二、研究計畫內容：

(一) 摘要

傳統手語翻譯系統受限於固定規則，難以有效處理語境變化與複雜語法，因此本計畫致力於開發基於多模態深度學習與大型語言模型的自動手語翻譯系統。透過解析手部動作、姿態、表情等多模態特徵，並結合語意對應與句子生成技術，提升翻譯的準確度與自然度。本計畫將與台灣手語翻譯協會合作，建立高品質的手語數據集，並進行專業標註，以確保模型的準確性與泛化能力。研究成果可廣泛應用於教育與公共服務的智慧翻譯系統，推動資訊無障礙環境的發展，並提升聽障者的社會參與與溝通便利性。

(二) 研究動機與研究問題

「人工智慧 (AI) 不僅是一種技術，更是一種以人為本的應用，能夠賦能社會並帶來實質改變。」然而，弱勢族群對新興科技的接受度較低，且缺乏相關技能與輔助工具，導致其生活質量進一步受限。在全球社會支持體系中，特別是身心障礙者輔助領域，長期面臨人力嚴重短缺的問題。由於專業人員數量有限，導致供需失衡，使許多有需求的人無法獲得即時且有效的支持。這不僅使弱勢群體更難融入主流社會，也加劇了他們的社會邊緣化。若能將科技導入現有的支持體系，不僅能作為輔助工具，更能為弱勢群體提供自我展現與發揮潛能的機會，進而減輕人力資源短缺的壓力，並成為社會支持體系的重要能量，實現更普及且公平的社會關懷。

手語的挑戰與社會現況

受文化差異、語言演變、社區影響、地理隔離及社會認可程度等多重因素影響，全球各地的手語發展出獨特的視覺語言特徵。然而，長期以來，許多國家的教育體系推行「口語至上主義」，強調聽障者應以口語溝通，而忽視甚至排斥手語的使用，導致手語文化逐漸邊緣化。這種將口語視為優於手語的偏見，不僅影響聽障者的文化與身份認同，也限制了其多元溝通的可能性。事實上，聽障者的溝通方式是多樣且個別化的，無論是手語、口語，或兩者結合，都應受到同等的尊重。聽人社會應摒棄將聽障者視為「需要修復」的觀念，轉而支持並尊重手語文化，營造一個更加多元包容的社會環境，讓聽障者能自在表達、積極參與社會生活，並在自我認同與相互理解中擁有更穩固的生活基礎。雖然手語已被列為台灣的法定語言之一，但在政策實踐上，仍常被視為一種「社會福利」，未能享有與其他語言翻譯服務同等的重視。例如，在國家通訊傳播委員會舉辦的聽障平權會議中，因預算不足未安排手語翻譯員，導致會議無法提供即時手語翻譯服務，這反映出台灣在推動手語平權上的困境與挑戰。

手語翻譯服務的不足

目前，除了實體手語翻譯服務外，中華電信與衛生福利部近年也積極推動「手語視訊轉譯服務」，讓手語使用者能透過社群平台與聽人順利互動。然而，專業手語翻譯員人數嚴重不足，導致服務的可用性與覆蓋範圍受限。此外，翻譯員需承擔高度心理壓力，特別是在報案電話或緊急醫療通報等高風險情況下，翻譯內容的不確定性更增加了職業倦怠的風險。目前台灣約有數十萬名聽障者，但全國僅有 616 名合格的手語翻譯員，平均每人需服務約 300 名聽障者，服務比例遠低於日本[1]。這使得許多聽障者在溝通上必須依賴文字，但文字的溝通方式不僅效率較低，還容易引發誤解或資訊延遲，進一步凸顯其日常交流的困境。手語不僅是一種溝通工具，更是一種擁有獨立語法與文化價值的語言。若將其僅視為輔助工

具，便會忽略其作為語言與文化資產的意義。因此，提升手語的社會地位與資源投入，不僅能促進聽障者的社會參與，也能展現多元與包容的社會價值。

研究動機與技術挑戰

為了緩解手語翻譯人力資源的不足，自動手語翻譯技術成為一種關鍵解方。透過人工智慧技術的應用，可望提升手語翻譯的可行性、減少等待時間，並提供更穩定且高效的溝通服務。此外，若能同時強化手語專業人才的培育，才能真正實現無障礙溝通的目標，確保聽障者能更方便地獲取所需要的資訊與服務。目前，手語學習的普及仍面臨諸多挑戰，包括學習資源的缺乏、師資不足，以及學習者難以獲得即時且準確的反饋。因此，若能透過手語教學與評分系統，提供隨時隨地的學習機會，將能有效突破時間與空間的限制，特別是對於缺乏手語教師資源的地區，這項技術能大幅提升手語的普及率。透過科技輔助推動數位平權，不僅能促進手語使用的便利性，還能縮小聽障者與聽人之間的語言隔閡。然而，現有的手語辨識與翻譯技術仍面臨諸多限制，包括：

1. 多模態資訊整合不足：多數手語辨識模型未能有效融合手勢、面部表情與身體動作，難以準確理解語境與語法結構。
2. 數據資源有限：缺乏大規模、高品質的手語數據集，使模型的訓練與優化變得困難。全球各地的手語種類繁多，且手語的數位化、標註過程需投入大量人力與資金，進一步限制了技術發展。
3. 應用場景不足：現有系統多偏向研究性質，未能廣泛應用於日常場景，例如即時對話、手語字幕生成等實際需求。

此外，人類在溝通時常伴隨各種肢體動作；手語辨識技術在實踐中面臨非語言資訊解析的挑戰，像是如何同時整合手勢、表情與身體動作，並精準萃取關鍵特徵。而在語言輸出的階段，系統仍可能因語法錯誤或語意不連貫而影響使用者體驗，特別是在處理複雜或模糊的手語表達時，更需要進一步的技術突破。

本研究將透過資料分析與現有技術的探討，了解手語辨識與翻譯的挑戰，並發展更精準且高效的解決方案，以縮小手語使用者與主流社會之間的溝通鴻溝，促進數位平權，並推動社會對手語文化的重視與尊重，創造更具包容性的環境。

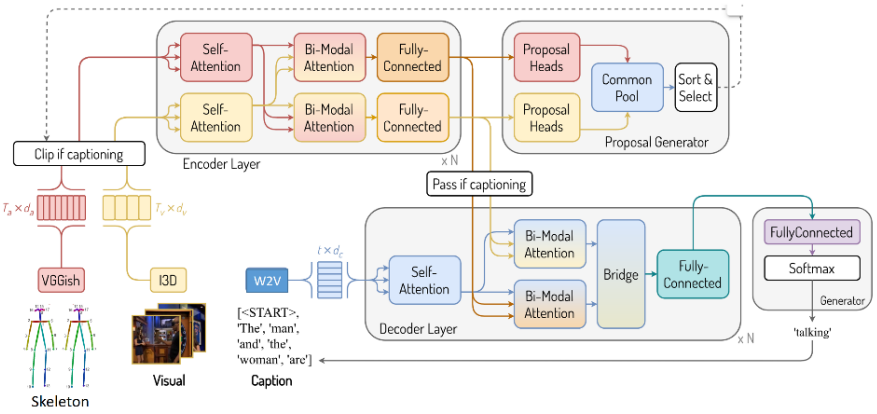
(三) 研究方法及步驟

早期的自動手語翻譯系統主要依賴規則或預設模式，透過固定的手勢字典進行比對。然而，這種方法在面對手語的上下文變化與複雜語法時，往往難以準確呈現完整的語意。隨著大型語言模型與多模態技術的快速發展，現今的系統已能從影像中捕捉關鍵場景，並生成相對應的文字描述，涵蓋更豐富的內容與細節。這些模型不僅能分析手勢動作，還能識別面部表情與肢體語言，例如手語表達問句時，會透過臉部表情來強調疑問語氣。透過這些技術，手語影像可轉換為更精準的文字描述，同時保持上下文的連貫性，從而為聽障者與聽人之間建立更流暢的溝通橋樑。手語是一種視覺語言，其語法結構與口語存在顯著差異，並非與文字一一對應。自然手語的表達多源自於生活事物或動作的模仿，其核心特徵是「表意不表字」。因此，傳統逐字翻譯的方法難以適應手語的靈活性與多樣性。基於深度學習與大型語言模型的手語翻譯系統，則能有效解決手語表達順序不固定的問題。由於不同手語使用者的表達方式可能不一，甚至語法結構完全不同，該方法的核心在於提取句子的關鍵詞，降低對手語順序的依賴

性，進一步提升語意的準確度。透過深度學習模型對多模態特徵進行提取，再由大型語言模型負責語句潤飾，系統能夠準確理解手語語意，並生成符合口語習慣的表達方式。這種架構整合了多模態深度學習與語言模型的優勢，兼顧語義準確性與系統靈活性，為手語翻譯應用提供創新的解決方案。

手部、姿態與表情的擷取與向量表達

在手語翻譯過程中，為了提升辨識準確度與上下文理解能力，系統將採用多模態融合技術。手語不僅包含手勢動作，還涵蓋臉部表情、身體語言，甚至某些情境下的語音或文字輔助。因此，將視覺與語言資訊結合，能夠顯著提升翻譯的精確度與流暢度。系統將對使用者的手語表達進行特徵提取，包括手部動作、手指細節、肢體姿態以及面部表情等資訊（可利用 mediapipe 來擷取），重點在於識別關鍵動作與表情。影像與文字特徵將透過編碼器進行嵌入處理（Embedding），確保多模態訊息能夠有效對應。為了提升模型對手語的理解能力，研究將採用預訓練特徵提取技術，基於現有的視覺與語言預訓練模型，如 I3D（Inflated 3D Convolutional Networks）與 BERT，使系統能夠從不同維度學習豐富的手語特徵。I3D 負責影像處理，確保系統能夠捕捉手語的時間與空間變化，而 BERT 則負責語言模態的特徵學習，使手語動作與自然語言能夠準確對應[2]。由於手語表達具有連續性，不同手勢之間的邊界往往不明確，因此，模型將能夠自動識別手語序列中的關鍵時間點。這種多模態預訓練策略，將使模型在正式訓練前即具備一定的手語理解能力，並透過微調（Fine-tuning）適應不同的手語變體，提高系統的適應性與泛化能力。



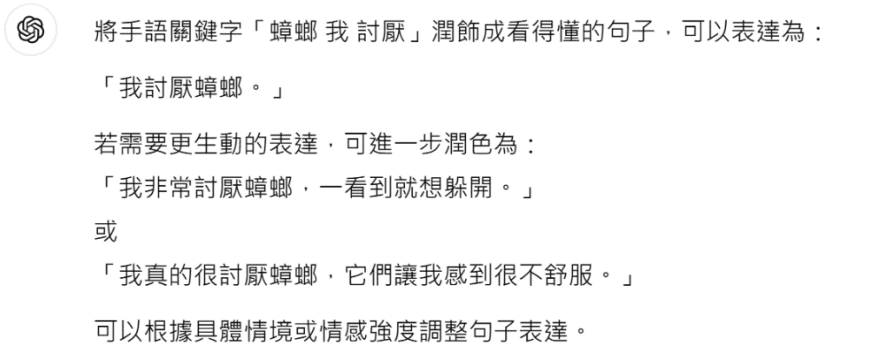
圖一　手語翻譯模型簡述[2, 3]

動作與文字的結合

本研究的模型架構將基於 Transformer 設計，以取代傳統的遞歸神經網絡（RNN）或長短期記憶網絡（LSTM）。由於手語具有強烈的時間序列特性，準確的語意理解需要掌握長距離的時序關係，而 Transformer 的自注意力機制（Self-Attention）能夠有效捕捉序列中各個時間步驟的權重，避免 RNN 常見的長期依賴問題（Long-term dependency problem）。透過交叉注意力機制（Cross-Attention）來建立視覺與語言之間的關聯性，融合技術將確保系統能夠精確對應影像資訊與語言輸出，從而在不同手語場景中達成更高的辨識準確性。在此基礎上，模型將進一步加入多頭注意力機制（Multi-Head Attention），以捕捉不同層級的手勢特徵，如手指細節、手掌運動軌跡、身體姿態變化等，使系統能夠更全面地學習手語的空間與時間資訊。

大型語言模型的語意潤飾

透過大型語言模型，系統將對提取出的手語關鍵詞進行語言生成，使其轉換為符合自然語言表達的句子。此階段的重點在於語義整合與語法潤飾，使翻譯結果更加符合口語或書面語習慣。語言模型具備上下文理解能力，能根據手語的語境提供適當的詞彙選擇與句子結構，使翻譯結果更加流暢且符合語境。



圖二 大型語言輔助手語關鍵字成為句子

資料整合與標註

本研究將與社團法人台灣手語翻譯協會合作，以獲取高品質的手語數據集。協會提供的多樣性生活手語影片涵蓋豐富的語境，包括日常對話與專業領域內容，確保模型能夠學習不同情境下的手語表達方式。透過專業手語翻譯人員的參與，數據標註與校對將確保準確性與一致性，進一步提升深度學習模型的訓練效果。此外，協會專家將參與系統的現場測試，記錄機器翻譯的語境適配能力，確保最終系統能夠有效支持手語使用者與一般聽人的溝通需求。

(四) 預期結果

本研究預期透過多模態深度學習與大型語言模型的結合，實現高準確度且具語境適應能力的自動手語翻譯系統。系統將能有效解析手語影像中的手部動作、手指細節、肢體姿態與面部表情，並透過預訓練模型與語言生成技術，轉換為符合自然語言習慣的流暢句子。預期成果可從技術、應用與社會影響三個層面進行分析。在技術層面，本研究以 Transformer 為核心的手語翻譯框架，取代傳統 RNN 或 LSTM 架構，提升時序資訊的捕捉能力。透過自注意力機制與交叉注意力機制，模型能更準確地對應手語影像與語言輸出，降低手語表達順序的影響。此外，系統將結合 I3D 影像處理技術與 BERT 語言模型，以自動識別手語的關鍵時間點與語意對應，提高翻譯準確度與語境適應能力，從而達到更自然的溝通體驗。為了強化模型的泛化能力，本研究亦將導入自監督學習與對比學習技術，使系統能夠學習不同使用者的手語變體，確保其適用於多樣化的手語表達方式。在應用層面，此研究將透過與台灣手語翻譯協會的合作，建立高品質的手語數據集，並確保系統的翻譯結果符合專業標準。研究成果將可應用於即時手語翻譯系統，並可進一步拓展至教育、醫療、公共服務等領域，幫助聽障者更順利與社會溝通。此外，本研究開發的手語語料庫與預訓練模型，將作為後續相關研究的重要基礎，並進一步推動手語技術的發展。本研究將降低手語使用者與一般社會大眾之間的溝通障礙，促進資訊無障礙環境的發展，透過提高手語翻譯技術的準確性與自然度，讓聽障者在教育與就業場域能夠更積極參與，並對智慧助理與人機互動技術的發展提供新的可能性。

(五) 需要指導教授指導內容

如何選擇適合的深度學習架構，有效整合影像處理與自然語言處理技

術，進而提升手語轉換的準確性與流暢度。此外，模型訓練流程的設計應包括超參數調整、特徵提取技術的應用，並探討如何進一步提升模型的適應性與泛化能力，以確保系統能應對不同的手語變體與語境。期望能透過老師的建議與協助，成功完成此關心社會議題的跨域科技研究計畫。

(六) 參考文獻

[1] https://www.chnnews-tv.com/16/36481/

[2] Iashin, Vladimir and Rahtu, Esa, A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer, British Machine Vision Conference (BMVC), 2020

[3] Iashin, Vladimir and Rahtu, Esa, Multi-Modal Dense Video Captioning, The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020