Distance 僅比較字元差異，未考慮語意，因此將給予較低的判斷權重。

商用 LLM 中 Gemini 2.5 flash 模型在 F1-score 與 Cosine Similarity 上表現皆是最佳，其 API latency 亦是最低，平均僅有 1.67 秒，而開源小模型中以 Qwen2.5VL 32b 的 F1-score 最高，Gemma3 12b 的 Cosine Similarity 最高，兩者皆可以作為使用開源模型的選擇。

## 三、實驗三：文字+圖片資訊下，LLM 還原失語症語句效果

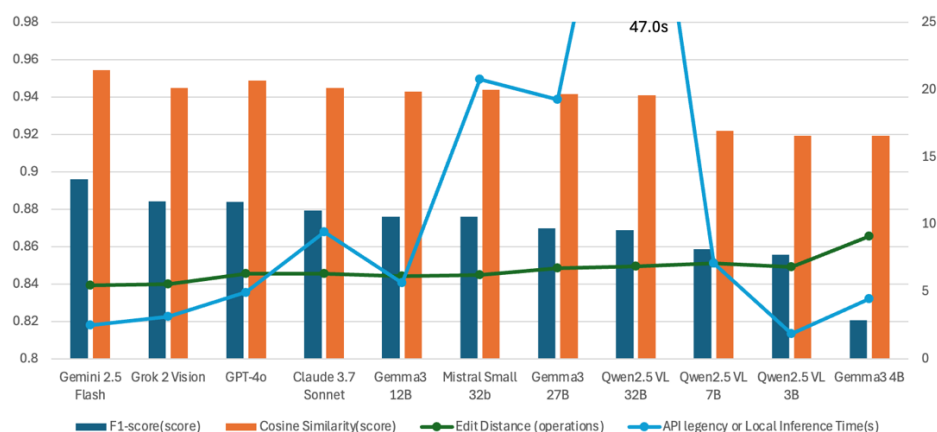


圖 5-2：LLM 還原模擬語句的結果評估（圖片來源：作者繪製）

如圖 5-2 所示，在整合圖像模態的場景中，Gemini 2.5 Flash 依然展現出最佳整體表現，不僅在 F1-score 與 Cosine Similarity 上皆取得最高分數，其 API 響應速度亦為所有商用模型中最快。因此，後續實驗將採用 Gemini 2.5 Flash 作為商用模型的系統實現方案。在開源模型部分，Gemma 3 12B 表現最為優異，成為精確度導向方案的首選。相較之下，實驗二中原本表現最佳的 Qwen2.5 VL 32B，此次排名下滑至第四。值得注意的是，Qwen2.5 VL 3B 在 API 響應速度上甚至超越 Gemini 2.5 Flash，且在 F1-score 上達到 0.86，僅較 Gemini 2.5 Flash 的 0.896 低約 4.5%，顯示其整體效能亦相當出色。因此，Qwen2.5 VL 3B 將作為開源模型中追求快速響應的實作方案。

#### 四、實驗四：關鍵幀算法比較

表 5-2：關鍵幀算法比較實驗結果（表格來源：作者自行製作）

method	Time(s)	Accuracy	Extra Frames / Total Extracted Frames
Optical Flow	17.01	43.62%	87.09%
Frame Difference	0.44	14.42%	84.40%
Block Matching	23.06	58.51%	86.31%
Feature Tracking	1.91	41.13%	87.08%
Histogram Difference	0.52	56.56%	85.34%
ORB	2.13	58.63%	86.16%
CLIP	0.65	39.78%	85.92%

從表 5-2 中可以看到，最高準確率的算法是區塊匹配法（Block Matching），其準確率為 58.51%，但計算時間較長，為 23.06 秒。相比之下，其他算法如光流法（Optical Flow）雖然計算時間為 17.01 秒，但準確率僅為 43.62%，而幀差法（Frame Differencing）雖然在時間上表現優異，僅需 0.44 秒，但準確率只有 14.42%。這些方法雖然在某些方面表現不錯，但無法在速度與準確率之間達到理想的平衡。ORB（Oriented FAST and Rotated BRIEF）則在兩者



之間提供了較好的折衷，準確率為 58.63%，計算時間僅為 2.13 秒，顯示出它在實時處理中能提供較高準確度，同時保證計算效率，特別適合作為對話輔助系統中的關鍵幀提取算法。

## 五、實驗五：系統效果驗證

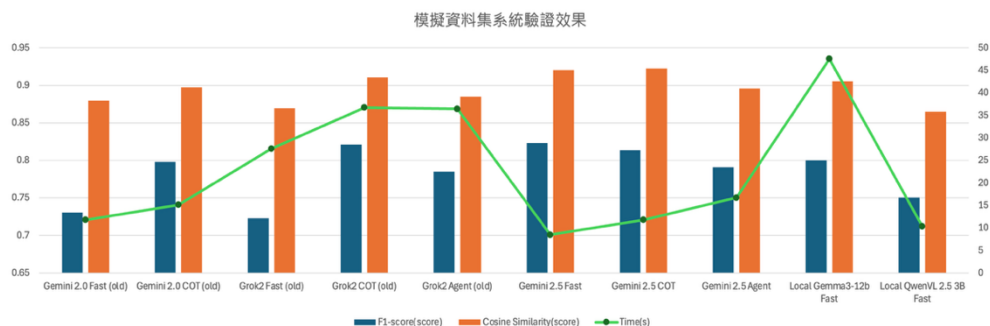


圖 5-3：  
對話輔助系統在  
Vs 資料集的效果  
（圖片來源：作  
者繪製）

如圖 5-3 所示，圖左側標示為「old」的部分代表於研究一開始，前代系統所採用的 LLM 在當時實驗出的結果，包括 Gemini 2.0 Flash 與 Grok2 在 Vs 驗證集上的表現。因為先進模型與增強大語言模型的方法出現，我們再次展開更進一步的實驗，如圖右側，涵蓋 Gemini 2.5 Flash、Gemma 3 12b 以及 Qwen 2.5 3b。其中表現最優的是 Gemini 2.5 Flash Fast 系統，其以僅 8.5 秒的響應時間，達到了超越前代 Grok-2 COT 模式的成績，且速度大幅優於後者。因此，搭配 Fast 模式的 Gemini 2.5 Flash 可視為目前在商用模型應用上的最優解。

進一步觀察 COT 與 Agent 模式的表現差異，在 Gemini 2.0 及 Grok2 系統中，導入這兩種提示詞策略能明顯提升模型對失語症語句的還原能力，F1-score 最高提升至 0.82。然而，在 Gemini 2.5 Flash 系統中，即便使用 COT 或 Agent 模式，F1-score 並未出現顯著提升。我們推測這是因為在未經任務特化微調的情況下，模型在 Vs 可達到的 F1-score 上限約為 0.8，當達到此一極限後，即使透過提示詞誘導 LLM 進行更完整的推論（例如產出更多 token），對準確度的提升亦有限。

至於開源小模型方面，Gemma 3 12b 表現穩定，其準確度已接近前代的 Gemini 2.0 Flash。然而，由於其視覺模態計算成本較高，在我們的實驗設備（M2 Max）上需耗費大量運算時間。若要應用於實際部署場景，則需配備更高效能的 CPU 與 GPU，以同時兼顧高準確率與即時反應的需求。

另一方面，Qwen 2.5 VL 模型在響應速度方面與 Gemini 2.5 Flash 相當，儘管準確度略遜一籌，但仍優於 Gemini 2.0 Flash 與 Grok2 的 Fast 模式表現。綜合考量，其在算力受限或邊緣運算等場景下，具有良好的應用潛力。

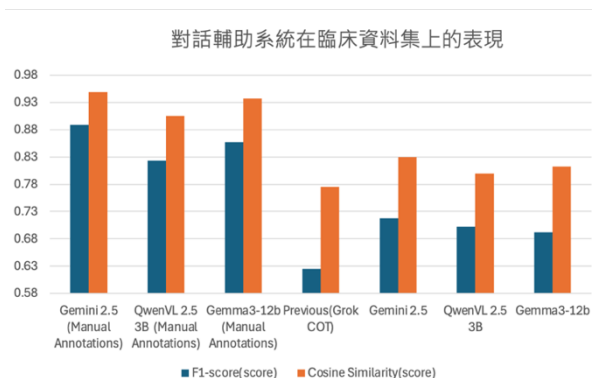


圖 5-4：對話輔助系統在 Vc 資料集的效果（圖片來源：作者繪製）

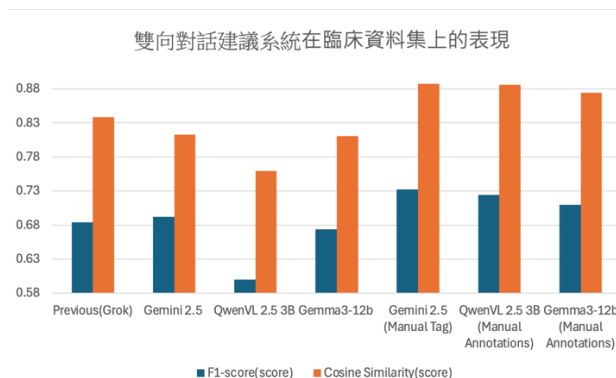


圖 5-5：雙向對話建議系統在 Vc 資料集的效果（圖片來源：作者繪製）

如圖 5-4 所示，Gemini 2.5 Flash Fast 系統在臨床資料集上的表現仍為最佳，其 F1-score 為 0.71，但相較於其在 Vs 驗證集上達到的 0.82，仍有明顯落差。經深入分析系統記錄檔後發現，表現下滑的主因在於 Aphasia Bank 資料集本身的品質問題：該資料集中的對話影片畫質較差，且背景雜音干擾嚴重，導致系統常無法準確偵測人臉或擷取失語症患者的語音內容。

為改善此問題，我們嘗試將 Aphasia Bank 所提供之 .cha 檔案中記錄的患者語句嵌入模型提示詞中，進行人工標註（Manual Annotations）輔助測試。結果顯示，無論是 Gemini 2.5 Flash、Qwen2.5 VL 3b 或 Gemma 3 12b，在加入人工標註輔助後，語句還原表現皆顯著提升。

在雙向對話建議系統的實驗中亦觀察到類似情形（見圖 5-5）。由於該系統需根據對方的發話內容及對話歷史推測失語症患者接下來的表達，若因背景干擾無法準確擷取對方語意，將導致模型生成錯誤回應，且錯誤會隨對話歷程不斷累積，進一步惡化後續預測準確性，整體表現也因此下降。基於上述觀察，為提升系統於實務場域中的效能，本研究所設計之應用程式將統一採用 1080p 畫質錄製失語症患者語音與影像，並進行錄音降噪處理，以確保資料品質，進而提升模型推論的準確性。

## 六、實驗六：系統效率優化

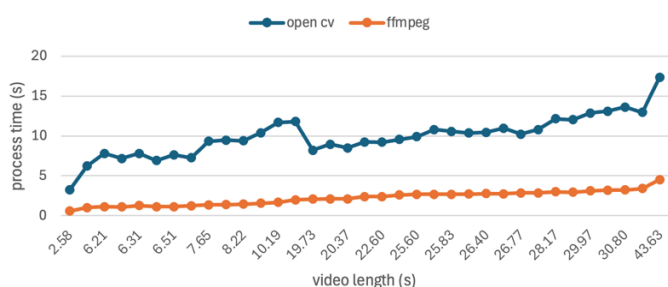


圖 5-15：影片取幀優化（OpenCV vs. FFmpeg）

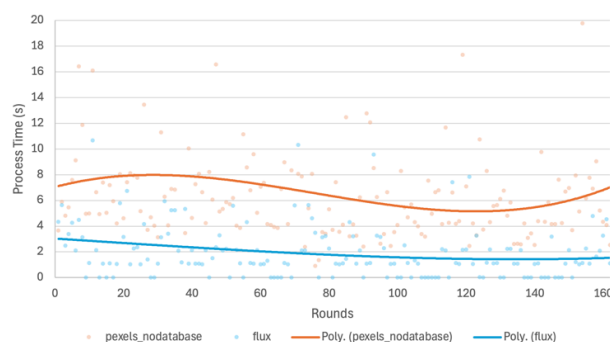


圖 5-16：圖卡即時建議系統優化（圖片來源：作者繪製）

### (一) 影片取幀優化

從圖 5-15 可以看出，改使用 FFmpeg 後，影片處理耗時顯著減少，且隨影片時長增加的計算耗時增幅較小，這樣的優化顯著提升對話輔助系統的計算效率。

### (二) 輕量 Flux 文生圖與圖像資料庫優化

如圖 5-16，採用輕量 Flux 文生圖與圖像資料庫優化比使用 Pexels API 更為迅速，且耗時趨勢隨查詢輪數穩定下降。

### (三) 異步優化

通過異步優化，系統的計算效率得到了顯著提升，Agent 模式下提升了 28.1%，COT 模式下最多提升了 22.1%，Fast 模式下提升了 27.6%。如圖 6-1 顯示，Gemini 2.5 flash，經過異步優化與 FFmpeg 的優化後，其響應效率皆有顯著提升，Fast 提示詞下最效率提升 55%，COT 下提升 47%，Agent 下提升 38%。異步優化確實有效改善了對話輔助系統的計算效率。

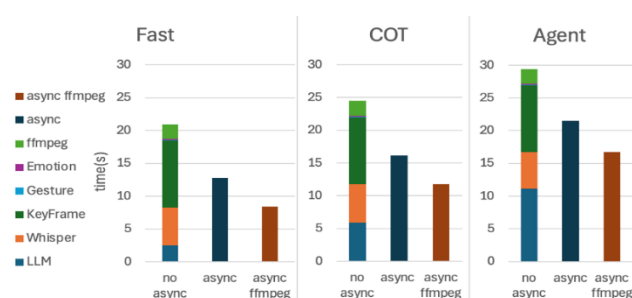


圖 6-1：  
Gemini 2.5 Flash 在不同模式下的計算耗時  
(圖片來源：作者繪製)

## 七、實驗七：輕量 LLM 微調效果

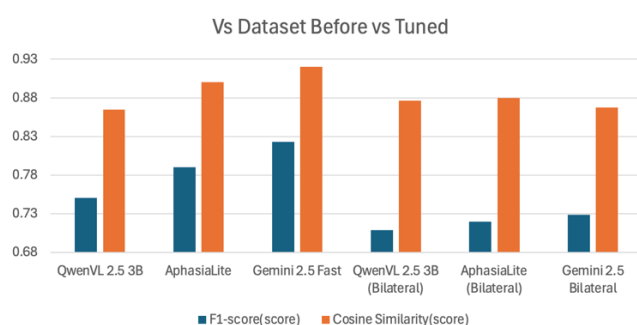


圖 6-5：微調輕量 LLM 在系統應用中的效果  
(圖片來源：作者繪製)

AphasiaLite-LLM 模型在經過微調後展現出明顯的性能提升。在對話輔助系統中，其 F1-score 提高了 0.04；在雙向對話建議系統中，表現更是接近 Gemini 2.5 Flash，僅相差 0.006，顯示透過資料集蒸餾微調的方式，能有效強化輕量開源模型於失語症語句還原任務上的表現。儘管目前可取得的應用實例仍有限、微調資料集規模也尚不足（少於 300 筆），導致整體效果仍略遜於 Gemini 2.5 Flash，但本實驗已實證此方法的可行性。未來若能持續累積更多實際應用資料，預期能進一步提升模型效能，實現高效能且低成本的開源部署方案。

## 陸、 討論與結論

### 一、AphasiaSim-LLM 的模擬方法能有效生成模擬失語症語句。

本研究針對中文失語症語料稀缺的問題，設計出結合語意遮蔽與錯詞替換的模擬演算法，並透過微調語言模型 AphasiaSim-LLM，自動生成與真實失語症語句相似的資料。經由三種 Naïve Bayes 模型進行分類測試，qwen2.6-7b 模型於 GaussianNB 下分類準確率降至 0.5508，與真實語句組的 0.9117 相比，下降 36.8%，顯示模擬語句難以與真實語句區分，達到高度擬真效果。相較 van Vaals et al. (2024)等人方法辨識準確率落差有限，我們的方法能更有效混淆分類器，證實其在模擬資料生成上的效益。此結果為後續進行失語症語句還原模型訓練與評估提供關鍵資料支援。

### 二、還原失語症語句的任務上 Gemini 2.5 Flash 效果最佳，而開源模型亦有其優勢。

在失語症語句還原任務中，不同 LLM 在語意復原的準確度與系統效能上展現出各自優勢。商用模型如 Gemini 2.5 Flash 不僅擁有最佳的 F1-score 與 Cosine Similarity，亦具備極快的響應速度，適合用於需要即時處理與高正確率的臨床應用場景；而在開源模型方面，Gemma 3 12B 提供高精度的語意還原能力，Qwen2.5 VL 3B 則兼顧準確性與速度，成為低成本、大規模部署的潛力方案。而在可調整性與成本控制上，開源 LLM 同樣具備高度實用性。

### 三、手勢識別模型的性能佳，且此模態確實有助於對話輔助系統還原語句。

我們的驗證結果顯示，手勢識別模型在各項指標上均達到 86%以上，分類性能表現優異。其種多數手勢參考 Aphasia Library 中所提及的失語症患者常見手勢，提升系統在實際應用中的效果。在系統測試中，手勢模態也能有效協助系統還原失語症患者的語意。

### 四、異步處理與 FFmpeg 影片取幀能顯著提升系統即時性。

研究針對系統效能瓶頸進行優化，引入異步處理避免 Whisper 和 LLM API 等待時間，最多可提升 28%的計算效率；同時將原先使用 OpenCV 的影片取幀方法改為使用 FFmpeg，大幅降低影片處理的延遲，效率提升 77.87%，提升了系統整體即時性，以符合臨床需求。

### 五、Flux 模型與個人化圖像資料庫有效改善圖卡生成效能。

研究團隊將原本不穩定且耗時較長的 Pexels API 方案替換為輕量級 Flux 文生圖模型與個人化圖像資料庫方案，能即時生成新圖卡，並在生成後儲存於資料庫內供重複使用，明顯減少後續重複生成的響應時間，圖片生成趨勢穩定下降，效率提升 70.5%，也改善了圖卡圖像品質與一致性。

六、ORB 關鍵幀法與伺服器資源分配及並行處理策略，有效降低多模態分析負擔並加快響應。

研究比較多種關鍵幀算法後，選擇以 ORB 作為影片關鍵幀提取方法，提升影片處理效率，並減少系統的計算負擔。同時，將伺服器任務依據特性與是否有衝突，分配至三個獨立佇列，成功避免擁塞與 API 衝突，進一步提升並行處理能力與系統運作效率。

七、商用模型 Gemini 2.5 Flash Fast 模式於對話輔助任務中表現最優，提示詞工程的效果趨於飽和。

在本研究中，我們觀察到 COT (Chain-of-Thought) 與 Agent 提示詞策略在先前的 Gemini 2.0 與 Grok2 系統中能有效提升模型對失語症語句的還原能力，最高可使 F1-score 提升至 0.82。然而，當相同策略應用於 Gemini 2.5 Flash 時，其準確度未見顯著改善。我們推測這是因為 Gemini 2.5 Flash 即便未經任務特化微調，在 Vs 資料集上即已接近其性能上限（約 F1-score 0.8），導致即使透過提示詞引導模型進行更深入推理，也無法有效突破既有表現瓶頸。因此，對於已達高準確度的先進模型，提示詞策略的邊際效益相對有限。

八、資料品質與提示設計是提升模型效能的關鍵。

實驗結果顯示，模型在 Vs 資料集上表現良好，但在 Aphasia Bank 等臨床資料集上的 F1-score 明顯下滑（例如 Gemini 2.5 Flash 從 0.82 降至 0.71），主因在於資料畫質差、雜訊大，造成視覺與語音模態辨識困難。為此，本研究引入資料集中的 .cha 失語症語句的人工標註作為輔助提示詞，成功提升了多個模型的還原效果。進一步實務應用中，採用 1080p 高畫質影像與錄音降噪處理是確保模型準確性與穩定性的關鍵策略。

九、採用資料集蒸餾策略微調輕量模型，顯著提升失語症語句還原效能

本研究採用資料集蒸餾 (Dataset Distillation) 策略，將 Gemini 2.5 Flash 所產生的多樣輸出語句，經排序與標準答案修正後，重組為具有語句品質階層的微調資料集，成功應用於輕量模型 Qwen2.5 VL 3B 的 LoRA 微調。實驗結果顯示，AphasiaLite-LLM 在失語症語句還原任務中，F1-score 提升 0.04，雙向對話建議任務中更與 Gemini 模型僅差 0.006，證明資料集蒸餾能有效強化輕量開源模型的應用表現，即使在資料量有限的情況下，仍具實質成效。此方法為未來低成本、高效率部署提供具體可行的方向。

十、失語症醫療專業人員訪談與建議

為確保本系統在臨床應用中的有效性，我們與神經內科專科醫師進行訪談，並收到以下三項重要建議：

（一）系統應用範圍：服務對象應限定於非流暢型且理解能力較完整的失語症患者，主要

針對布洛卡失語症（Broca's Aphasia）與跨皮質運動型失語症（Transcortical Motor Aphasia）。

(二) 提升語意選項多樣性：不同患者的表達能力差異大，增加語意選項數量與多樣性能提高語句生成的準確性。

(三) 降低對文字閱讀的依賴：部分患者有閱讀困難，系統應設計自動選擇預設語句功能，減少文字閱讀需求，提升使用友善度。

針對以上建議，我們已明確界定系統服務對象與範圍，並透過優化 Prompt 設計擴增選項數量至五項，提升互動效率。此外，系統設定中新增了選項，讓閱讀困難的患者可直接略過語句選擇，提升使用體驗。

## 柒、系統其他應用層面與未來方向

### 一、本作品其他應用層面

#### (一) 長者照護與其他身心障礙輔助

本系統所採用的整合唇語、手勢、表情辨識技術，除了能應用於失語症患者外，還可以擴展到其他有溝通困難的身心障礙人士，如自閉症、唐氏症等。它也可用於長者照護，協助長者進行語言復健或提升日常生活中的對話流暢性。

#### (二) 語言學習輔助工具

App 中的唇語識別、語音識別和手勢識別技術，能作為中文學習的輔助工具。即便初學者無法完整表達意思，仍可以透過片段語句與肢體動作，讓系統反饋完整語句，幫助使用者更準確地掌握中文的發音和語法結構。

### 二、未來研究方向

(一) 結合自動資料生成與長期蒐集以擴充蒸餾資料集規模。

(二) 強化中文唇語識別準確性。

(三) 擴充肢體語言資料庫以支援非語言輔助溝通。

## 捌、 參考資料

- [1] What is aphasia? (n.d.). National Aphasia Association. <https://aphasia.org/what-is-aphasia/>
- [2] Manir, S. B., Islam, K. M. S., Madiraju, P., & Deshpande, P. (2024). LLM-based text prediction and question answer models for aphasia speech. IEEE Access. <https://doi.org/10.1109/ACCESS.2024.3443592>
- [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. arXiv. <https://arxiv.org/abs/1706.03762>
- [4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. <https://arxiv.org/abs/1810.04805>
- [5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. arXiv. <https://arxiv.org/abs/2201.11903>
- [6] Azrien, E. A., Hartati, S., & Frisky, A. Z. K. (2024). Regularized Xception for facial expression recognition with extra training data and step decay learning rate. <https://ijai.iaescore.com/index.php/IJAI/article/view/25153/14297>
- [7] Ma, P., Petridis, S., & Pantic, M. (2021). End-to-end audio-visual speech recognition with conformers. arXiv. <https://arxiv.org/abs/2102.06657>
- [8] Communication Tips (n.d.). the Aphasia library. <https://www.aphasia.com/>
- [9] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., Bernstein, M. S., & Li, F. F. (2016). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. arXiv. <https://arxiv.org/abs/1602.07332>
- [10] Kapitanov, A., Kvanchiani, K., Nagaev, A., Kraynov, R., & Makhliarchuk, A. (2024). HaGRID - HAnd Gesture Recognition Image Dataset. arXiv. <https://arxiv.org/abs/2206.08219>
- [11] Visual Intelligence and Pattern Analysis Group. (2018). CMLR: Chinese Multi-Modal Laughter Recognition Dataset. Zhejiang University. <https://www.vipazoo.cn/CMLR.html>
- [12] The Aphasia Bank Project. (n.d.). Aphasia Bank. TalkBank. Retrieved March 21, 2025, from <https://aphasia.talkbank.org/>
- [13] Misra, R., Mishra, S. S., & Gandhi, T. K. (2022). Assistive completion of agrammatic aphasic sentences: A transfer learning approach using neurolinguistics-based synthetic dataset. arXiv. <https://arxiv.org/abs/2211.05557>
- [14] van Vaals, S., Matuskevych, Y., & Tsiwah, F. (2024). Generating completions for fragmented Broca's aphasic sentences using large language models. arXiv preprint arXiv:2412.17669. <https://arxiv.org/abs/2412.17669>

## 【評語】 052516

本作品參考了文獻上有關失語症語言特徵描述，使用 BERT 注意力分數引導的 token 替換、遮蔽和重複，進行失語症溝通系統的研發，題目具應用性且是大眾所會關注議題之一。

作品有良好的研究重要性說明，文獻探討與研究架構都相當完善。研究結論有呼籲所提出的研究問題。作品書撰寫論述邏輯性佳，相當容易閱讀。

目前僅為採用語料庫資料進行實驗，未來可結合臨床測試，建立更有說服力的研究成果。



作品海報

# 基於多項人工智慧之 多模態失語症溝通系統探究



## 研究動機

家人曾提及，對於某位知名好萊塢影星因罹患失語症而不得不息影一事感到惋惜。在對失語症有了更多了解後，我們不禁好奇近年來快速發展的人工智慧模型，是否能為這些病患提供協助，減輕他們的生活負擔？

失語症患者因中風、腦部創傷等原因喪失語言能力，溝通困難對照顧者和患者都是巨大挑戰。現行借助肢體語言、圖片等輔助方式仍有其局限，難以準確理解患者想法。

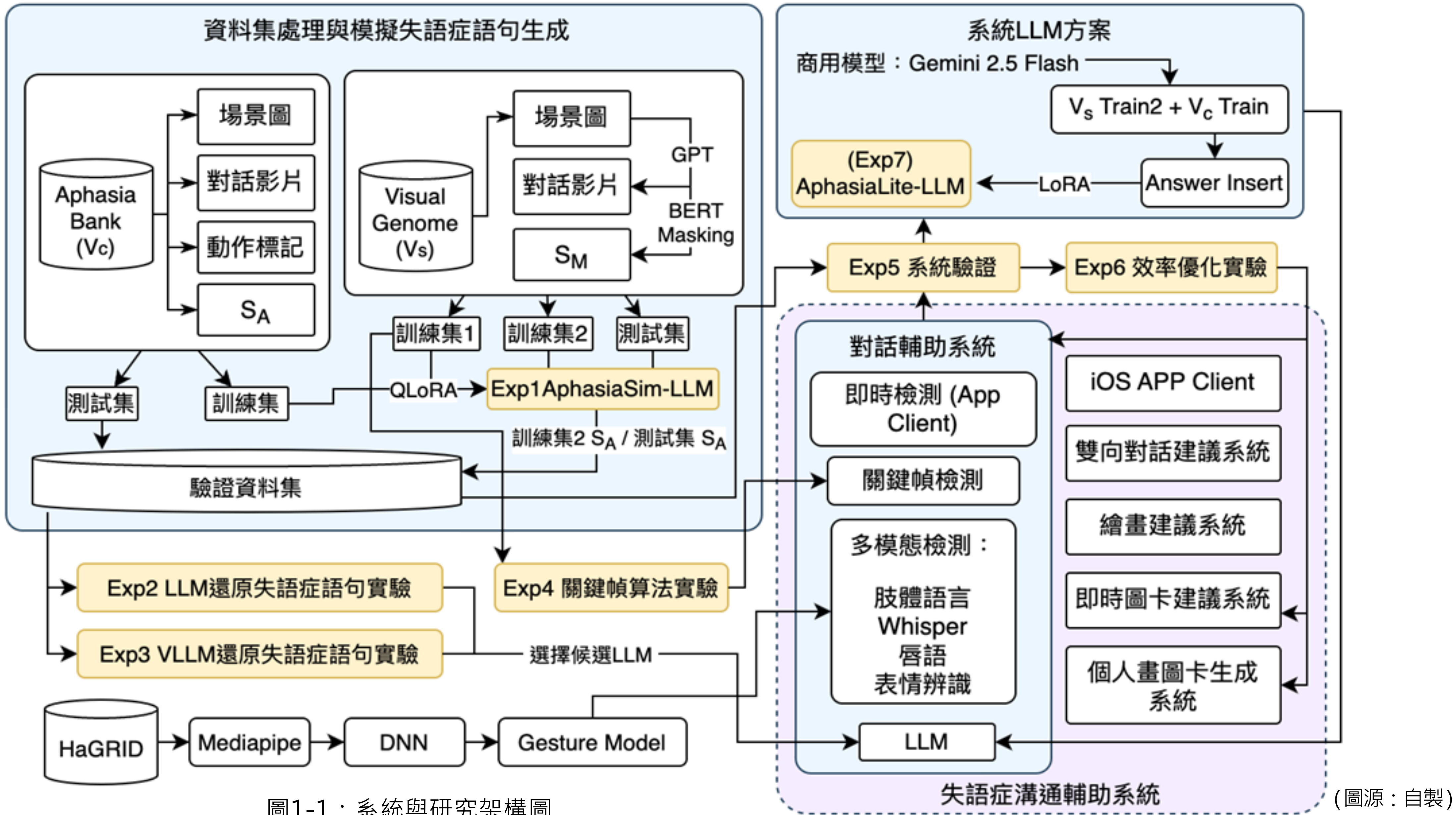
隨著AI與大語言模型的發展，解決此問題出現新契機。我們的目標是開發一款iOS App，結合環境影像、語音、唇語、手勢及情緒識別的多模態模型，協助醫護人員更精確理解患者需求，提升溝通效率。

## 研究目的

- 微調出 AphasiaSim-LLM，藉此模擬高擬真失語症語料，以利進行還原效果評估。
- 使用客觀量化方式來評估模型，比較各VLLM在還原失語症語句效果，以找出系統最適合模型。
- 分析多種關鍵偵提取演算法，以設計系統最適合的提取機制。
- 微調開源 VLLM 得出 AphasiaLite-LLM，以利輕量化硬體需求。
- 分析各種效率優化策略與增強模型能力的技巧，探討對於系統效率與還原正確率的提升程度。
- 整合各個子系統與VLLM交互，驗證失語症溝通輔助系統的整體效果。

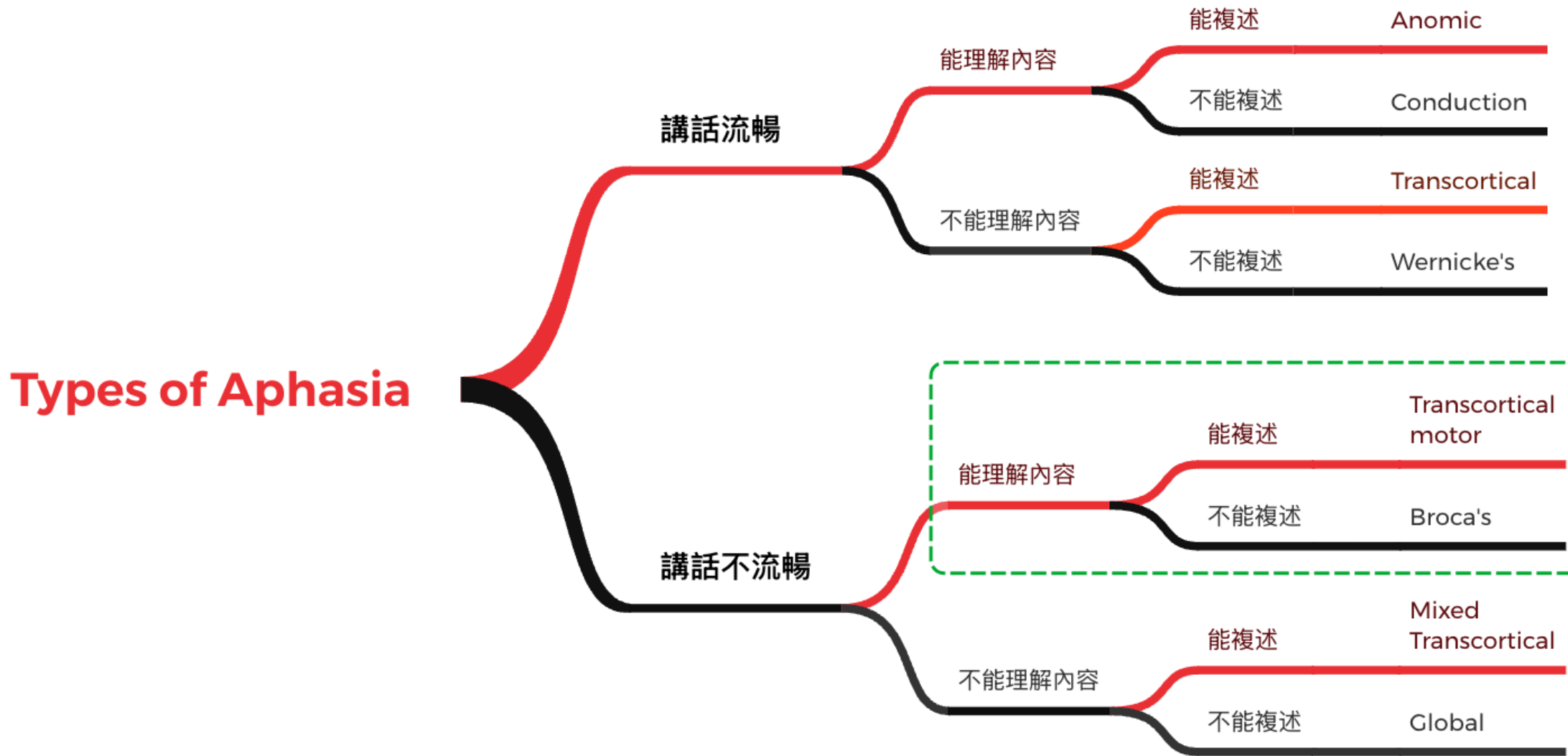
## 研究架構與背景知識介紹

### 一、研究架構



### 二、失語症介紹

本研究所設計的系統，將患者不流暢的話語轉換成順暢的語句，因此能幫助的對象僅限非流暢型的失語症，如圖1-2，失語症按照功能可分為以下幾種：



(圖源：自製)

圖1-2：失語症分類

### 三、資料集介紹

- Visual Genome：包含108,077張生活場景圖像
- Aphasia Bank：具有約180名失語症患者和140名非失語症對照者以統一的任務腳本的話語影片。

### 四、模擬語句還原效果評估方法

使用LLM還原失語症資料的方法，現有論文以人為評分的效果近似於隨機評分，可參考性較低且費時。我們透過以下三種方法，評估LLM還原語句( $S_R$ )與流暢語句( $S_F$ )的相似性。

表1：效果評估方式比較 (表源：自製)

方法	用途	優勢
BERTScore	語義匹配、文本評估	捕捉語意相似度，不受詞序影響
Cosine Similarity	語句級語義比較	測量整體語意關係，適合句子比較
Edit Distance	字元級或詞級相似度計算	計算簡單，適用於錯別字校正

## 實驗設計與結果

### 實驗一：AphasiaSim-LLM 模擬失語症語句

因Aphasia Bank僅包含失語症患者的語句，無法直接用於以「對話場景 - 失語症語句 - 真實語意」為結構的實驗。我們設計一套結合低精度微調技術的模擬失語症語句方法。詳細流程如下。

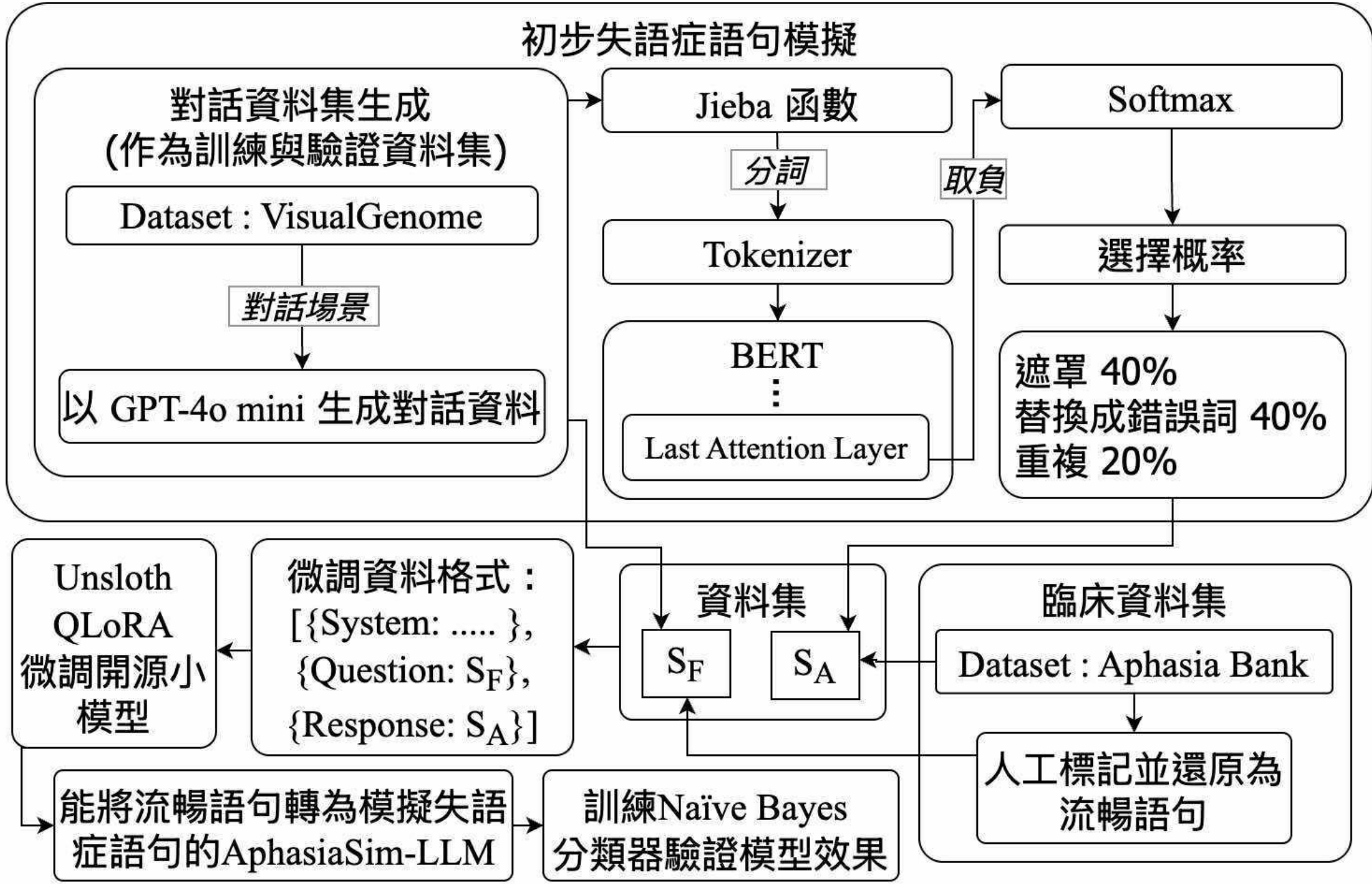


表2-1：不同AphasiaSim-LLM的分類器訓練結果 (實驗一) (表源：自製)

Accuracy (Score)	previous paper	llama3.2-3b	llama3.1-8b	qwen2.5-7b	qwen2.5-14b	Mistral v0.3-7b	fluent vs aphasia
BernoulliNB	0.83/0.81	0.6840	0.6558	0.6321	0.5892	0.6769	0.6965
GaussianNB	(experimental)	0.5982	0.5756	0.5508	0.6038	0.9323	0.9117
MultinomialNB	/control)	0.6806	0.6512	0.6535	0.6738	0.7246	0.6978

最後我們用GPT-4o-mini為Visual Genome資料集中的日常場景生成流暢語句，並透過AphasiaSim-LLM生成對應的模擬語句，產生資料集。

表2-2：本研究中不同類型的語句資料簡寫 (表源：自製)

中文	英文	簡寫
AphasiaBank 的臨床失語症語句資料	Clinic Aphasic Sentence Samples	S <sub>C</sub>
AphasiaSim-LLM 生成的模擬失語症語句	AphasiaSim Simulated Aphasic Sentence	S <sub>A</sub>
S <sub>C</sub> 人為標註還原為流暢語句的結果或從模擬失語症語句之演算法產生的流暢語句	Fluent Sentence	S <sub>F</sub>
失語症語句經過系統還原為流暢語句的結果	Restored Sentence	S <sub>R</sub>

表2-3：模擬失語症資料範例 (表源：自製)

流暢語句 (S <sub>F</sub> )	對話場景
這個房間看起來有點亂，需要好好整理一下。 是啊，我的書桌上堆滿了東西，都沒地方放了。	
模擬失語症語句 (S <sub>A</sub> )	
這 房 間 看 起 來 看 來 沒 有 整 理 是 啊 ， 我 的 上 都 堆 了 東 西 ， 都 了 沒 有 放	



實驗二、實驗三：LLM還原模擬語句(S<sub>A</sub>)實驗

我們設計了兩個實驗來驗證不同LLM將S<sub>A</sub> 還原為S<sub>R</sub> 時的效果。實驗二僅提供S<sub>A</sub> 並由LLM生成S<sub>R</sub>。而實驗三中提供S<sub>A</sub>，的同時提供Visual Genome 中的對話場景，讓 Vision LLM 生成S<sub>R</sub>。然後找出還原模擬語句表現最佳的LLM。

(一) 純文字資訊下，LLM還原失語症語句效果

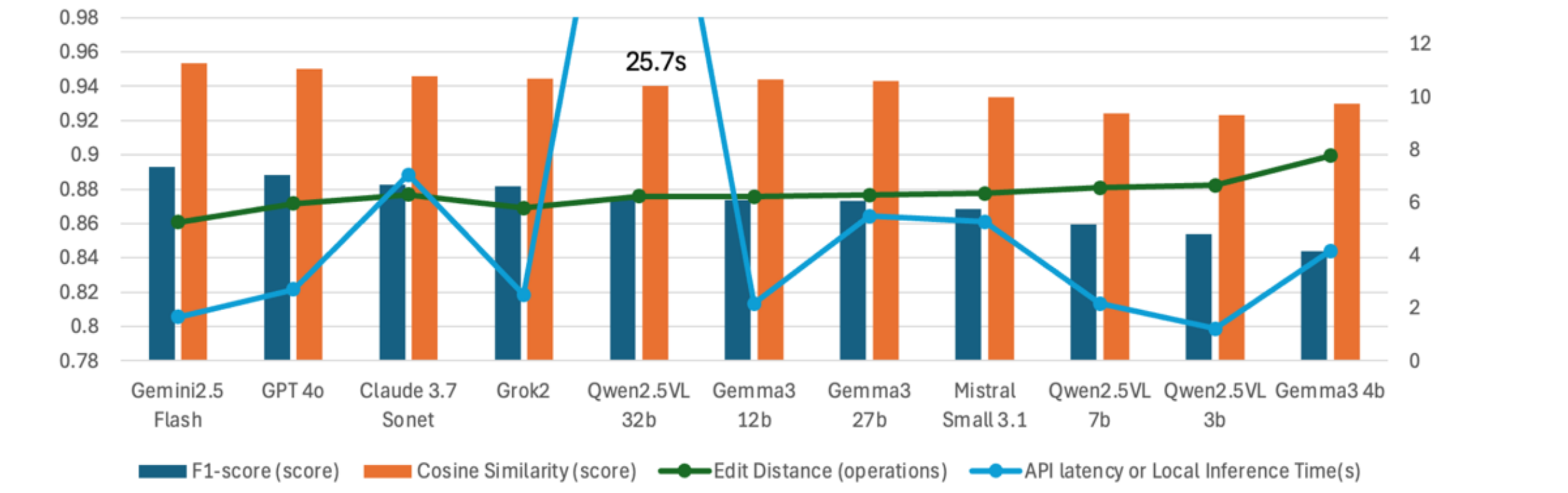


圖2-2：LLM還原模擬語句的結果評估 (圖源：自製)

(二) 文字 + 圖片資訊下，LLM還原失語症語句效果

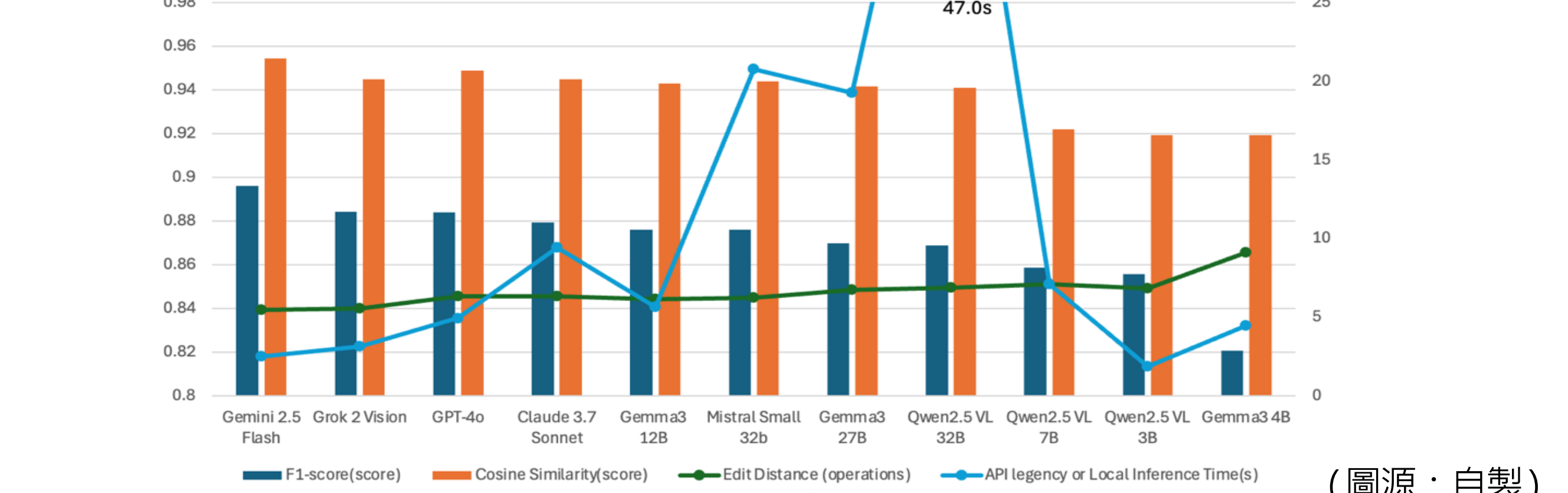


圖2-3：LLM還原模擬語句的結果評估 (圖源：自製)

綜合結果，後續實驗將採用 Gemini 2.5 Flash 作為商用模型的系統實現方案。在開源模型部分，Gemma 3 12B 表現最為優異，成為精確度方案的首選。其中，Qwen2.5 VL 3B 在 API 響應速度上甚至超越 Gemini 2.5 Flash，且在 F1-score 上達到 0.86，僅稍低於 Gemini 2.5 Flash 的 0.896。因此將作為開源模型中追求快速響應的實作方案。

實驗四：關鍵幀提取技術比較實驗

此實驗的目的是探討不同幀間差異算法在即時計算場景下，提取失語症患者動作與表情關鍵幀的效果。為符合臨床情況，採用Aphasia Bank資料集來評估不同關鍵幀算法的效果。

實驗首先提取影片中的動作敘述形成動作序列，以及對影片每0.2秒取一幀，形成幀序列。然後由LLM對兩個序列交叉評分，透過峰值檢測定位該動作的確切幀。如圖2-4所示。

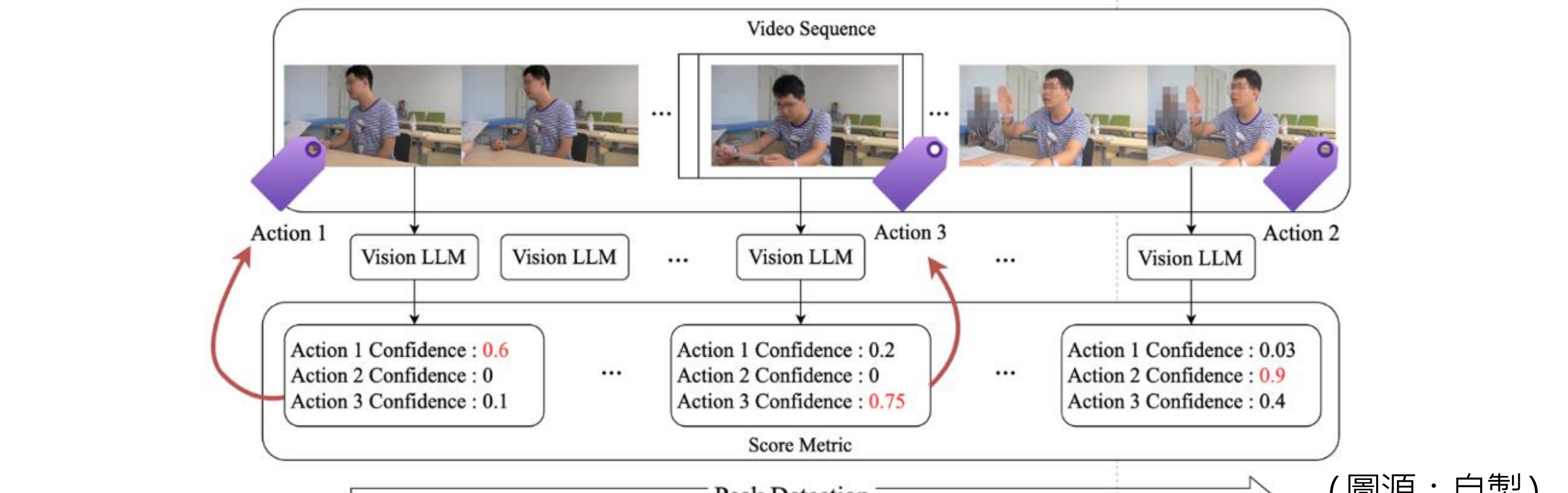


圖2-4：動作確切幀演算法 (圖源：自製)

以關鍵幀算法計算幀序列各幀的分數後，以此從影片中提取出關鍵幀序列。計算關鍵幀找出確切幀中的幀數，再除以確切幀序列長度作為準確度。

比較各種關鍵幀算法，如表2-4，ORB算法準確率為 58.63%，耗時僅 2.13 秒，顯示它能提供較高的準確度，同時維持計算效率，適合作為系統中的關鍵幀提取算法。

表2-4：不同關鍵幀算法準確度比較 (表源：自製)

method	Time(s)	Accuracy	Extra Frames / Total Extracted Frames
Optical Flow	17.01	43.62%	87.09%
Frame Difference	0.44	14.42%	84.40%
Block Matching	23.06	58.51%	86.31%
Feature Tracking	1.91	41.13%	87.08%
Histogram Difference	0.52	56.56%	85.34%
ORB	2.13	58.63%	86.16%
CLIP	0.65	39.78%	85.92%

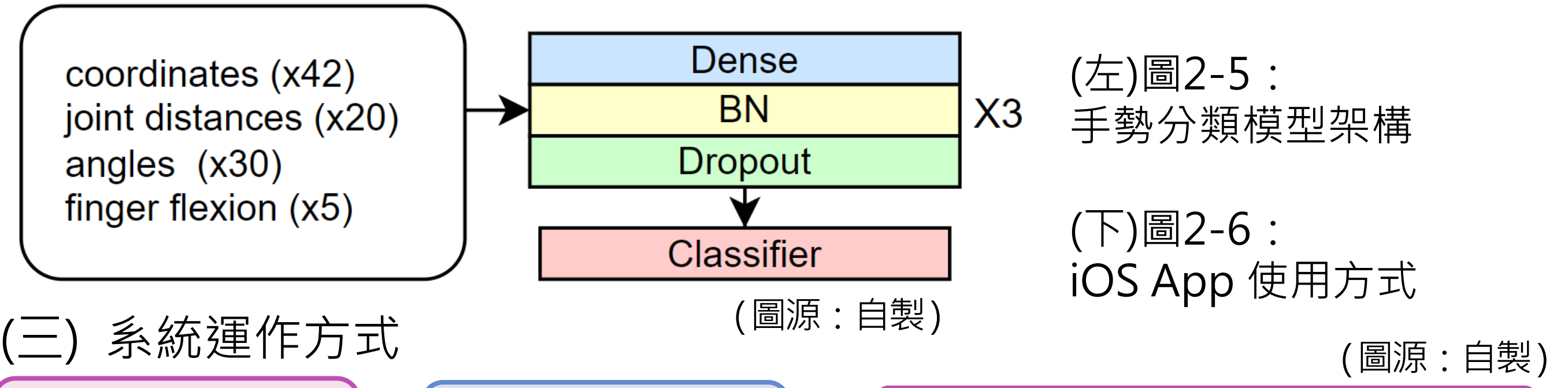
實驗五：系統設計與驗證

(一) 系統目標

- 市面上的輔助系統要求患者逐字選擇來組句，對於失語症患者來說並不直觀。因次系統導入LLM並整合多模態輸入，讓患者只需
- 「說出」、「選擇」或「畫出」概念就能表達。  
圖卡內容應貼近日常生活，然而現有系統多仰賴人工上傳圖片。
- 此系統設計能根據對話情境，自動擷取新物件並生成個人化圖卡。市面上輔助系統只仰賴單一模態，無法即時適應不同應用場景。

(二) 訓練手勢分類模型

模型採用HaGRID資料集訓練，透過MediaPipe取得手部關鍵點間的各種特徵，強調模型的輕量與準確性，以符合系統即時計算的需求。

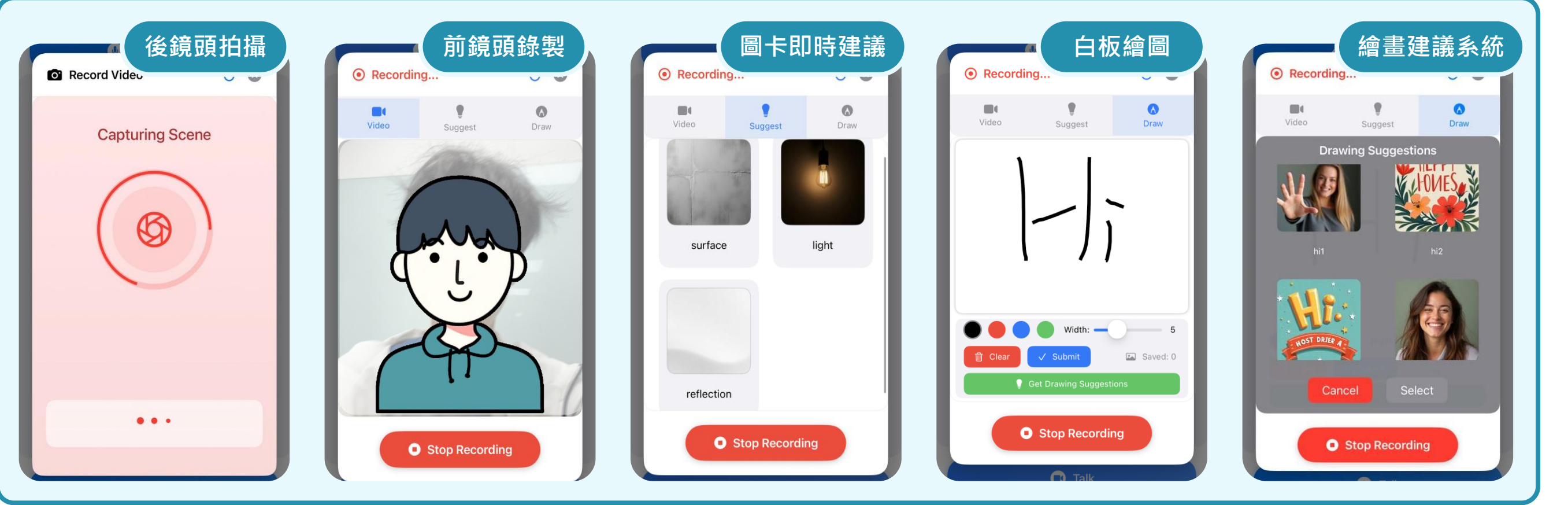


(左)圖2-5：手勢分類模型架構

(下)圖2-6：iOS App 使用方式



(圖源：自製)



當對方說完話後，被動對話模式生成患者可能想要回應的句子，可選擇最符合意圖的回應作為下一句話。若沒有符合選項，可切換至主動敘述模式，表達自己的想法，交由系統還原成流暢語句。

如下圖所示，系統架構由手機App與Flask伺服器兩部分協作運行。(圖源：自製)

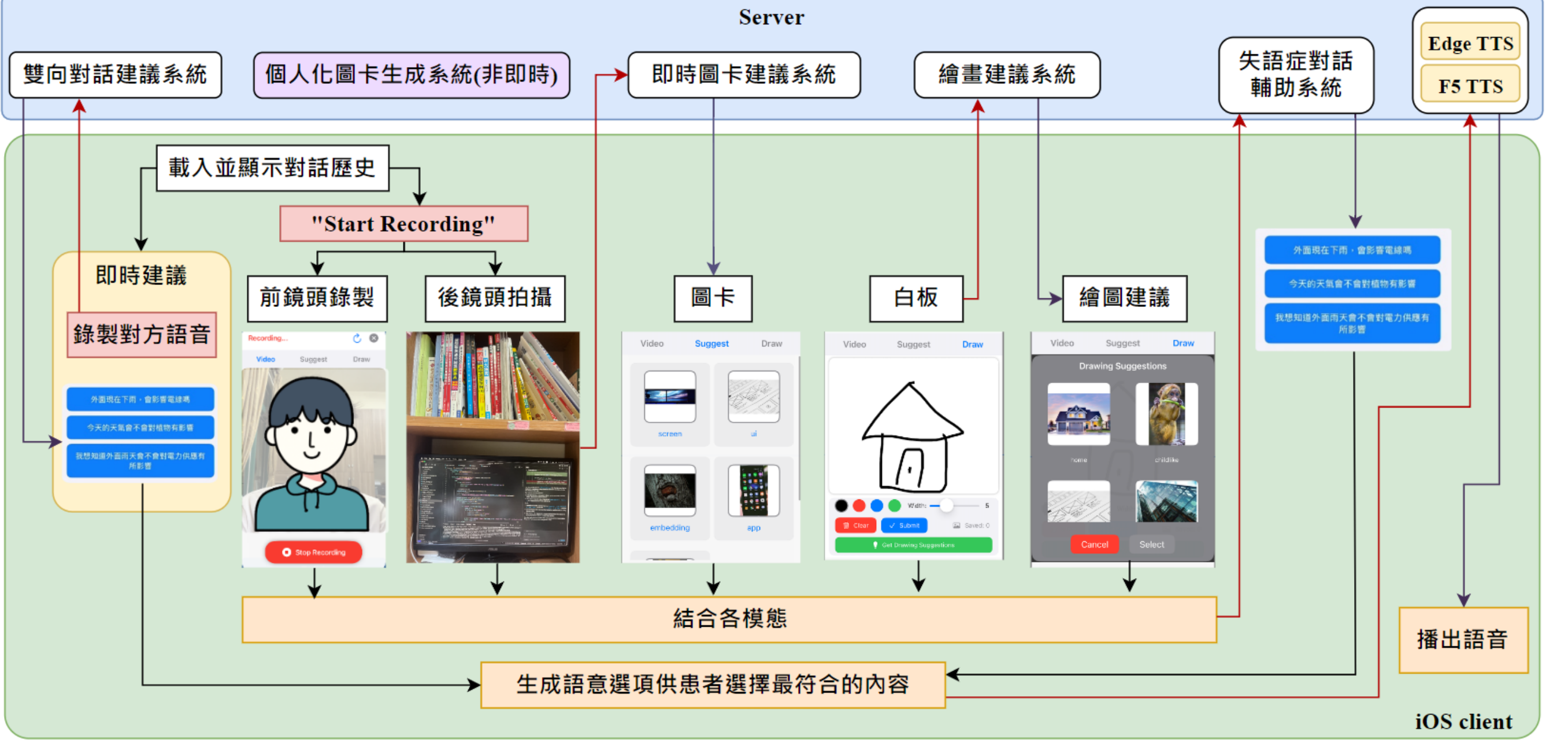


圖2-7：App運作流程圖

以下是五大子系統的技術細節與實現方式：

子系統一：繪畫建議系統

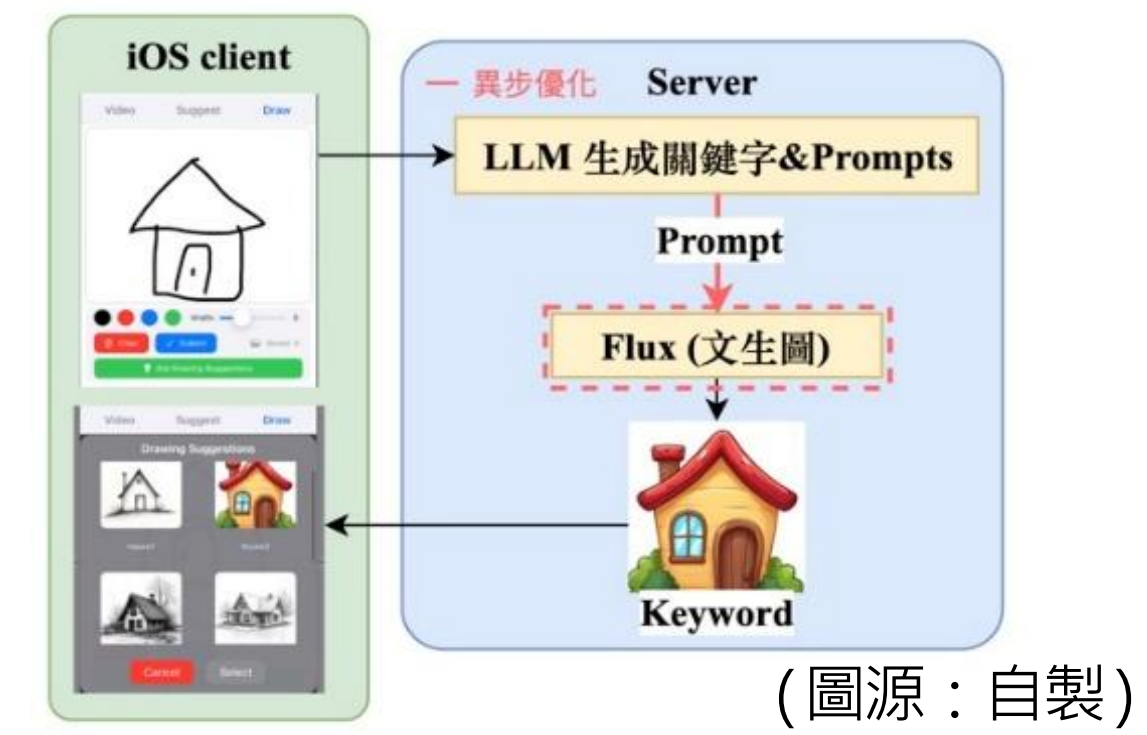


圖2-8：繪畫建議系統

子系統二：圖卡即時建議系統

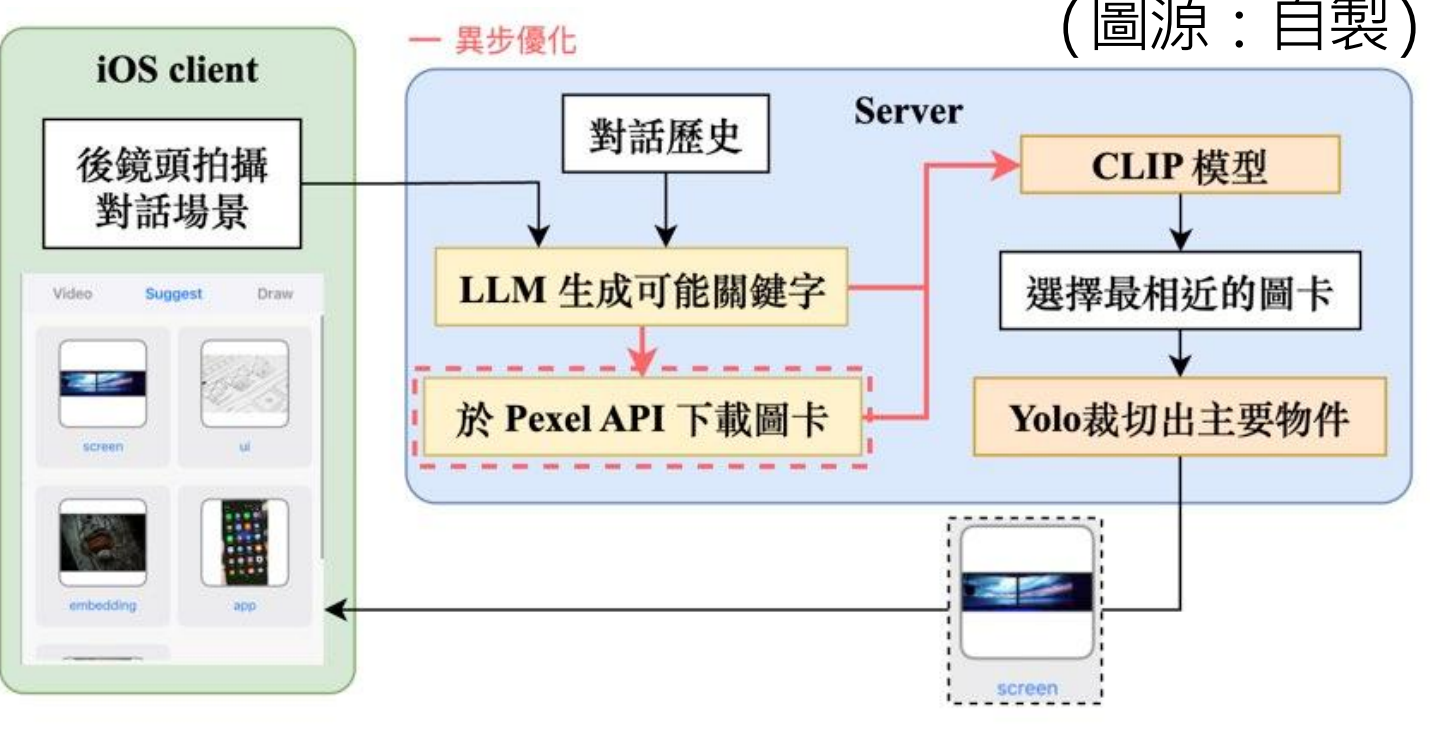


圖2-9：圖卡即時建議系統

子系統三：個性化圖卡生成系統

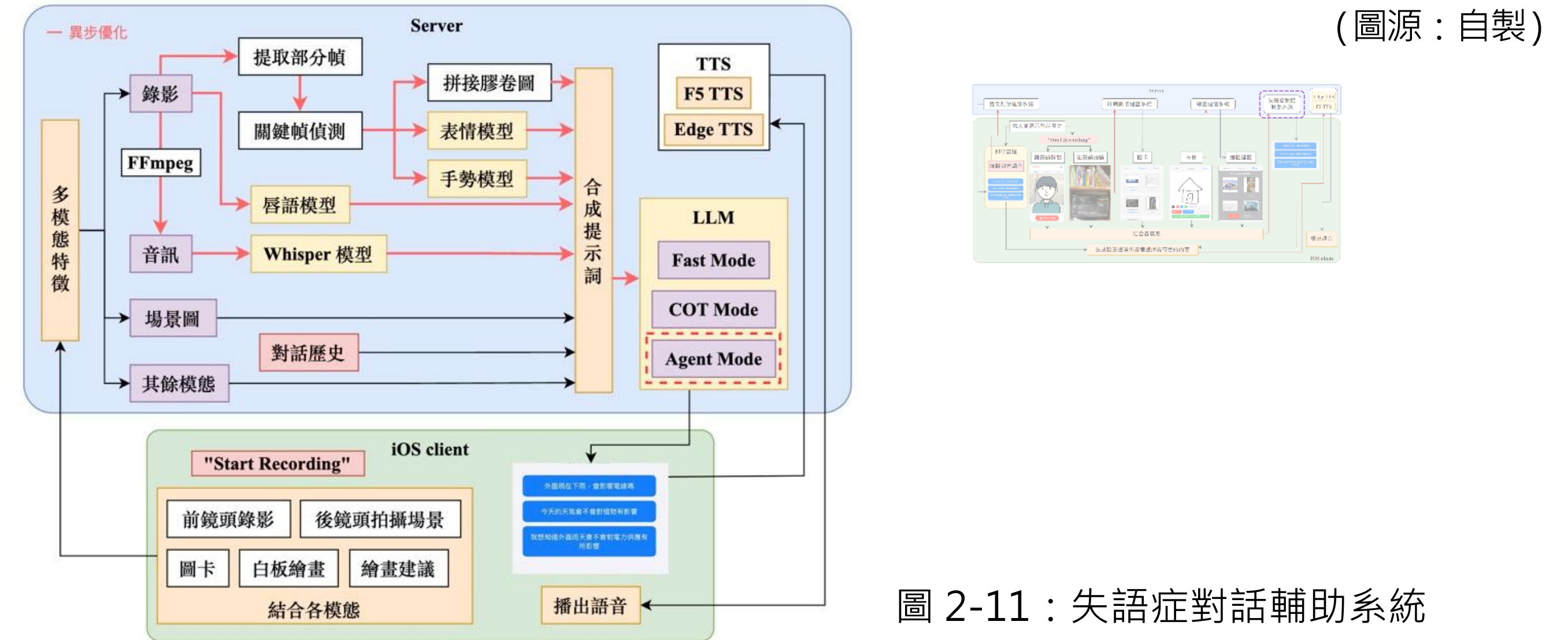
系統透過Yolo檢測對話歷史的場景與影片中出現的物件。裁切後上傳至Gemini 2.0 Flash生成對應的關鍵字與修改建議，透過Gemini-2.0-Flash-exp-image-generation依據修改建議修正圖片視角、大小並提升畫質。

圖2-10：個性化圖卡生成系統 (圖源：自製)

子系統四及五：雙向對話建議系統與失語症對話輔助系統

執行雙向對話建議系統時，App會將錄製語音傳給Whisper模型轉換為文字，結合對話歷史後直接由LLM生成患者可能的回應選項。

執行失語症對話輔助系統時，說話錄影不只進行語音轉文字，也會進行中文唇語識別。說話錄影每0.2秒取一幀，以ORB進行關鍵幀提取。將關鍵幀以多模態模型檢測生成特徵序列。關鍵幀橫向拼接成一個膠卷圖作為LLM的其中一個輸入，提供LLM 更多對話細節。如圖2-11。



(圖源：自製)

圖 2-11：失語症對話輔助系統



(四) 思維鏈 (COT) 思考流程與代理人架構

1. COT 思考流程

本系統的 COT 流程透過多層次推理資訊：先判斷提問類型確立語句結構，接著辨識關鍵詞與語意，並分析場景圖與膠卷圖以補充語境線索，然後結合情緒與手勢判斷語氣，若有手繪圖則檢視相關符號。最後生成忠於患者原意的五句繁體中文選項，語法完整且不添加額外內容。

2. 創新代理人架構



圖2-12：Agent模式架構 (圖源：自製)

(五) 系統效果驗證

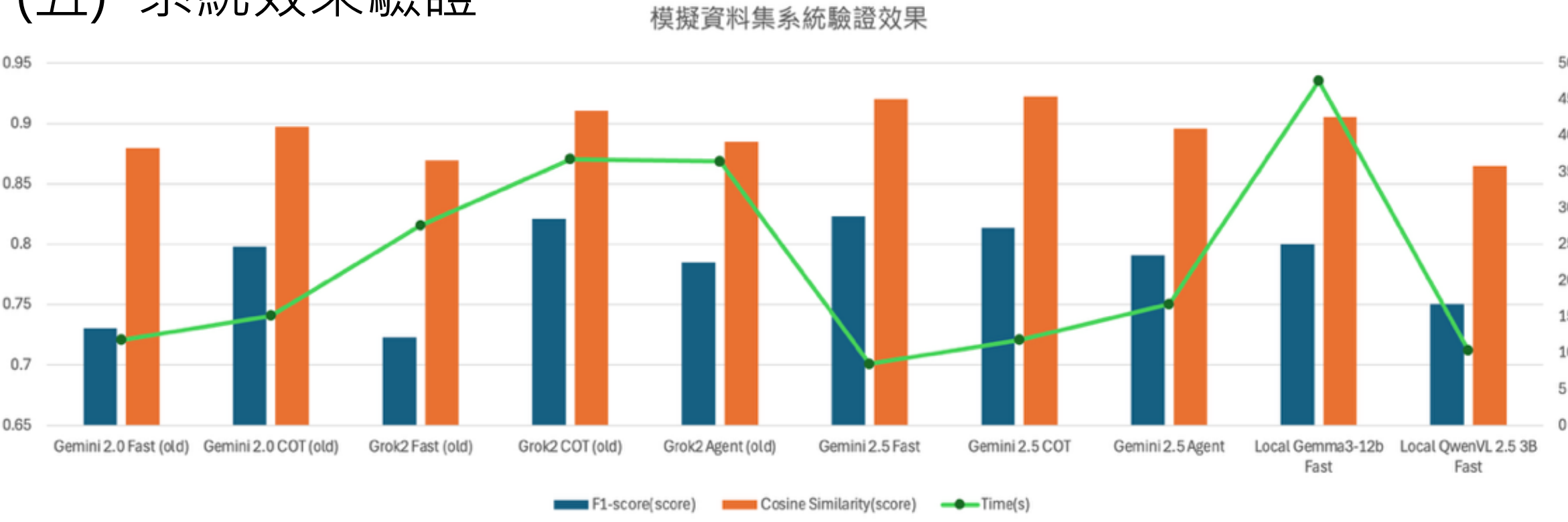


圖2-13-1：對話輔助系統在V<sub>S</sub>資料集的效果

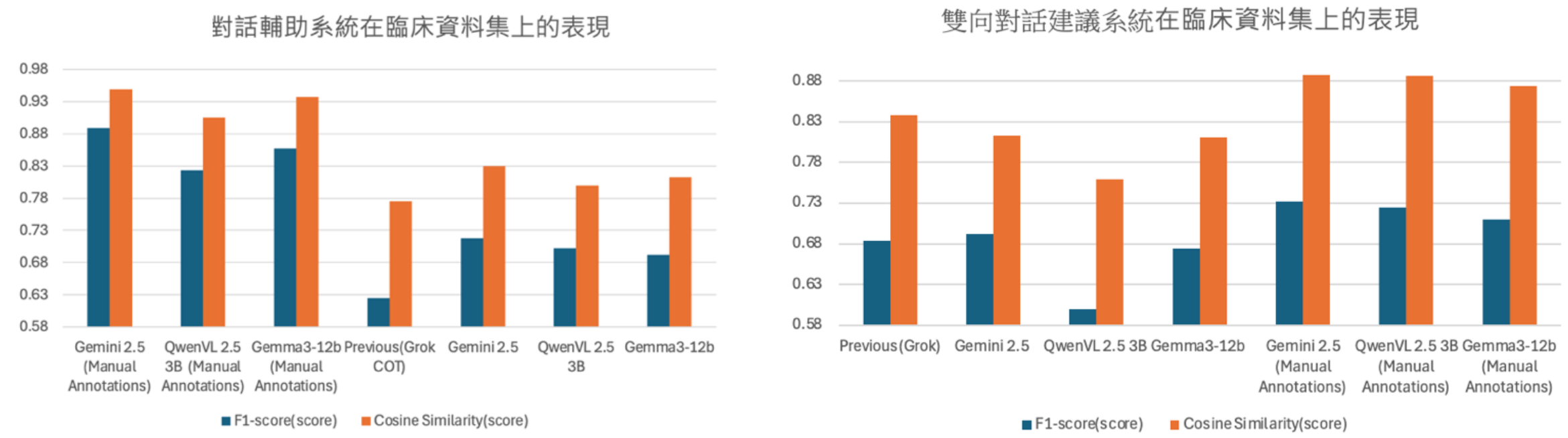


圖2-13-2：對話輔助系統在V<sub>C</sub>資料集的效果

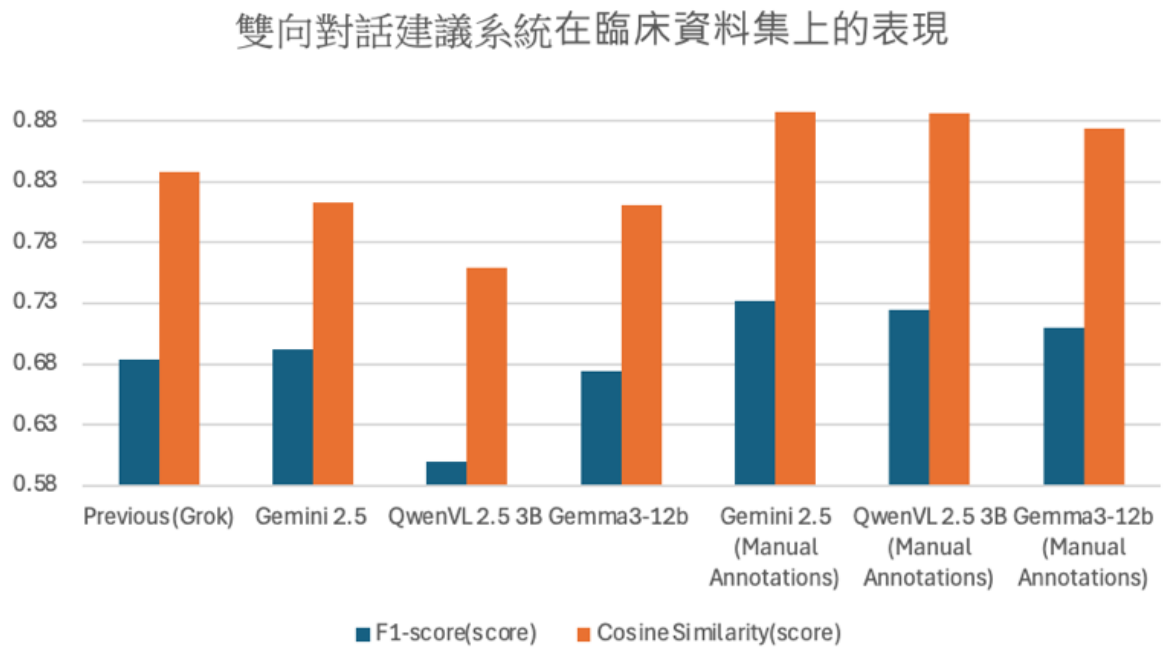


圖2-13-3：雙向對話建議系統在V<sub>C</sub>資料集的效果 (圖源：自製)

圖2-13-1中，Gemini 2.5 Flash響應最快且具有最高的準確度。COT與Agent策略在舊模型上能提升表現，但在Gemini 2.5 上效果有限，推測因模型已達性能上限，單純增加推論步驟難以再突破。開源模型部分Gemma3 12b準確度最佳而Qwen2.5VL3b響應速度與頂尖模型相當。圖2-13-2與2-13-3為系統在V<sub>C</sub> 資料集的效果，再排除背景雜音干擾後，系統表現良好。

實驗六：系統效率優化策略

(一) 影片取幀優化

(圖源：自製)

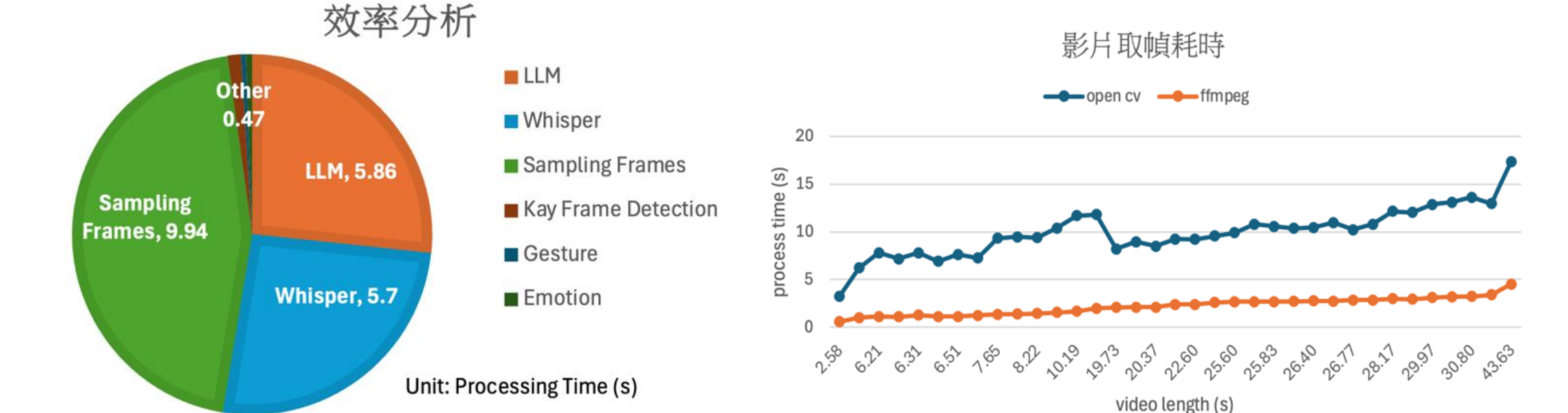


圖2-14-1：初代系統效率分析

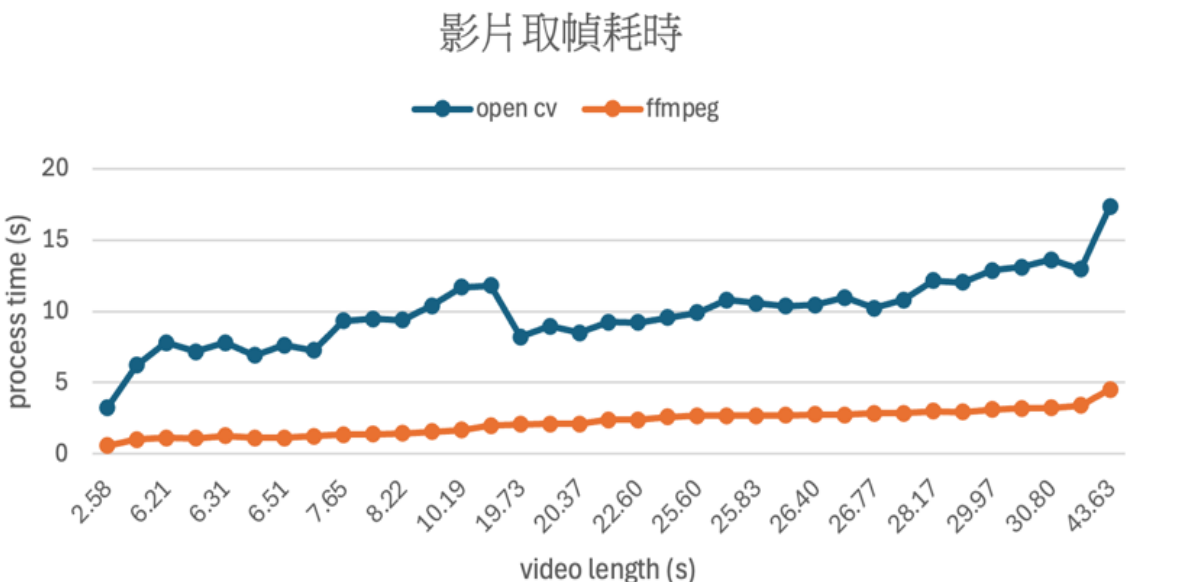


圖2-14-2：影片取幀優化

(二) 異部優化：以此提升系統效率 (見圖2-7至2-11)。

(三) 輕量Flux 文生圖與圖像資料庫優化

如圖2-14-3所示，圖卡即時建議系統雖加入異步處理，Pexels API 處理時間仍然偏長(橘色線段)。因此結合個人圖像資料庫與輕量級Flux生成模型，使系統可直接從資料庫中取得已生成的圖卡，同時Flux模型加快獲得圖卡的時間(藍色線段)。

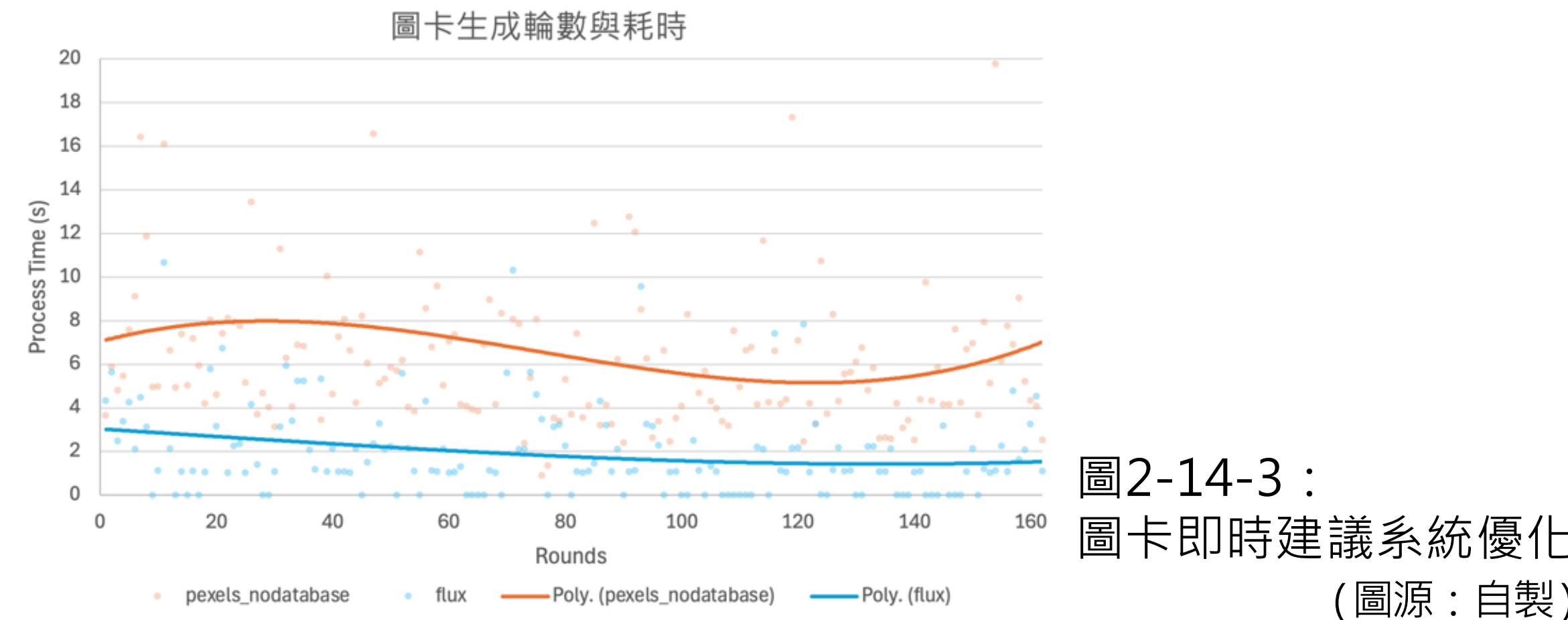


圖2-14-3：圖卡即時建議系統優化 (圖源：自製)

(四) 停用唇語識別模型

由於E2E AVSR模型泛化效果不佳，甚至誤導系統做出錯誤的語意猜測。然而自行訓練唇語模型需要大量算力以至於較難實現。因此系統驗證階段暫時停用該模態。

(五) 伺服器資源分配與平行處理

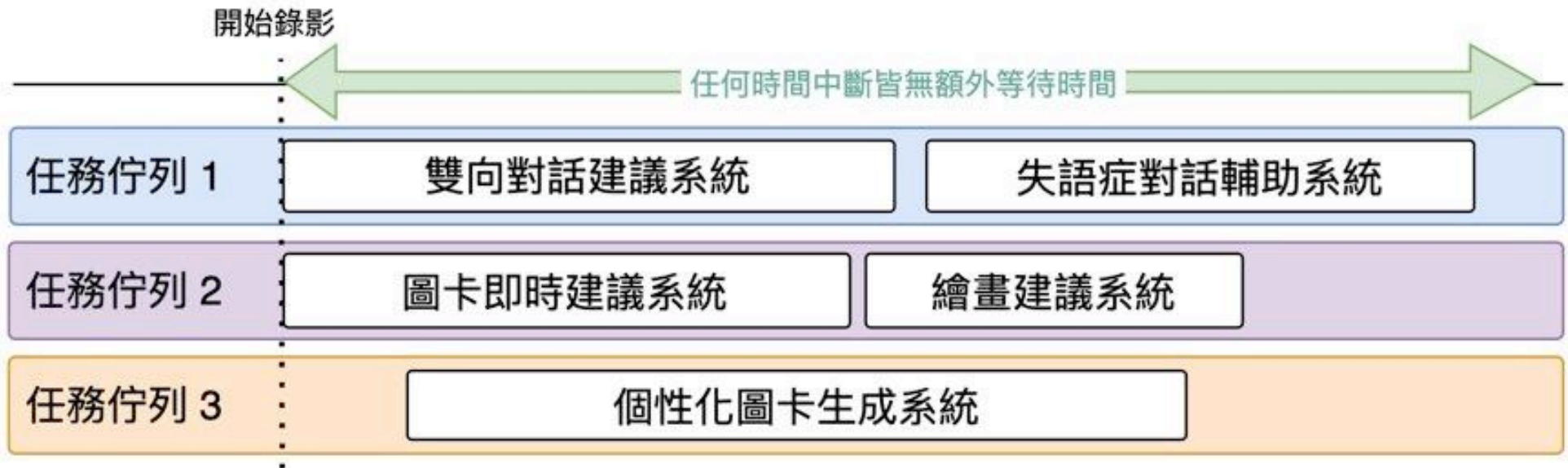
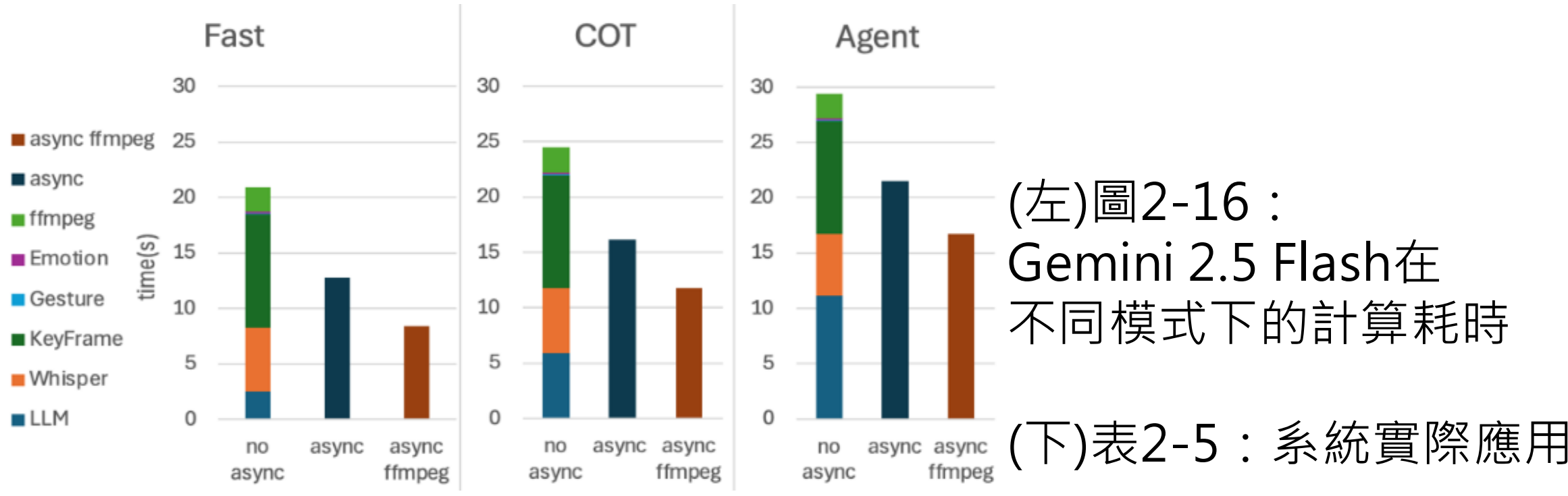


圖2-15：伺服器任務佇列優化 (圖源：自製)

(六) 優化效果驗證

(圖源及表源：自製)



(左)圖2-16：Gemini 2.5 Flash在不同模式下的計算耗時

(下)表2-5：系統實際應用

對話場景	
對方	您好，請問需要什麼嗎？
患者 (主動敘述)	我我 大大 水
系統回覆	我想要大瓶的水。/ 我要買大瓶的礦泉水。/ 請給我大罐的水。/ 我需要大瓶裝的水。/ 有沒有賣大瓶的水？
對方	有的，我們有 1 公升和 1.5 公升的，您要哪種？
患者 (被動敘述)	(被動敘述模式，無需表達)
系統回覆	我要 <b>1.5</b> 公升的。/ 請給我一公升的。/ 給我大的。/ 我要大瓶的。/我要 1.5 公升。
對方	您是要買一瓶還是兩瓶呢？
患者 (主動敘述)	(Action: 手勢比出 2) ..... 才夠
系統回覆	我要兩瓶。/我總共要買兩瓶。/是的，請給我兩瓶。/我要買兩瓶，謝謝。/兩瓶。
對方	你還要買什麼？
患者 (主動敘述)	(選擇圖卡) No/Price/Pay/Bag/Cookies
系統回覆	我還想要餅乾和一個袋子。/我需要買餅乾，還有袋子。/給我餅乾和袋子，謝謝。/我想要一些餅乾，請給我一個袋子。/還有餅乾，也要一個袋子。
對方	好的，總共是 70 元，請問要用現金還是電子付款？
患者 (主動敘述)	(白板手繪) (畫了一個手機)
系統回覆	我要用現金。/我要用電子付款。/我要付現。/我要刷卡。/我要算一下。

實驗七：基於Qwen2.5VL 3b的AphasiaLite-LLM微調

我們採用資料集蒸餾策略，以Gemini 2.5 Flash的輸出為基礎對Qwen2.5 VL 3B進行微調。

- (一) 系統要求Gemini2.5 Flash產出五個可能的還原語句。
- (二) 將五個語句與S<sub>F</sub>計算F1-score，挑選出F1-score最低者，以Ground Truth (即S<sub>F</sub>) 替換之。
- (三) 依照F1-score高低排序。
- (四) 結合多模態資料合成微調用資料集並進行LoRA微調。

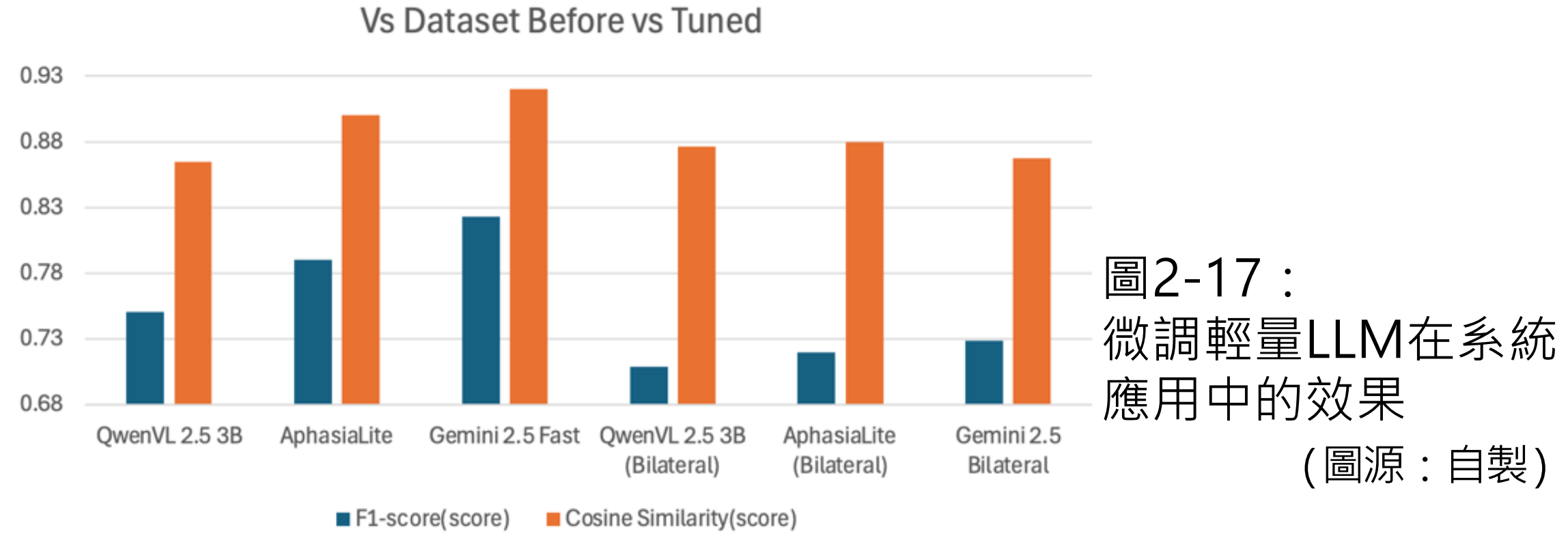


圖2-17：微調輕量LLM在系統應用中的效果 (圖源：自製)

如圖2-17，AphasiaLite-LLM經微調後，F1-score提高0.04，接近Gemini 2.5 Flash。未來若能若能持續累積更多實際應用資料，預期能進一步提升模型效能，實現高效能且低成本的開源部署方案。

失語症醫療專業人員訪談與建議

我們與神經內科專科醫師訪談，並得到以下建議後於本研究中改進：

- 1. 系統應用範圍應限定於非流暢型的失語症患者。
- 2. 不同患者的表達能力差異較大，增加選項數量與提升語意多樣性可提升輔助能力。
- 3. 部分患者有閱讀困難，系統應減少文字閱讀需求，降低對文字閱讀的依賴。

結論

- 一、AphasiaSim-LLM 的方法能有效生成模擬失語症語句。
- 二、還原失語症語句的任務上Gemini 2.5 Flash效果最佳，而開源模型亦有其優勢。
- 三、ORB關鍵幀法兼具快速與準確的特性，適合作為即時影片處理時的關鍵幀算法。
- 四、商用模型 Gemini 2.5 Flash Fast 模式於對話輔助任務中表現最優，提示詞工程的效果趨於飽和。
- 五、透過異步處理、FFmpeg 編解碼優化、Flux 模型應用、個人化圖像資料庫的建置，以及伺服器資源的有效分配與並行處理策略，能夠大幅提升整體系統效能。
- 六、採用資料集蒸餾策略微調輕量模型AphasiaLite-LLM，能顯著提升失語症語句還原效能。
- 七、本研究結合五大子架構實現之系統能有效地協助失語症患者進行流暢溝通。