第二十三屆旺宏科學獎

成果報告書

參賽編號：SA23-609

隊伍名稱：芝麻街伯特

作品名稱：自監督學習在臺灣手語辨識上之應用研究

參賽類別：資訊

關鍵字：機器學習、自監督學習、臺灣手語

摘要

在臺灣手語辨識，先前研究所使用的監督式學習需要大量標記樣本而限制可辨識詞彙量。為此，本研究借鑒自然語言處理領域中 BERT 的遮罩想法，將未標記手語影片隨機遮蓋部分幀數，並讓模型學習預測被遮蓋的幀數以學習臺灣手語的特徵，並透過遷移學習來訓練辨識模型，此作法可克服現有臺灣手語資料缺少的問題。經過實驗，本研究訓練之詞彙辨識模型達成了 242 個詞彙量，94.8%的準確率。

此外，先前研究皆未在手語句子翻譯上有成果。因此本研究基於預訓練模型，整合設計手語翻譯的系統，在實驗中，整個系統在 100 個句子的翻譯表現達到 88%的準確率，證明自監督學習的方式在手語辨識、翻譯上是有效的。並展現出樣本需求少與辨識詞彙量可輕易擴大的潛力。

壹、前言

一、研究動機

在 2020 到 2022 新冠疫情肆虐期間，衛福部每日會定時召開防疫指揮中心記者會，直播說明有關新冠疫情相關之案例報告、政策。研究者時常關注防疫直播，也時常注意到螢幕右下方有一位手語翻譯員，為聾人翻譯與會人士的發言。研究者十分好奇手語翻譯員所比出的手語意義，想要尋求軟體翻譯，但是在搜尋網路之後，發現市面上並沒有臺灣手語的翻譯軟體，因此本研究希望可以自己研究並開發臺灣手語翻譯的系統。

在此以前，不管是國內外，各種的手語都有做過類似的嘗試。國內亦有許多臺灣手語翻譯的相關研究，研究的主題主要是針對靜態手語詞彙的辨識，以及日常手語詞彙的辨識，但其所提出的方式均只能辨認少數手語詞彙，且不能翻譯手語句子。

目前研究普遍採用的是監督式學習來進行手語詞彙辨識，這種方式需要準備大量標記樣本來訓練模型。然而，準備這些樣本的過程非常耗時，且現有資料量十分稀少，這限制模型能夠識別的手語詞彙數量。而且，現在研究普遍都僅僅是單詞的辨識，對於在日常情況的手語句子翻譯未有成果。因此本研究希望能夠解決目前上述臺灣手語辨識研究上所遇到的問題。

二、研究目的

本研究希望透過自監督學習的方式訓練模型，使模型能夠自行學習到手語的特徵。此方案大幅降低所需要的標記樣本的資料，以擴大模型可辨識的手語詞彙。並自行設計手語翻譯系統，達到句子的翻譯。

（一）探討自監督學習應用在動態手語辨識任務上。

（二）探討如何用少數的標記資料完成手語辨識任務。

（三）研究如何將預訓練模型應用於手語翻譯。

貳、文獻回顧

一、遷移學習

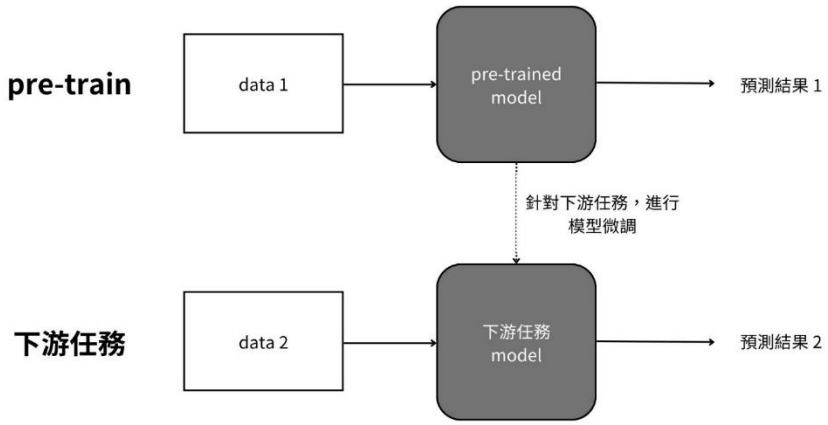

圖 2-1：Fine Tuning 示意圖（來源自行製作）

遷移學習（Transfer Learning）是一種機器學習方法，旨在將一個領域學到的知識應用到另一個相關領域。這種方法特別適用於當標註數據稀缺或獲取標註數據成本高的情況。遷移學習的核心思想是利用在一個大型數據集上預訓練的模型，然後將該模型應用到特定的目標任務上。這樣，不僅能夠節省訓練時間，還能提升目標任務的性能。

在遷移學習中，其中一個關鍵技術是微調（Fine Tuning）。微調是指在預訓練模型的基礎上，對模型的某些層進行進一步訓練，使其更適應新的目標任務。首先，選擇一個在大型

數據集上預訓練好的模型。這些模型通常是如 VGG、ResNet 或 BERT 等，已經在如 ImageNet 這樣的大型數據集上進行了充分的訓練，具備了良好的泛化能力。接著可根據目標任務的需求，對預訓練模型進行調整，然後，將調整後的模型在新的數據集上進行訓練。

微調的優勢在於，它能有效利用預訓練模型的豐富知識，顯著提升目標任務的模型性能，特別是在目標數據集較小的情況下。此外，微調還能節省大量的計算資源和訓練時間，因為大部分計算密集型的學習已經在預訓練階段完成。

二、Transformer 模型

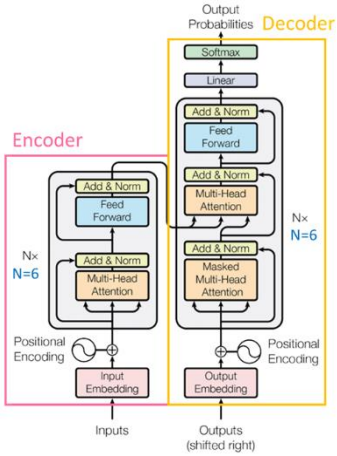

圖 2-2：Transformer 模型架構圖（來源：Vaswani, 2017 [1]）

Transformer（Vaswani et al., 2017）一開始用於自然語言處理 （NLP） 任務，如機器翻譯。它的設計摒棄傳統的循環神經網路 （RNN） 和長短時記憶網路 （LSTM） 等序列模型，採用全新的自注意力 （self-attention） 機制，使其在處理長序列時表現更好。

（一）自注意力機制 （Self-Attention）

Transformer 使用自注意力機制來捕捉輸入序列中不同位置的依賴關係。 每個輸入位置都與其他所有位置建立關聯，這允許模型在處理不同距離的依賴關係時保持高效率。

（二）位置編碼 （Positional Encoding）

由於 Transformer 沒有明確處理輸入序列的順序訊息，需要添加位置編碼來幫助模型理解單字的相對位置。

（三）編碼器-解碼器結構 （Encoder-Decoder）

Transformer 由編碼器和解碼器組成，適用於序列到序列的任務，在機器翻譯。 編碼器負責將輸入序列轉換為特徵，解碼器則將這個表示轉換為輸出序列。

Transformer 模型的出現在 NLP 領域引起革命性的變化，它不僅在翻譯任務上取得令人矚目的成果，也成為許多其他 NLP 工作的基礎模型，如 Bert、GPT 系列等等。

三、Bidirectional Encoder Representations from Transformers （BERT）

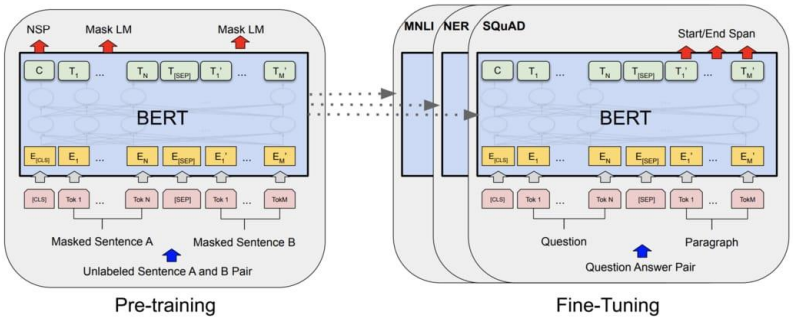

圖 2-3：BERT 示意圖（來源：Jacob Devlin, 2018 [2]）

BERT （Bidirectional Encoder Representations from Transformers）是由 Google 在 2018 年提出的預訓練語言模型，它在 NLP 領域取得巨大成功。 與傳統 NLP 方式不同，BERT 的獨特之處之一是它採用自監督學習的方法進行預訓練。 自監督學習是一種無監督學習的形式，其中模型從輸入資料中學習，而無需標籤。

（一）預訓練與微調

BERT 首先在大規模文字語料上進行預訓練，學習通用的語言表示。 然後，可以透過微調在特定任務上，例如文字分類、命名實體識別等，以適應特定的應用場景。

（二）Transformer Encoder

BERT 模型通常由 Transformer Encoder 組成，每個編碼器層都有多頭自註意力機制和前饋神經網路。 這些層允許模型學習不同層次的語言表示。

（三）掩碼語言模型 （Masked Language Model，MLM） 訓練任務

BERT 在預訓練階段使用一個掩碼語言模型任務，其中一些輸入詞被隨機遮蓋，模型需要預測這些掩蓋詞的標籤 。 這鼓勵模型學習更豐富的上下文表示。

自監督學習任務使得 BERT 能夠捕捉大量的語言知識，並且預訓練階段的學到的參數可以在各種 NLP 任務上進行微調，從而獲得更好的性能。自監督學習的想法是透過模型本身產生標籤，因此無需手動標註大量標籤資料。 這種方法使得模型能夠從大規模的未標記資料中學到有用的特徵，然後在特定任務上進行微調。

四、Vision Transformer

Vision Transformer（ViT）是一種將 Transformer 模型應用於電腦視覺任務的架構，由 Google 團隊在 2020 年提出。 與傳統的捲積神經網路（CNN）不同，ViT 利用 Transformer 的自註意力機制來處理影像的全局關係，將影像劃分為一系列均勻的圖塊，然後將每個圖塊的表示作為輸入序列。

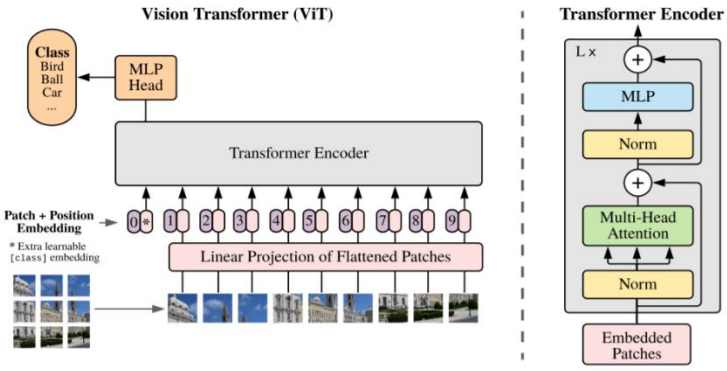

圖 2-4：Vision Transformer 模型架構圖（來源：Alexey Dosovitskiy, 2021[3]）

（一）影像分割為圖塊 （Patching）

輸入影像被分割為大小一致的圖塊，每個圖塊被視為一個序列元素。 這樣，ViT 將影像的全局資訊引入 Transformer 模型中。

（二）Transformer 編碼器

與 NLP 中的 Transformer 類似，ViT 有著 Transformer 的 Encoder，用於學習影像中的特徵表示。

（三）類頭 （Classification Head）

ViT 的輸出通常透過一個分類頭 （classification head） 來處理，產生最終的分類結果。 這個分類頭通常包括一個全連接層，將 Transformer 的輸出對應到類別機率。

ViT 的引入對電腦視覺領域產生深遠的影響，尤其是在影像分類任務上超越傳統 CNN 的表現。 此外，ViT 的研究也展現 Transformer 應用在 NLP 領域以外的可行性，甚至超越。

五、Masked autoencoders （MAE）

Masked Autoencoders（MAE）是一種深度學習模型，主要用於無監督學習，特別是在處理圖像數據方面。這種方法靈感來自於自然語言處理（NLP）領域的成功技術，例如 BERT（Bidirectional Encoder Representations from Transformers）。MAE 的核心思想是在輸入數據中隨機遮蓋（mask）一部分內容，然後訓練模型重建（reconstruct）被遮蓋的部分。

MAE 通常由兩部分組成：一個編碼器（encoder）和一個解碼器（decoder）。

（一）Encoder

Encoder 的作用是處理輸入數據，但在此之前，會先隨機選擇並遮蓋數據的一部分。例如，在處理圖像時，會隨機遮蓋圖像的一些像素或區域。編碼器只對未被遮蓋的數據進行處理，從而學習到數據的內在特徵和結構。

（二）Decoder

Decoder 的目標是根據編碼器處理過的數據來重建原始數據的遮蓋部分。這個過程迫使模型學習數據的重要特徵，因為它需要理解和推斷遮蓋部分的內容。

在 MAE 的訓練過程中，首先會隨機選擇並遮蓋輸入數據的一部分，然後編碼器對剩餘的未被遮蓋數據進行處理，提取特徵。接著，解碼器嘗試重建被遮蓋的部分。這一過程涉及到損失函數的計算，用以衡量重建數據與原始數據之間的差異。最後，通過反向傳播和參數更新，模型逐漸學會如何準確重建遮蓋的數據。

MAE 因其強大的特徵提取能力，在諸如圖像分類、物體識別和圖像生成等領域展現出卓越的性能。此外，由於它屬於無監督學習，對於那些標記樣本較少的應用場景有用。

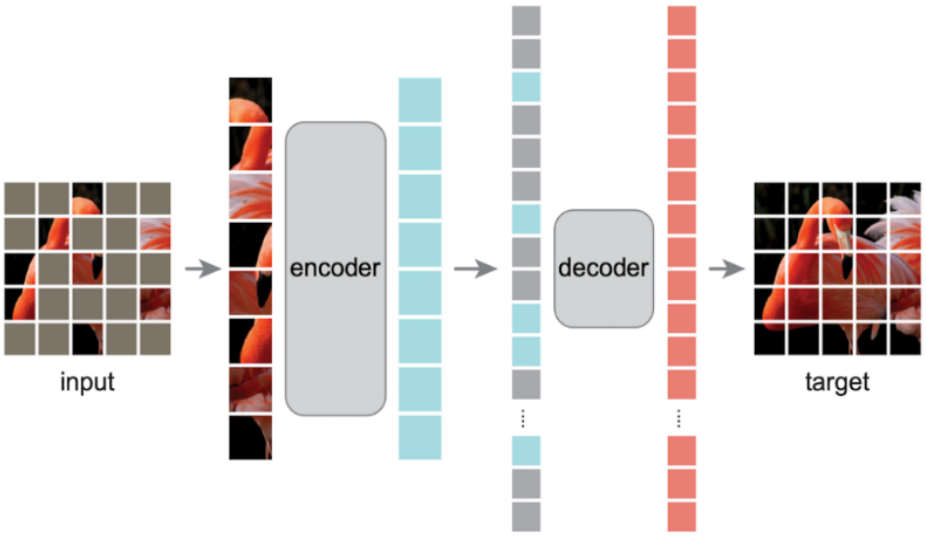

圖 2-5：Masked Autoencoder 架構圖（來源：Kaiming He, 2021[4]）

參、研究設備器材

一、設備：

	系統	GPU	CPU
筆記型電腦	Windows 11	RTX 3060 latop	i7 - 12400H
Google Colab	Ubuntu 22.04.3	Nvidia A100	Intel（R） Xeon

二、軟體：

軟體／套件	python	cuda	pytorch	opencv-python	mediapipe	GPT
版本	3.9.4	11.8	2.0	4.7.0	0.10.9	4

肆、研究過程與方法

受到前述 Masked Autoencoders 與 Vision Transformer 工作的啟發，本研究認為 Transformer 模型具有處理手語影片序列的潛力。本研究借鑑 Vision Transformer 處理圖片的方法來處理手語序列片段，同時利用 Masked Autoencoders 的自監督訓練方法，使整個模型能夠有效地學習手語的特徵，以應用在詞彙辨識任務中。

為了達到手語翻譯，本研究將預訓練模型進行 Fine tune，得到手語詞彙辨識模型。接下來，將手語影片分段辨識出詞彙，之後應用本研究者設計之滑動窗口演算法得出句子中所含詞彙，再將使用大型語言模型重組文句。

一、研究及實驗架構流程圖

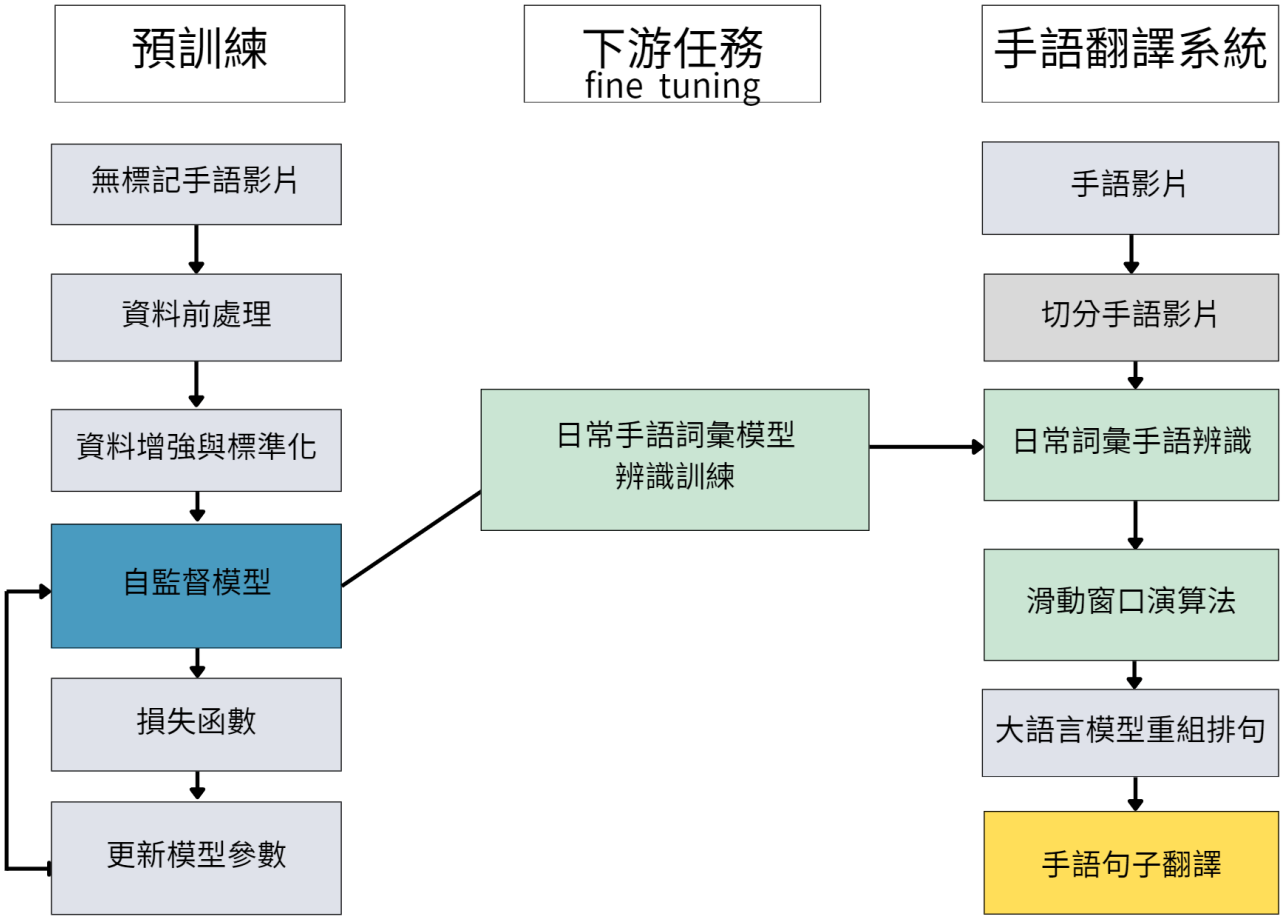

圖 4-1：研究流程架構圖（來源自行製作）

二、訓練資料的收集與處理

（一）資料來源

本研究發現疫情指揮中心定期召開的防疫記者會 （中央流行疫情指揮中心嚴重特殊傳染性肺炎記者會），旁邊配有手語翻譯人員，提供豐富的資料來源。本研究從防疫指揮中心平台上下載 481 部影片。

（二）MediaPipe 進行影像前處理

本研究將所有影片以 64 幀為一單位進行切分成片段子集$segment\ subset$ ，並且將畫面裁切成合適的大小，其中所有片段v_i為總長 64 幀，畫面大小為 640x640 的影片。在經過 python 套件 mediapipe 處理，標出翻譯人員手部的關節點。

在上一篇研究當中，研究者發現，如果輸入預訓練模型的資料中僅包含手部的點座標（如圖 4-2），那麼模型將會失去身體骨架的資訊，即無法區分手部在在身體的相對位置，然而，如果訓練資料包含了身體關節座標，則會失去對於手型的精細度，因此在新的研究當中，研究者決定同時訓練兩個模型，一個是訓練資料僅包含手部點座標的手型預訓練模型(Hand-shape Pretrained Encoder)，另一個則是骨架預訓練模型(Body Pretrained Encoder)，包含手部點座標以及身體骨架。

綜上所述，本研究定義兩種訓練資料，第一種是每筆資料P_m是儲存 64 幀中，每幀雙手 40 個點（一隻手各 20 點）以及關節點 4 點，臉部 1 點的(x, y)座標，共 45 個點。另一種P_m則是僅包含每幀雙手 40 個點（一隻手各 20 點）的的座標，共 40 個點的點座標（如圖 4-2）。（在此論文之中的研究過程以及研究結果與討論，皆以優先展示包含 45 個點的P_m為範例）

$$segment\ subset\ = \left\{v_1, v_2 \ldots v_n\right\}, v_i \in R_{64\times640\times640\times3}$$

$$dataset\ = \left\{P_{m1}, P_{m2} \ldots P_{mn}\right\},\ P_{mi} \in R_{64\times45\times2}$$

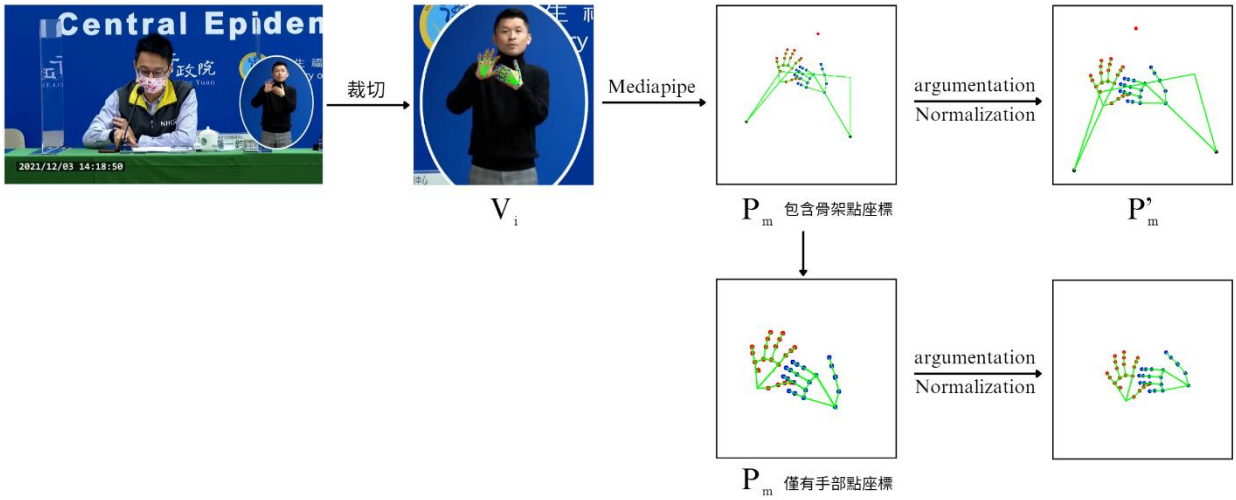

圖 4-2：資料前處理示意圖（來源取自防疫指揮中心）

三、預訓練過程

（一）資料增強與標準化

每一筆進入模型的資料P_m，在進入模型之前須經過縮放（Scaling）、旋轉（Rotation）、平移（Translate）、標準化（Normalization），以進行增強。本研究視P_m為$\{P_1, P_2, \ldots, P_{64*45}\}$（其中$P_i$為$P_m$中的所有點）以進行資料增強，具體操作公式如下：

$$\arg u\,mentation(P) = R(\theta) \cdot \big(s \cdot (P - C)\big) + C + v, -20° \le \theta \le 20°, 0.7 \le s \le 1.5$$

其中，$C = (x,y)$為錨點，$R(\theta)$表示旋轉矩陣，s是隨機的縮放因子，v是隨機的平移向量。這個過程對每一個點P_i都進行同樣的變換。

為了消除特徵間的尺度差異以及提高收斂速度，P_m在經過資料增強後，還需要進行標準化（Normalization)，公式如下：

$$P_m^{'} = \frac{P - \mu}{\sigma}, \mu = 335.49, \sigma = 134.28$$

在經過資料增強和標準化後，本研究得到轉換過後的資料P'_m。

（二）模型架構

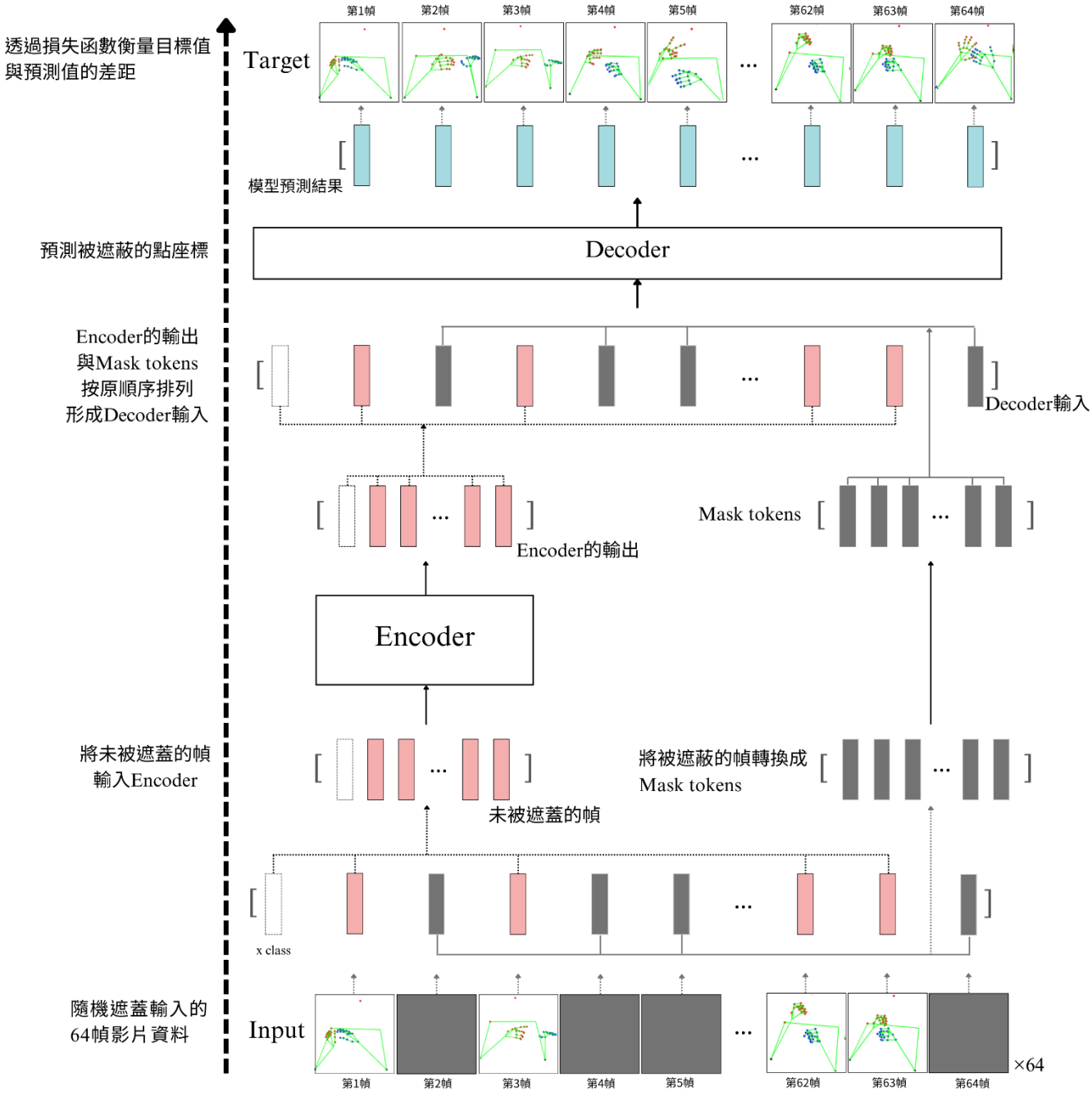

圖 4-3：模型流程概覽（來源自行製作）

本研究參考 MAE 研究的架構，在預訓練期間，將輸入的資料$P_m^{'}$中的幀依遮蓋率（mask ratio）隨機遮蓋，其中未被遮蓋的幀輸入進 Encoder，被遮蓋的幀轉換成 Mask tokens，隨後將 Encoder 的輸出與 Mask tokens 按原順序排列形成 Decoder 的輸入資料，將之輸入 Decoder 以預測被遮蓋的點座標，以此重建原始手語幀的座標形成預測值。最後透

過損失函數衡量目標值與預測值的差距，更新模型參數以最小化損失，使模型能夠學習到手語的特徵。在預訓練之後，Decoder 被移除，而未被隨機遮蓋的手語序列幀則輸入進 Encoder 以進行識別任務。

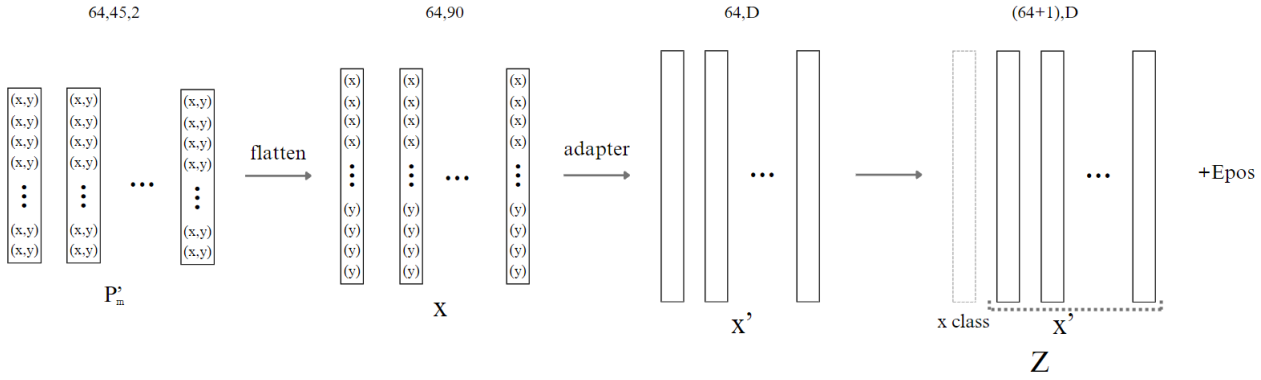

圖 4-4：步驟一的流程圖（來源自行製作）

如圖 4-4，本研究將資料$P_m^{'} \in R_{64\times45\times2}$轉換成序列$x \in R_{64\times90}$ 以符合 Encoder 所要求的 shape，而為了使輸入的x投射到設定的 Hidden size D，本研究通過設計一個轉接層 Linear Projection Adaptor 使$x \in R_{64\times90}$變成$x' \in R_{64\times D}$，公式如下：

$$x^{'} = adaptor(x) = xW + b, x \in R_{64\times90}, W \in R_{90\times D}, b \in R_{64\times D}$$

在 position embedding 的部分上，本研究選擇與 vision transformer 一樣的方式，在序列x'前連接一個 lernable embedding x_{class}，再加上 position embedding E_{pos}，使整個序列保留位置的資訊，公式如下：

$$z = \left[x_{class}; x^{'}\right] + E_{pos}, E_{pos} \in R_{(64+1)\times D}$$

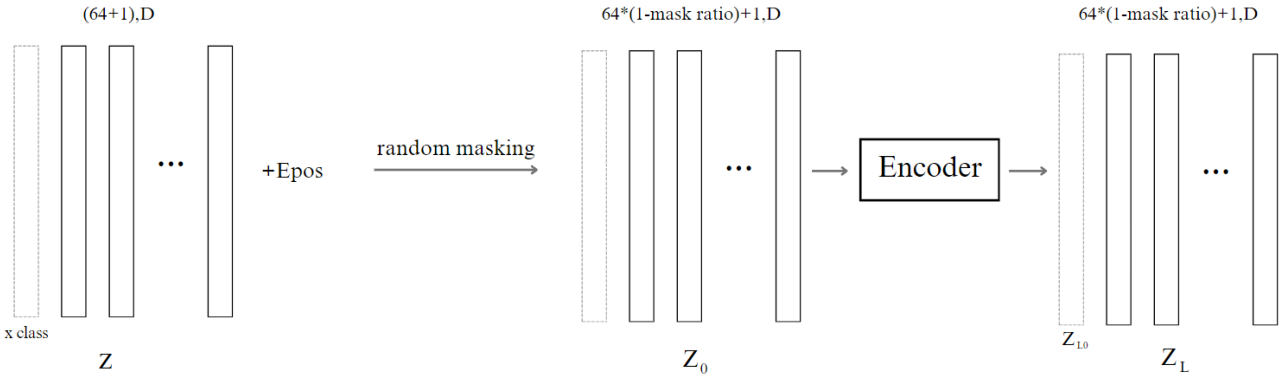

圖 4-5：步驟二的流程圖（來源自行製作）

如圖 4-5，在經過 position embedding 後，本研究遵循 Masked auotencoder 論文的方式，透過生成一個抽樣表，裡面包含隨機的編號，按照編號依遮蓋率（masked ratio）隨機將序列z中的部分 token 抽離，剩餘的形成新的序列 $z_0 \in R_{1+64*(1-mr)\times D}$（其中$mr$為設定的 masked ratio）作為 Encoder 的輸入。

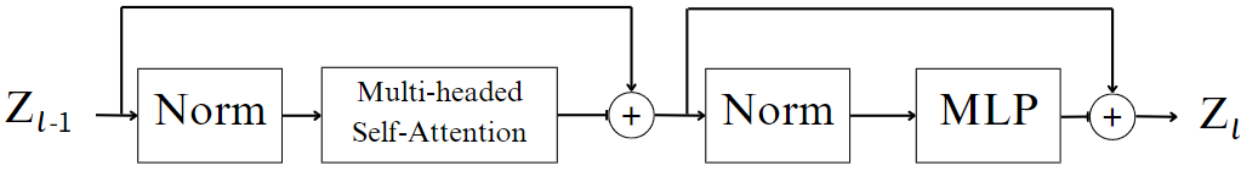

圖 4-6：Pre-Norm Transformer 的架構（來源自行製作）

如圖 4-6 所示，本研究選擇採用改良版的 Transformer encoder，稱為 Pre-Norm Transformer，與原始論文中使用的 Post-Norm Transformer 有所不同，其特性是在訓練的過程中對 learning rate 不那麼敏感，較為穩定。這個 Encoder 結構包括交替排列的 Multiheaded Self-Attention 和 MLP 層（參見公式 1、2）[3] 。在每一層的前面，本研究進行 Layer Normalization（LN），並在每一層的後面加入 Residual Connection。接下來，本研究將被隨機遮蓋的序列z_0輸入進 Encoder，並且得到 Encoder 的輸出$z_L \in R_{1+64\cdot(1-mr)\times D}$。 而$x_{class}$在 Encoder 輸出端的狀態$z_{L0}$（$z_L$的第 1 項）在經過 Layer Norm（$LN$)得到的預測輸出$y_{feature} \in R_{1\times D}$將作為手語的特徵向量（參見公式 3）。

公式 1 $z'_l = MSA(LN(z_{l-1})) + z_{l-1}, l = 1,2\ldots, L$

公式 2 $z_l = MLP(LN(z'_l)) + z'_l, l = 1,2\ldots, L$

公式 3 $y_{feature} = LN\left(z_{L0}\right)$

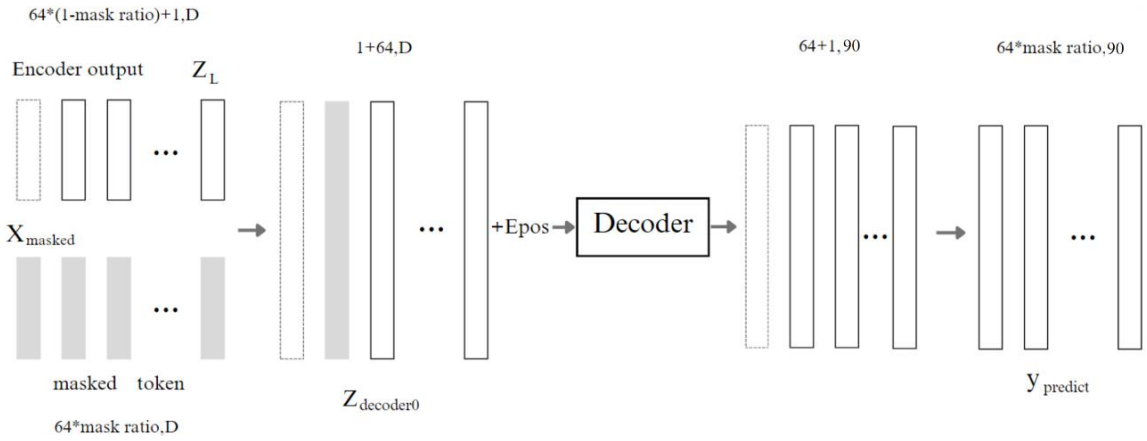

圖 4-7：步驟三的流程圖（來源自行製作）

以$x_{masked} \in R_{1\times D}$與 Encoder 的輸出$z_L$，按原先的順序排列後組成$Z_{decoder0} \in R_{(1+64)\times D}$，加上$E_{pos}$以保留位置資訊，再輸入進 Decoder。$Z_{decoder0}$包含完整長度的序列（如圖 4-7 所示），其中$x_{masked}$為共享的向量，表示需要預測的缺失訊息的存在。

本研究的 Decoder 也採用 Pre-Norm Transformer encoder，其中的層數比起 Encoder 較少。Decoder 的輸出在經過 prediction layer 後，按照抽樣表編號抽樣，將得到遮蔽部分的預測結果$y_{predict}$，公式如下。

$$y_{predict} = prediction\ layer(x) = xW \quad x \in R_{64\times D}, W \in R_{D\times 90}$$

其中x為 Decoder 的輸出

（三）損失函數（Loss Function）

為了衡量模型預測與實際目標之間差異，本研究選擇$MPJPE()$[6]作為指標，公式如下，其中$\hat{P}_i$為模型預測值的點，P_i為目標值的點。

$$MPJPE\ loss\left(\hat{P}, \mathrm{P}\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\left|P_i - \hat{P}_i\right|\right|_2$$

模型在預訓練中的目標是最小化這個差異。$MPJPE()$ 計算每個關節的預測位置與真實位置之間的歐幾里德距離，然後對所有關節的這些距離求平均，得到的結果是模型在二維空間預測關節位置的平均誤差。為了計算模型預測的 Loss，本研究將預測值$y_{predict} \in R_{64\cdot mr\times 90}$ Reshape 成符合計算$y'_{predict} \in R_{64\cdot mr\times 45\times 2}$，並將目標值$P'_m$並按抽樣

表的編號抽樣組成$target \in R_{64*mr\times45\times2}$，接下來計算$MPJPE\left(y'_{predict}, target\right)$，之後 Optimizer 會進行梯度下降，更新模型參數以最小化損失，使模型能夠學習到手語的特徵。

（四）預訓練實驗設置

在預訓練階段，本研究將$dataset$分成訓練集與測試集兩部分，其中訓練集有 78.8 萬筆，測試集有 4.1 萬筆。此外，在遮蔽率 mask ratio 設置上，我們在上一篇研究中已經證明 50%是最佳的數值，因此本研究在這選用了 50%進行預訓練。

	Encoder	Decoder
Layers	12	4
Hidden size	768	512
MLP size	3027	2048
Heads	12	16
Params	86M	40M

learning rate	1e-6
mask ratio	50%
batch size	128
optimizer	AdamW

表 4-1：模型訓練各項超參數設置

四、下游任務

預訓練過後，本研究將模型的中的 Decoder 移除，模型僅保留 Pretrained Encoder。一樣將單詞手語片段經過 Mediapipe 處理得到 P_m ，並進行標準化後輸入進 Pretrained Encoder ，此時，Encoder 的輸入是完整的 64 幀，不會被隨機遮蓋。

接著，在下游任務進行手語辭彙 fine tuning 階段，本研究實驗了四種不同的模型，以分析比較實驗數據。

（一）手型預訓練模型與骨架預訓練模型

本研究將手型、骨架預訓練模型分別接上 MLP 層與 Softmax 層進行 fine tuning（如圖 4-8），比較辨識準確率。

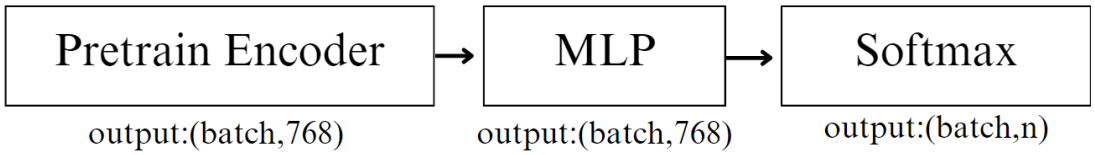

圖 4-8：模型架構圖（來源自行製作）

（二）融合模型

為進一步探索預訓練模型的潛力，本研究認為，骨架預訓練模型(Body Pretrained Encoder)較有能力關注到整體，而手型預訓練模型 (Hand-shape Pretrained Encoder) 較有能力關注到手型的細節，如果能夠整合此二模型，將會達到更好的辨識準確率。

因此，本研究將 Body Pretrained Encoder 以及 Hand-shape Pretrained Encoder 所輸出的 Feature 取出並且連接在一起組成$connective\ feature$，然後再接上 GELU Activation unction 跟 MLP 層，以及最後的 Softmax 層以用於分類（如圖 4-9）。

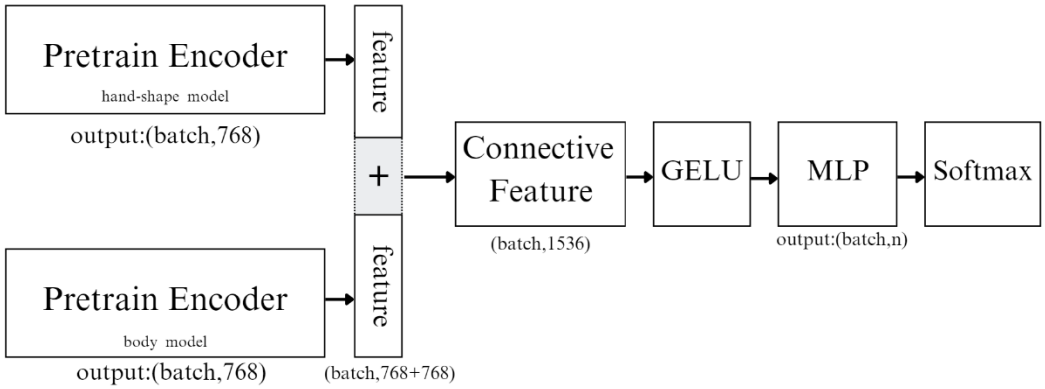

圖 4-9：模型架構圖（來源自行製作）

（三）照組組：在此模型，本研究將不會使用預訓練模型，僅使用隨機初始化的 Encoder，以此來對比是否使用預訓練模型的準確度差距。

五、日常詞彙手語辨識實驗

在辨識實驗中，本研究評估模型在 4 個日常詞彙手語辨識的能力，從中正大學手語辭典所提供的 500 多個例句中，挑選了 100 個句子，其中包含了 242 個日常使用字彙（表 4-2），因此本研究選擇了這 242 個詞彙進行實驗，實驗流程如下（圖 4-10）：

1. 請實驗者為 242 個日常詞彙手語，每一個類別錄製 5 次作為訓練集，並將四個模型設定 batch_size 為 64，進行 fine tuning。

2. 請受試者為 242 個日常詞彙手語，每一個類別錄製 1 次作為測試集，輸入進四個模型，以分析比較模型性能。

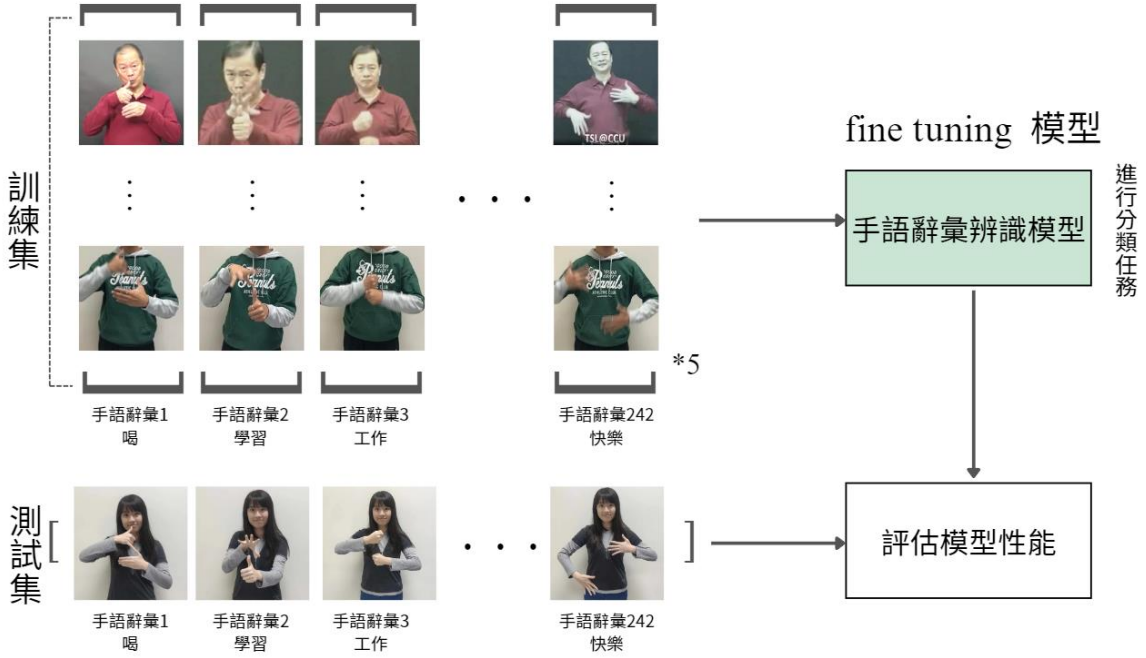

圖 4-10：日常詞彙實驗流程圖（來源自行製作）

2	夏天	記得	朋友	幫_N	錢	每天	如果	直	我們兩個	清楚	硬_A	英文	晴天_A
3	互相	完了	卡片	床	胖_A	家_B	昨天	不要_S	欣賞	刷牙	金	介紹	也許
4	可以	跑_B	旅行	他們	老師_N	整理	結束	停	作弊	答案	臉	大樓_A	
5	森林	安靜	出爾反爾	照相_A	走	標準	奇怪_S	工作	開花	做	時間	老師_S	
6	油_S	剛剛	環境	他們兩個	好_A	寫	安排	抱	今天	學校	禁止	裙子	
7	綠	每	問題	夫妻_B	很	長大	還沒	明天	有	一定	叫	像	
8	回家_B	讀書	下雨	兒子	名字_S	加	快樂	平靜	真	鈴_A	客廳_A	眼鏡	
9	小孩	原來	幾時_B	忙	人_A	幸福	殘忍	倒	游泳_A	電影	結婚	過去的最近	
10	危險	拉	自己_N	腳踏車_	爸爸	不要_N	立刻	上癮_B	手語	我們	她	漂亮_S	
11	生氣_A	旅館_A	負責_B	少	未來_B	回答	希望	地方	雞	碰見	知道_S	變	
12	菜_N	好_B	愛	肯	想	去_B	難	不好_B	生病	怕_N	外面	妹妹	
13	棒_A	其中	紅_S	告訴	參加	兩個	更	見面_B	衣服	奶_S	動物	吃	
14	什麼_A	努力	結果	放	香煙	那	皮	困難	行動	比賽	身體_B	抽煙	
15	馬上	種類	見	會_N	早上_B	媽媽	獲得	經過	提供_B	繼續	學生	會議	
16	正確	日本	要_S	決定	現在	哭_A	是	作業	保護	報紙	玩	考試_S	
17	睡_A	他_A	完全	責任	出	有沒有	你	再_A	台北	快_S	健康	開車	
18	同學_A	舒服_A	去_A	嬰兒	代替	我_B	邀請	依然_A	失約_A	燈	這	棒球_B	
19	鐵	認真	張	西瓜	亮_A	收集	以後_B	貴_A	目的	吃飯_A	桌子	輸	
20	牙齒	第一名_	郵票	不能	講	炒	不知道_	眼睛	加入_B	送	近	來	
21	世界	學習	哪裡	整天_B	關心	還不錯	答應	看_N	有錢	一起_A	事情	豐富	

表 4-2：日常詞彙手語表（來源自行製作）

六、手語翻譯

為了在真實生活進行手語翻譯，本研究整合了手語詞彙辨識模型，自行設計了手語句子翻譯系統，並測試系統的整體準確度，實驗流程如下（圖 4-11）。

1.請受試者錄製句子影片 ，並將影片以每 20 幀切分輸入進手語詞彙辨識模型後，得到每個片段所代表的手語詞彙。

2.將步驟 1 輸出的所有手語詞彙，輸入進本研究設計的滑動窗口演算法，得出影片中所包含的所有中文詞彙。

3.將每個中文詞彙輸入進大型語言模型，重新排列成中文句子，與原句子比較分析，並計算準確度。

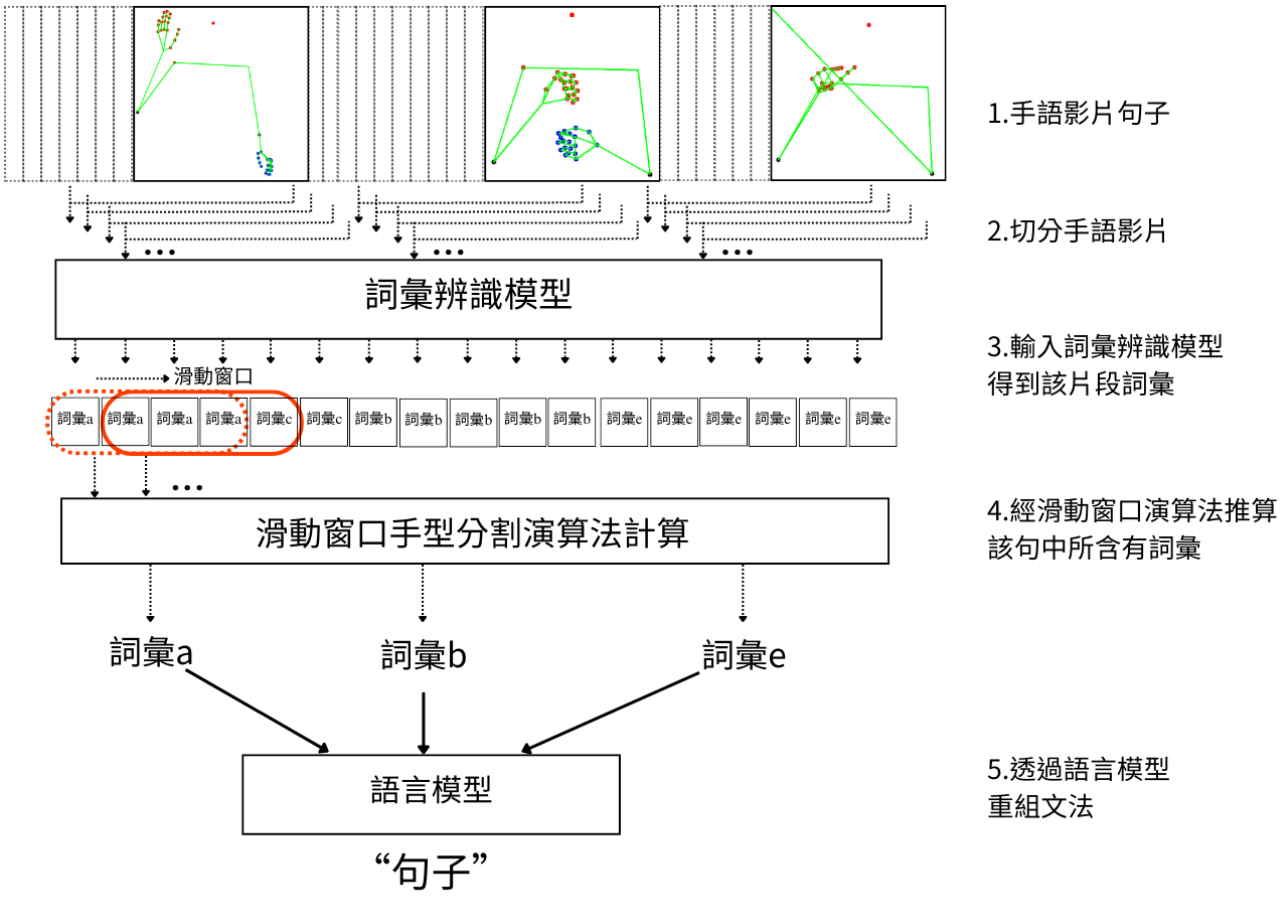

圖 4-11：手語翻譯流程圖（來源自行製作）

（一）滑動窗口

由於手語句子是由許多手語詞彙所組成的，本研究透過切分影片後，輸入進辨識模型得出可能詞彙，並且設計了滑動窗口演算法來判斷哪些詞彙存在於手語句子中，實現方法如下：

1.將每 20 幀的手語影片輸入詞彙辨識模型並記錄辨識結果，作為詞彙序列（如圖 4-12）。

2.設定 18 大小的滑動窗口，搜尋詞彙序列，若窗口內某詞彙的總數大於 75%就記錄該詞彙，若沒有詞彙的總數低大於 75%就記為-1（如圖 4-13）。

3.將每個窗口的標記整理成序列後，找尋連續相同的標記，合併成結果。（如圖 4-14）。

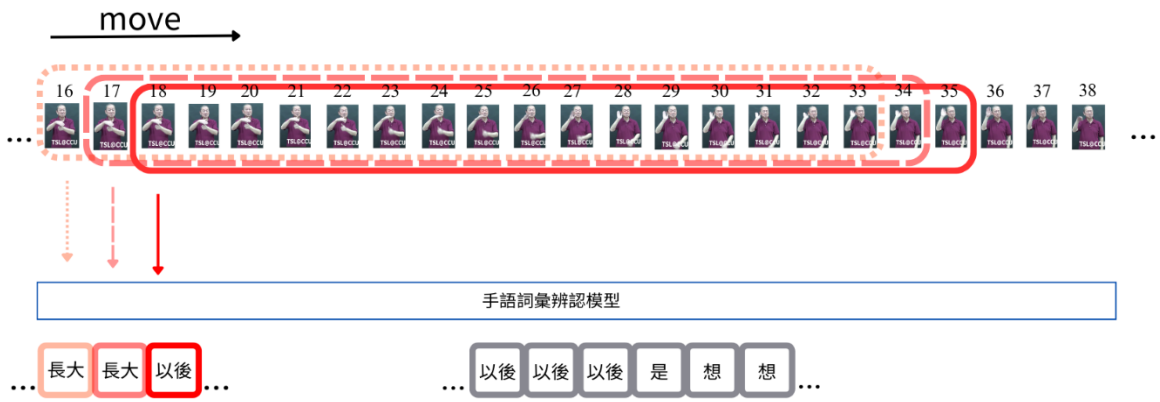

圖 4-12：影片切分辨識示意圖（來源自行製作）

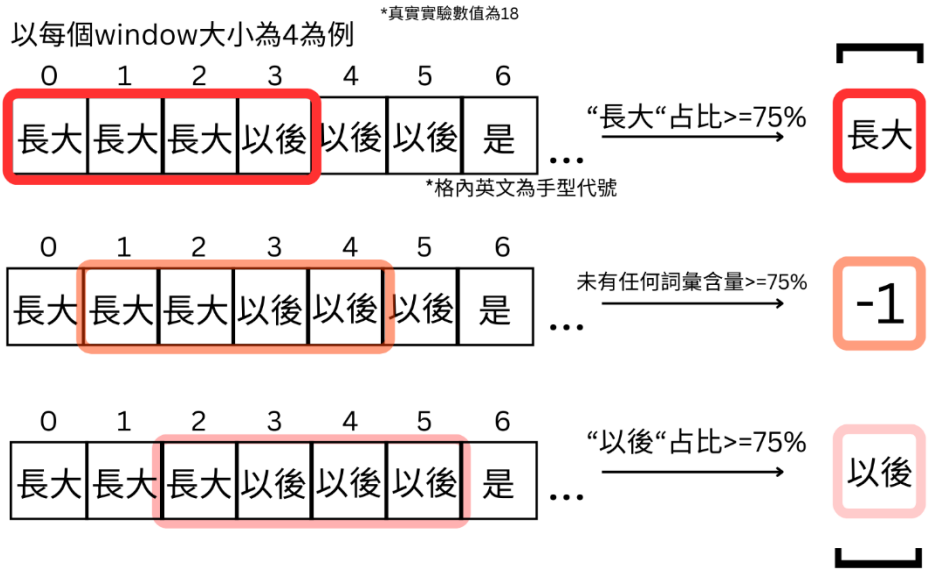

圖 4-13：滑動視窗示意圖（以每個 window 取 4 幀為範例）（來源自行製作）

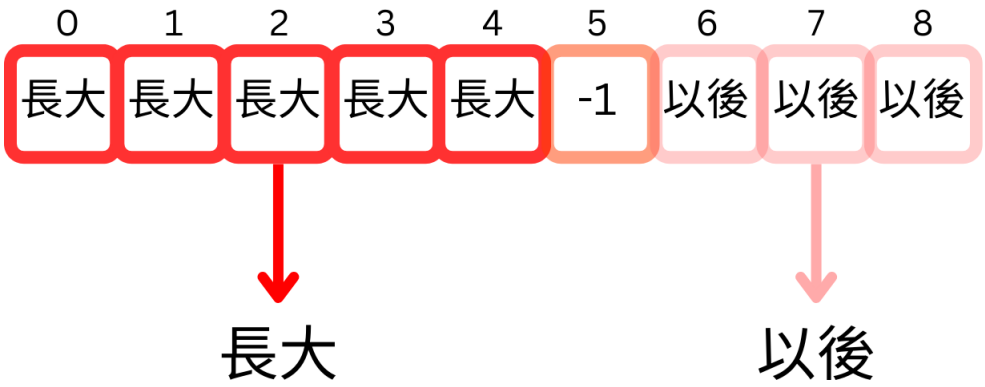

圖 4-14：合併相同標記示意圖（來源自行製作）

（一）大型語言模型翻譯

透過上述方式即可得出每個手語詞彙，最後將詞彙辨識結果輸入進大語言模型重組成句子，達成手語翻譯。

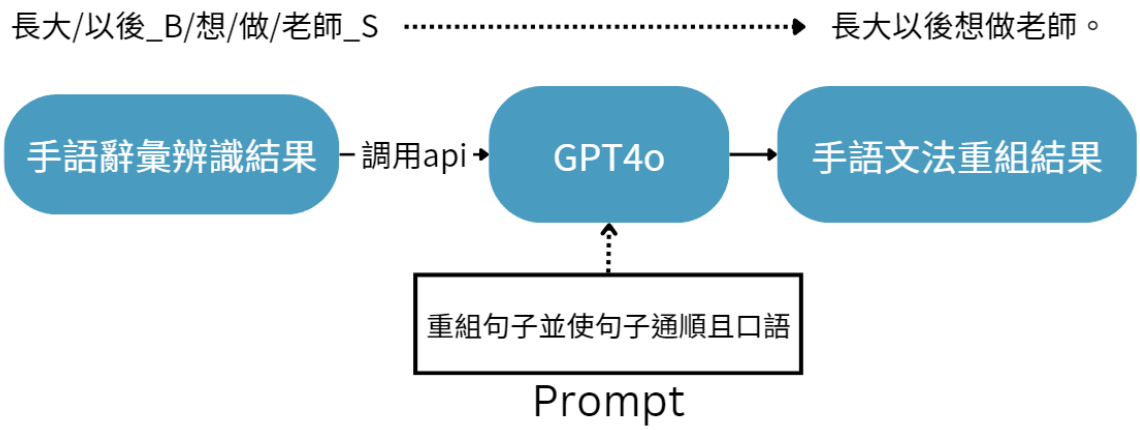

圖 4-15：大型語言模型調用流程圖（來源自行製作）

伍、研究結果與討論

一、預訓練結果

在預訓練中，本研究隨機遮蓋訓練資料中的座標輸入進模型，模型將預測遮蓋部分的座標。本研究使用 MPJPE loss function 衡量預測值與目標值的差異以更新模型參數，實驗訓練兩種不同輸入資料的模型，一個是包含骨架座標的模型 (Body Pretrained Encoder)；另一個是僅包含手部點座標的模型(Hand-shape Pretrained Encoder)。

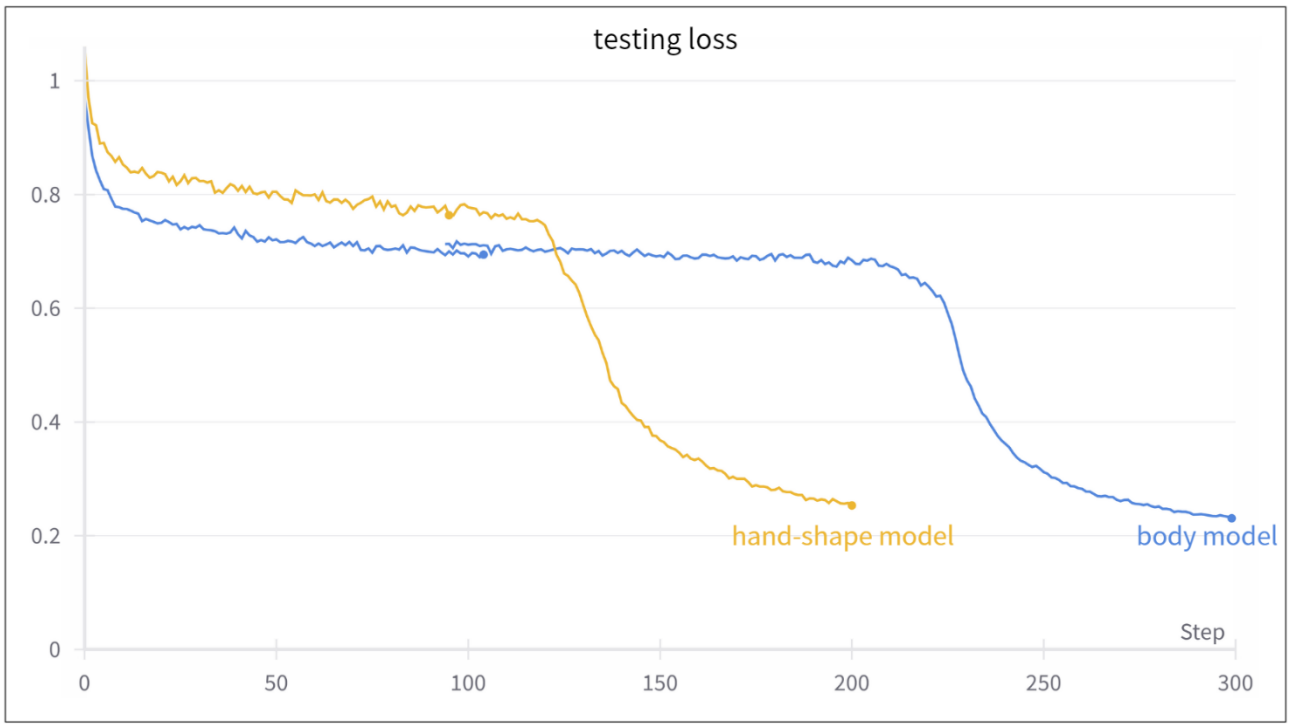

圖 5-1：不同 Batch Size，模型訓練的平均 testing loss（來源自行製作）

從圖 5-1 可以觀察在預訓練過程中，兩個模型分別在 step120 以及 step220 時，loss 值出現一次巨大的轉折，急遽下降，模型是在這期間學習到了手語的特徵。

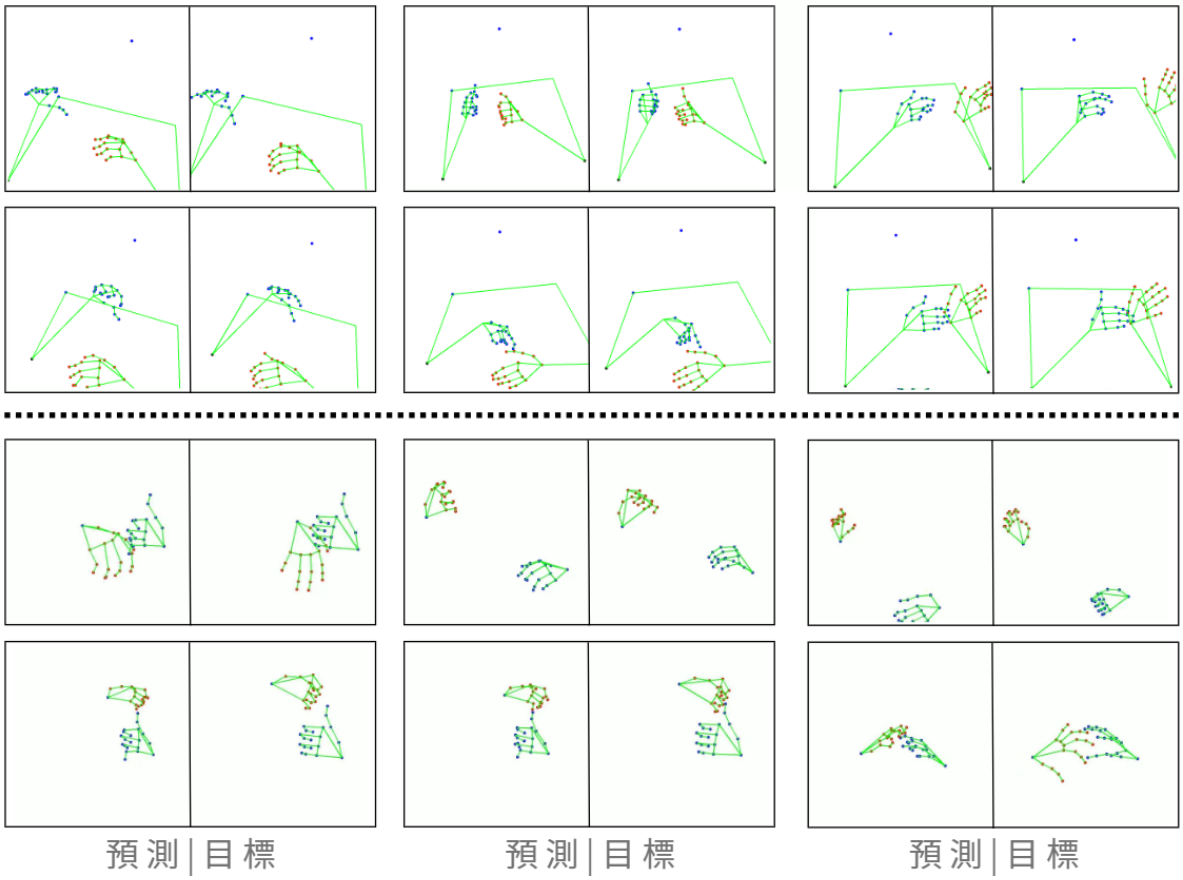

圖 5-2：模型預測的結果（左），目標值（右），上半部分為骨架預訓練模型，下半部分為手型預訓練模型。（來源自行製作）

圖 5-2 很好的展示出兩個模型對於遮蓋部分的預測與目標值相當接近，可以發現模型在雙手位置與關節點的部份預測得相當精準，而儘管在每個點預測上與目標值仍有些差異，不過仔細觀察，這是合理的誤差，模型仍然有預測出與目標值相同的手勢與姿態，足以說明模型學習到手語的特徵。

二、手語詞彙辨識實驗

在詞彙辨識實驗中，本研究 fine tune 了四種模型，在 242 個日常詞彙手語資料中進行訓練，並且透過測試集來評估模型的準確率。

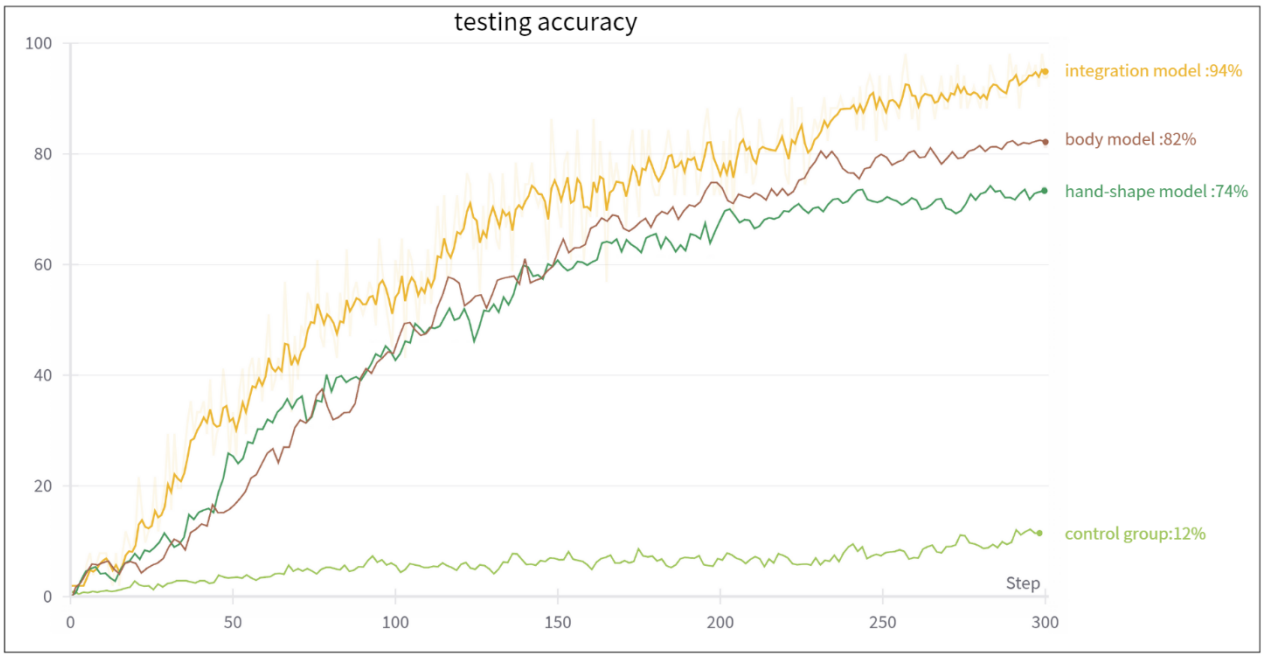

圖 5-3：四種詞彙辨識模型在訓練階段，測試集辨識準確率（來源自行製作）

模型/準確率	手型模型	骨架模型	融合模型	對照組
Testing accuracy	74.3%	82.6%	**94.8%**	12.3%

表 5-1：四種詞彙辨識模型最終在測試集辨識準確率（來源自行製作）

從表 5-1 可以看到，四種模型裡以融合模型的辨識準確率是最高的，相較於對照組提升了 82.5%的辨識準確率，也比手型、骨架模型的辨識準確率來的更高，足以說明融合模型可以將手型、骨架模型各自的優勢很好的加在一起，完成 94.85%的辨識準確率。因此，後續的手語句子翻譯系統，本研究都選用表現最好的融合模型。

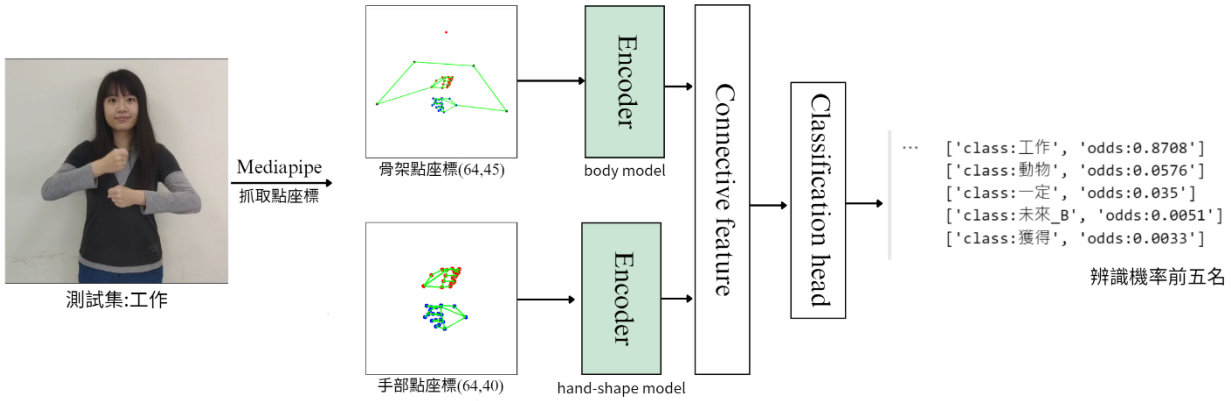

圖 5-4：詞彙辨識實驗圖（來源自行製作）

圖 5-4 受試者比出手語，經過 mediapipe 處理得到手部點座標，輸入進融合模型辨識，以"工作"為例。

三、滑動窗口演算法

本研究為了將每個手語詞彙從句子中辨識出來與去除雜訊，開發了滑動窗口句子分割演算法。實現方法為：每一幀往後取 20 幀輸入詞彙辨識模型，紀錄辨識結果，並將辨識結果由滑動窗口進行篩選，最後將連續詞彙進行合併，得到句子中所含詞彙。（如圖 5-5）。

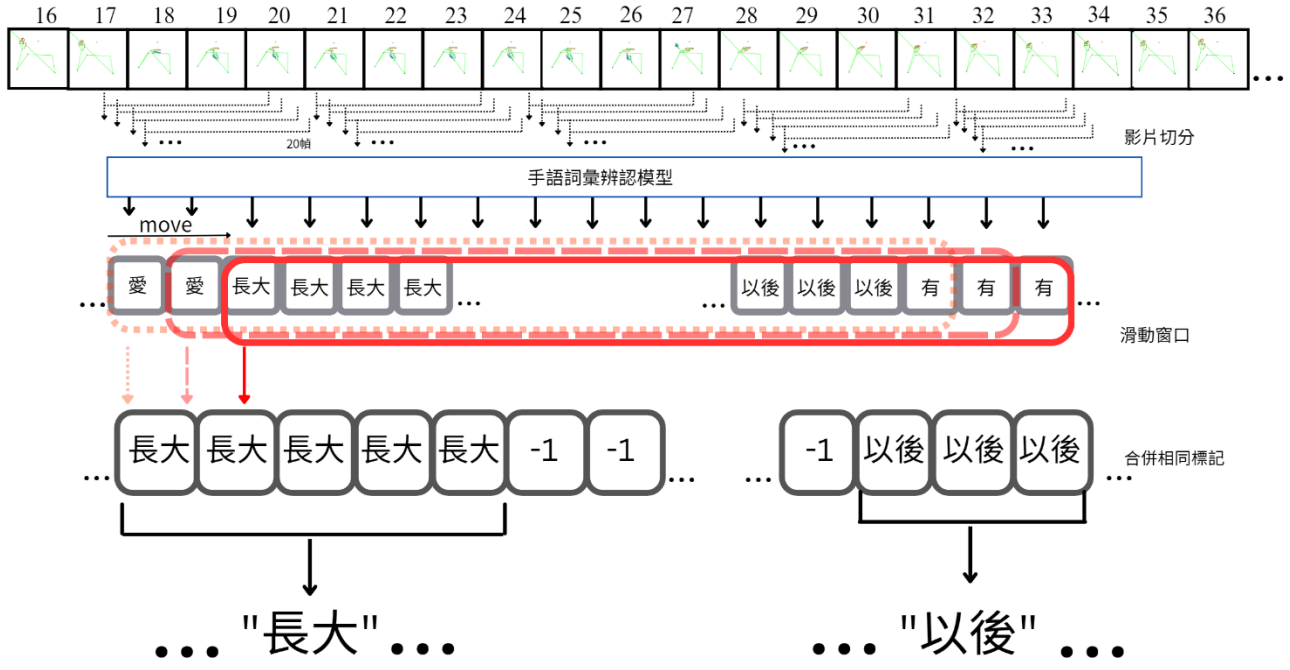

圖 5-5：滑動視窗手型分割演算法實驗圖（來源自行製作）

四、手語文法及大型語言模型實驗

臺灣自然手語的文法架構及順序與中文的文法大相逕庭，並且有大量省略詞彙與需要意會的動作。由於目前手語與中文的對照資料十分不足，資料量不足以支撐專門為台灣自然手語訓練一個語言模型，因此本研究必須另闢蹊徑。本研究的在測試後發現 LLM 能夠良好的解決此問題，並且相較於傳統語言翻譯模型 LLM 能夠根據現實世界的邏輯來還原省略詞彙與需要意會的部分，其中以 GPT-4 表現最佳，因此本實驗手語文法翻譯調用的是 GPT-4 的 API。

圖 5-6：GPT4-API 調用（來源自行製作）

五、手語翻譯實驗結果

為了在真實生活進行手語翻譯，本研究整合了手語詞彙辨識模型，自行設計了手語句子翻譯系統，並邀請受試者錄製手語句子影片來測試系統的整體準確度。

圖 5-7 展示了手語句子"我長大以後想當老師"的翻譯實驗圖，將影片切分為 300 個 20 幀的片段輸入進詞彙辨識模型得到對應的詞彙，並經滑動窗口演算法處理，隨後得到"我_B、長大、以後_B、想、工作、老師_S"。將此輸入大模型處理翻譯，最後模型翻譯結果為"我長大以後想做老師"。其結果顯現出，手語詞彙辨識模型辨識的相當準確，且大型語言模型是能夠理解手語中意會的部分，儘管與原句不是一字不漏地翻譯，其翻譯結果仍然與原句意思相同。

表 5-2 展示了手語句子翻譯的實驗最後的統計表，全部實驗了 100 個手語句子，並取得88%的翻譯準確率，其中詞彙辨識模型失誤率 9%，大型語言模型失誤率 6%。

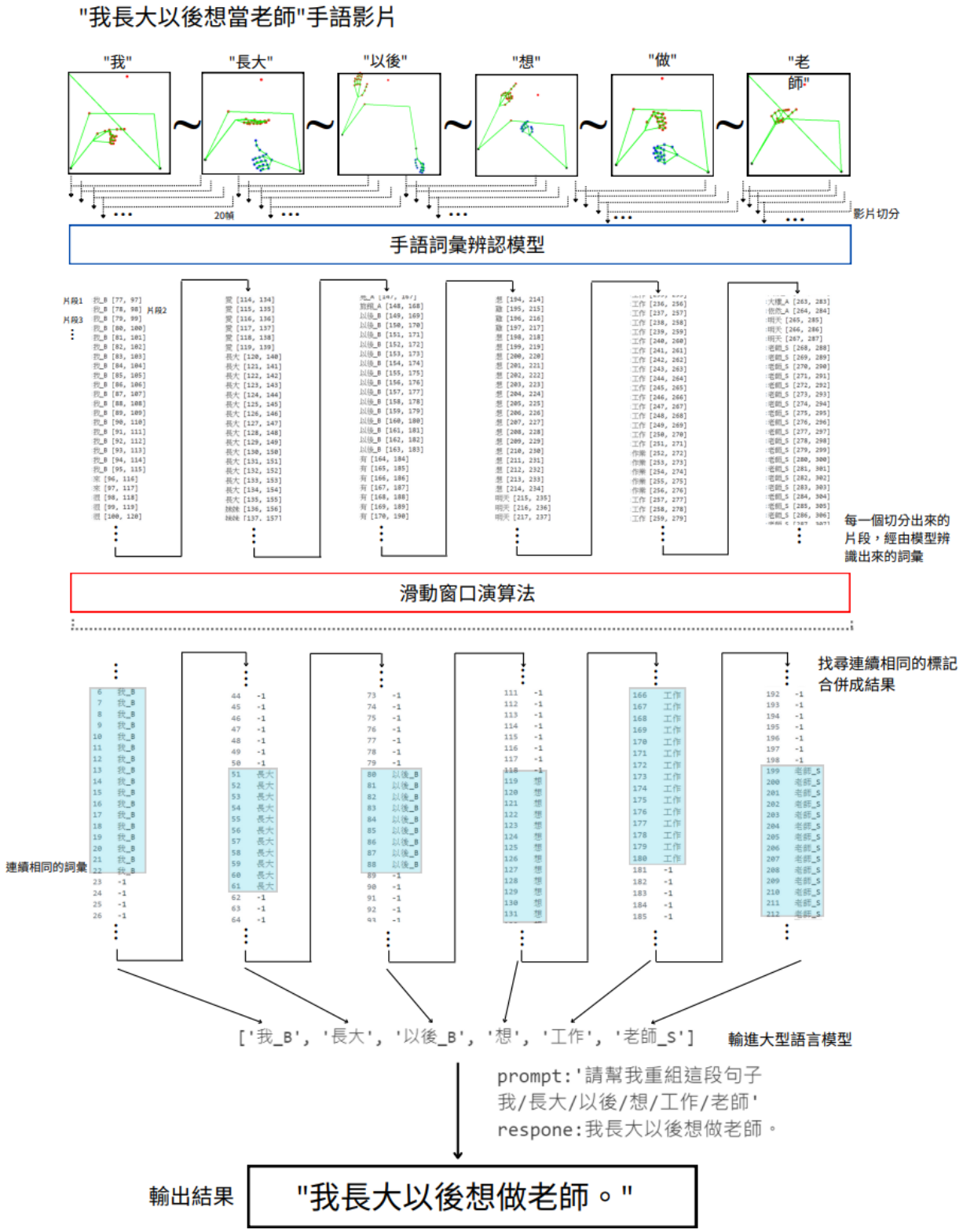

圖 5-7：手語翻譯實驗圖（來源自行製作）

原句	原詞	模型通行輸出	大型語言模型重組輸出
我有一件裙子	我_B/裙子/有/一	我_B/裙子/有/一	我有一條裙子。
這是標準答案	這/答案/標準	這/答案/標準	這是標準答案。
工作要認真	工作/認真/要_S	工作/認真/要_S	工作要認真。
下雨不能出門	下雨/出/不能	下雨/出/不能	下雨時不能出去。
他跑得很快	他_A/跑_B/直/快_S	他_A/跑_B/直/快_S	他跑得很快。
他們倆工作忙，很少見面	他倆/工作/忙/見面_B/少	他倆/工作/忙/見面_B/少	他們工作忙，所以見面少。
爸爸要我馬上回家	爸爸/告訴我/馬上/回家/馬上	爸爸/告訴我/馬上/回家/馬上	爸爸告訴我馬上回家
請自我介紹一下	(叫人)/我_B/自己/介紹	我_B/自己/介紹	我可以自己來介紹自己
這雞肉飯覓強壇可以	這/雞/吃飯_A/吃/壇不錯	這/雞/吃飯_A/吃/壇不錯	這雞吃起來壇不錯
每個地方都有好人和壞人	各處/人_A/好/人_A/不好_B/有	人_A/好/人_A/有	人很好。
今天是星期幾呢	今天/星期幾	今天/星期幾	今天是星期幾？
現在是幾點呢	現在/幾時_B	現在/幾時_B	現在幾點了？
紅燈停，綠燈行	紅_S/燈/停//綠/燈/走	紅_S/燈/停//我/燈/走	紅燈停，我走。
牙齒是很堅硬的	牙齒/硬_A	牙齒/硬_A	牙齒很硬。
睡覺前要刷牙	睡_A/前/刷牙/要_S	睡_A/來/刷牙/要_S	要睡覺了，來刷牙。
他願意和你結婚嗎	他_A/你/結婚/他_A/嗎	他_A/你/結婚/他_A/嗎	你覺得他願意和你結婚嗎？
寫作業會增加壓力	作業/寫/壓力/加/會_N	作業/寫/壓力/加/會_N	寫作業會增加壓力
爸爸很生氣	爸爸/生氣_A/很	爸爸/生氣_A/很	爸爸很生氣。
鐘響之後，老師就叫學生回來上課	鈴_A/完了/老師_N//叫/大家過來/拿書	鈴_A/完了/老師_N/一起/拿書	鈴聲響了，老師和我們一起讀書。
你做事一定要考慮清楚，不要反反覆覆	你/做/事/什麼//你/想/清楚/再/一定//出爾反爾_B/不要	你/做/事/什麼/你/想/清楚/再/一定/出爾反爾_B/不要	你做什麼事，一定要先想清楚，不要再出爾反覆
這個答案正確	這/答案/正確	這/答案/正確	這個答案是正確的，
我決定努力學習手語	手語/學/我_B/決定/努力	手語/學/我_B/決定/努力	我決定努力學手語。
你一定要幫我	你/幫/我/一定	你/幫/我/一定	你一定要幫我，
他努力得到第一名	他_A/認真/努力/獲得/第一名_N	他_A/認真/努力/獲得/第一名_N	他認真努力以獲得第一名。
環境保護，人人有責	環境/保護/每/人/責任	環境/保護/每/人/責任	保護環境是每個人的責任。
昨天的棒球比賽，結果我們輸了	昨天/棒球/比賽/結果/我們/輸	昨天/棒球/比賽/結果/我們/輸	我們昨天的棒球比賽輸了。
我比賽輸了	我_B/比賽/輸	我_B/比賽/輸	我輸了比賽。
你最好不要失約	你/失約_A/不要_S/記得	你/失約_A/不要_S/記得	別記得你失約的事情
他唸書的目的很明確	他_A/讀書/目的/清楚/很	他_A/讀書/目的/清楚/很	他讀書目的明確。
爸爸催促我立刻上床睡覺	爸爸/催/立刻/去/睡覺	爸爸/催/立刻/去/睡覺	爸爸趕快叫我去睡覺了。
也許明天會下雨	明天/下雨/也許/會_N	明天/下雨/也許/會_N	明天可能會下雨呢。
她可以跟我們一起去旅遊嗎	我們/一起/去/旅行/玩/她/加入/可以	我們/一起/去/旅行/玩/她/加入/可以	她可以加入我們一起去旅行玩。
原來他是老師	他_A/老師_S/原來	他_A/老師_S/原來	原來他是老師。
妹妹生病了	妹妹/生病	妹妹/生病	妹妹生病了。
昨天的考試很難	昨天/考試_S/難	昨天/考試_S/難	昨天的考試很難。
他今天替我代班	他 A/今天/幫我/代替/工作	他 A/今天/幫我/代替/工作	他今天幫我工作了。
我住在台北	我_B/家_B/台北	我_B/家_B/台北	我家在台北。
他們兩個人是我的朋友	他們兩個/我_B/朋友	他們兩個/我_B/朋友	他們兩個是我朋友。
希望世界和平	希望/世界/和平	希望/世界/和平	希望世界能夠和平。
他近來的行動怪異	他_A/過去的最近/行動/奇怪_S	他_A/過去的最近/行動/奇怪_S	他最近做的事情真的很奇怪。
自己去森林很危險	自己_N/去_B/森林/危險/很	自己_N/去_B/森林/危險/很	自己去森林很危險。
這麼繁重的工作，真叫我害怕	這/工作/這些/我_B/怕_N/很	這/工作/這些/我_B/怕_N/很	我對這些工作感到很害怕。
爸爸完全不知道這件事	這/事情/爸爸/完全/不知道_N	這/事情/爸爸/完全/不知道_N	爸爸完全不知道這件事情。
我剛剛已經回答了你的問題	你/問題/我_B/剛剛/回答你/完了	你/問題/我_B/剛剛/回答你/完了	我剛剛已經回答了你的問題。
明天我安排了一場會議	明天/會議/我_B/安排/完了	明天/會議/我_B/安排/完了	明天的會議我已經安排好了。
媽媽昨天整理好了客廳	媽媽/昨天/整理/客廳/地方/整理/結束	媽媽/昨天/整理/客廳/地方/整理/結束	媽媽昨天已經把客廳整理完了。
我每天騎腳踏車上學	我_B/每天/騎腳踏車_A/去/讀書	我_B/每天/騎腳踏車_A/去/讀書	我每天騎腳踏車去讀書。
他跟我是同學	我倆/同學_A/是	我倆/同學_A/是	我倆是同學。
衣服在桌子上	桌子/衣服/放	桌子/危險/放	桌子放得很不穩。
你有收集郵票嗎	郵票/張/郵票/你/收集/有沒有	郵票/張/郵票/你/收集/有沒有	你有沒有收集郵票呢？
工作還沒完成，繼續把它做完	工作/完/還沒/再_A/繼續/工作/完	工作/完/再_A/繼續/工作/完	工作完成後，再繼續工作。
早安，你好	早/安靜/好_A	早/安靜/好_A	早安，這裡真的很靜
我長大以後想當老師	長大/以後_B/想/做/老師_S	長大/以後_B/想/做/老師_S	
如果你不出席會議，記得要跟我說	如果/你/會議/參加/不要_S//你/記得/告訴我	如果/你/會議/參加/不要_S/你/記得/告訴我	如果你不參加會議，記得告訴我。
他的名字看起來很不錯	他_A/名字_S/看_N/棒_A	他_A/名字_S/看_N/棒_A	他看起來名字很棒。
看電影時請保持安靜	電影/看/請/安靜/依然_A	電影/看/請/安靜/依然	安靜看電影，依然不請，
他很有錢，到處亂花	他_A/富裕/有錢/花錢(到處)	他_A/富裕/有錢/花錢(到處)	他錢多，所以花得很大方
兒子長得很像爸爸	兒子/他_A/爸爸/他倆/臉/像	兒子/去/爸爸/他倆/臉/像	兒子像他爸爸，他倆的臉很像。
用動物皮製作產品很殘忍	動物/皮/做/那/哭_A/殘忍	動物/皮/做/那/哭_A/殘忍	那種以動物皮為原料的行為實在太殘忍，讓人無法忍住淚水
我本來是老師	我_B/原來/老師_S	我_B/原來/老師_S	我原來是老師呢
今天要考英文	今天/英文/考試S/要_S	今天/英文/考試S/要_S	今天我們要考英文了
我們一起走吧	我們兩個/來/走	我們兩個/來/走	我們兩個一起走吧！
夏天的時候我們會吃西瓜	夏天_B/夏天_A/西瓜/吃	夏天_B/夏天_A/西瓜/吃	夏天就該吃西瓜啊！
他的眼睛亮了起來	他_A/眼睛/這/亮	他_A/眼睛/這/亮	他的眼睛真的好亮啊
考試不可作弊	考試_S/作弊/不能	考試_S/作弊/不能	考試時不能作弊
遇到困難時，我可以幫你	碰見/困難/有//我_B/幫你/可以	碰見/困難/有/我_B/幫你/可以	如果你碰見困難，我可以幫你，
我答應他的邀請	他_A/邀請我/我_B/答應	他_A/邀請我/我_B/答應	他邀請我，我答應了，
努力學習是為了有更好的將來	學習/努力/目的/未來_B/更/棒_A	學習/努力/目的/更/棒_A	努力學習是為了更棒的目標。
每次出門都要帶錢包	每/出/錢/皮包/放進/要_S	每/出/錢/皮包/放進/要_S	每次出門，要把錢放進皮包。
每天游泳對身體很好	每天/游泳_A/幫我/身體/好_A	每天/游泳_A/幫我/身體/好_A	每天游泳對身體有好處。
我可以做完	做/結束/我_B/可以	做/結束/我_B/可以	我可以做到結束。
金是一種貴重金屬	鐵/金/種類/其中/一/金/貴	鐵/金/種類/其中/一/金/貴	金是其中一種貴金屬的種類。
我很快樂	我_B/快樂	我_B/快樂	我很快樂。
希望明天是晴天	希望/明天/晴天_A	希望/明天/晴天_A	希望明天是晴天
我是學生	我_B/學生	我_B/學生	我是學生。
他瞎了我一眼	他_A/見我	他_A/見我	他見到我了。
他說我很漂亮	他_A/告訴/我_B/漂亮_S	他_A/告訴/我_B/漂亮_S	他告訴我我很漂亮。
他說會永遠愛我	他_A/告訴/我/愛/我_B/整天_B	他_A/告訴/我/愛/我_B/整天_B	他一整天都在對我說愛我。
我倆是好朋友，從小一起長大	我倆/朋友/長大	我倆/朋友/長大	我倆是從小一起長大的朋友。
我們一起照相	我倆/一起/照相	我倆/一起/照相	我們一起來拍照吧
有好多小孩在跑來跑去	小孩/他們/跑來跑去	小孩/他們/跑來跑去	他們的小孩在到處跑來跑去
他對菸酒上癮	他_A/這/香煙/酒/上癮_B	他_A/香煙/酒/上癮_B	他對香煙和酒上癮。
校園禁止抽煙	學校/這/抽煙/禁止	學校/這/抽煙/禁止	這裡的學校是不允許抽煙的喔
媽媽抱抱小孩	小孩/媽媽/抱	小孩/媽媽/抱	媽媽抱著小孩
日本的櫻花全都開花了	日本/櫻/(到處)/開花	日本/櫻/開花	日本的櫻化就像花開一樣
他真的是我朋友 (他是一個真誠的朋友)	他_A/我_B/朋友/真/是	他_A/我_B/朋友/真/是	他真的是我朋友
我的朋友在日本	我_B/朋友/那/日本/那	我_B/朋友/那/日本/那	那是我在日本的一位朋友
他送我一張卡片	卡片/他_A/送我	卡片/他_A/送我	他送我卡片。
你知道我的眼鏡在哪裡嗎	我_B/眼鏡/哪裡/你/知道_S	我_B/眼鏡/哪裡/你/知道_S	你知不知道我眼鏡放在哪裡？
嬰兒要喝奶	嬰兒/奶_S/要_S	嬰兒/奶_S/要_S	他拉了我
出門記得要帶錢	嗨/出/記得/錢/放口袋	嗨/出/記得/錢/放口袋	嗨，出門記得把錢放口袋
互相關心是婚姻幸福的關鍵	互相/關心/這兩個/夫妻/幸福/重要/這	互相/關心/這兩個/夫妻/幸福/重要/這	這對夫妻互相關心，這對他們的幸福很重要
肥胖對身體不好	胖_A/對/身體_B/健康/不好_B	胖_A/身體_B/健康/不好_B	身體太胖對健康不好。
炒菜時要用油	菜_N/炒/油_S/倒/要_S	菜_N/炒/油_S/倒/要_S	炒菜的時候，要先倒油哦
他把我拉過去	他_A/拉/我	他_A/拉/我	我們兩個的地方很近。
外面下雨了	外面/下雨	外面/下雨	外面正在下雨。
我們兩個離得很近	我們兩個/地方/近	我們兩個/地方/近	我們兩個的地方很近。
旅館提供舒服的床	舒服_A/床/舒服_A/那/旅館/那/負責_B/提供_B	舒服_A/床/種類/那/旅館/那/負責_B/提供_B	那家旅館負責提供各種類型的舒服床。
我要看報紙	報紙/我_B/欣賞/要_S	報紙/我_B/欣賞/要_S	我想看看報紙。
他開車經過了許多高樓	大樓_A/他_A/開車/經過	大樓_A/他_A/開車/經過	他開車時經過了那棟大樓。

表 5-2：手語翻譯實驗成果圖（紅色底格子為錯誤輸出，正確率計算以最右欄計算）

六、問題與討論

本研究在手語翻譯上達到了 88%的準確率，本研究者詳細調查剩下 12%的錯誤，並歸納出以下幾點：

（一） Mediapipe 的不準確

在本研究的手語詞彙辨識模型實驗中，大部分的辨認錯誤都來自於 Mediapipe 的辨認失誤，而導致模型輸入資料完全錯誤。可惜的是，現在市面上仍無比 Mediapipe 更準確的手部點座標辨認模型。

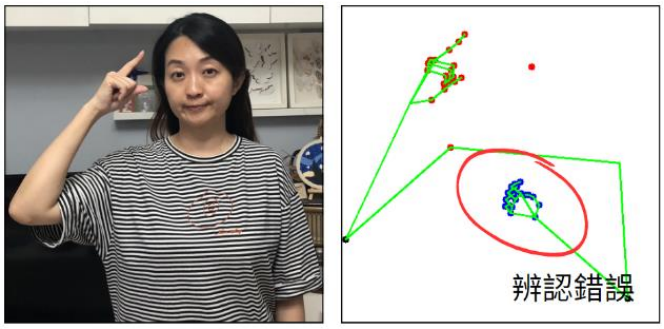

圖 5-8：Mediapipe 辨認錯誤（來源自行製作）

（二） LLM 翻譯問題

雖然 LLM 可以很好的協助我們重組中文句子，但是還是有其翻譯失誤的時候。 比如在"你最好不要失約"中，其拆解單字為"你/失約_A/不要_S/記得"，但輸入進 LLM 輸出的結果卻是"別記得你失約的事情"。

然而， 本研究仍採用 LLM 最主要原因來自於他的推理能力。像是"肥胖對身體不好 "中，原拆解文字為"胖_A/對/身體_B/健康/不好_B"，模型辨認錯誤導致輸出變成"胖_A/身體_B/健康/不好_B"，但是 LLM 有完整理解語意，翻譯為"身體太胖對健康不好 "。

本研究者推斷原因為在中文及手語的文法架構上的出入，導致 LLM 會錯意。日後期待更進提示詞，讓 LLM 的表現更精準。

原句	原詞	模型運行輸出	大型語言模型重組輸出
你最好不要失約	你/失約_A/不要_S/記得	你/失約_A/不要_S/記得	別記得你失約的事情
肥胖對身體不好	胖 A/對/身體 B/健康/不好 B	胖 A/身體 B/健康/不好 B	身體太胖對健康不好。

圖 5-9：LLM 翻譯問題，（來源自行製作）

陸、結論

本研究貢獻在於，第一次將自監督學習應用在台灣手語辨識，擺脫了過去的研究需要大量標記樣本進行辨識的困境。成為台灣第一個手語詞彙量突破百位數，達到了 242 個可辨識詞彙以及 94.8%的辨識準確率，而且本研究之作法僅需 5 個標記樣本即可訓練模型辨認詞彙。

本研究證明遮蔽一定資訊並使模型預測遮蔽內容的作法可適用於手語辨識任務。實驗結果顯示，結合手型模型與骨架模型的融合模型表現最佳，比沒有採用預訓練模型的對照組高出 82.5%的辨識準確率。

本研究也基於此自監督的預訓練模型，開發了首個可實際應用的手語翻譯的系統，在手語句子翻譯的表現達到優秀的 88%的準確率，證明了本研究的手語翻譯系統可真實應用在日常使用上。本研究者期待此技術在更妥善的完善後，加入人機介面，可以投入實際應用的場合，幫助聾人與聽人的交流、溝通，增進弱勢族群的福祉，同時也可為手語教育帶來貢獻，增進社會的共榮和諧。

柒、參考文獻資料

[1] Vaswani, A.（2017, June 12). *Attention Is All You Need*. https：//arxiv.org/pdf/1706.03762.pdf

[2] Devlin, J.（2018, October 11). *BERT： Pre-Training of Deep Bidirectional Transformers for Language Understanding*. https：//arxiv.org/pdf/1810.04805.pdf

[3] Alexey, D.（2020, October 22). *An Image Is Worth 16x16 Words： Transformers for Image Recognition at Scale*. https：//arxiv.org/pdf/2010.11929.pdf

[4] Kaiming, H.（2021, November 11). *Masked Autoencoders Are Scalable Vision Learners*. https：//arxiv.org/pdf/2111.06377.pdf

[5] Hvilshøj, F.（2023, March 3). *What Is One-Shot Learning in Computer Vision*. https：//encord.com/blog/one-shot-learning-guide/

[6] Huang, C. chiu.（2020, December 8). *論文閱讀筆記 — 3D 人體姿態辨識 Camera Distance-Aware Top-down Approach for 3D Multi-Person Pose Estimation from a Single RGB Image*. https：//williamchiu0127.medium.com/%E8%AB%96%E6%96%87%E9%96%B1%E8%AE%80%E7%AD%86%E8%A8%98-

3d%E4%BA%BA%E9%AB%94%E5%A7%BF%E6%85%8B%E8%BE%A8%E8%AD%98-camera-distance-aware-top-down-approach-for-3d-multi-person-pose-estimation-from-3d89a33eeb33

[7] Li, M.（2021, October 29). *Transformer 论文逐段精读*. https：//youtu.be/nzqlFIcCSWQ?si=5bXdhqd8Q3S__zff

[8] Mu, L.（2021, December 10). *MAE 论文逐段精读【论文精读】*. https：//youtu.be/mYlX2dpdHHM?si=JzMmuL3Y1bj6-15L

[9] Mu, L.（2021, November 30). *ViT 论文逐段精读【论文精读】*. https：//youtu.be/FRFt3x0bO94?si=8Xe34URtNDwvd5H9\

[10] Lee, H.（2019, June 1). *Transformer*. https：//youtu.be/ugWDIIOHtPA?si=udow_2gw22RXRB5a

[11] 中正大學.（n.d.). 臺灣手語線上辭典. https：//twtsl.ccu.edu.tw/TSL/index.php

[12] 教育部.（n.d.). 常用手語辭典. https：//special.moe.gov.tw/signlanguage

[13] 劉秀丹、曾進興（2007）。文法手語構詞語句法特性對聾生詞義與句義理解的影響。特殊教育研究學刊。 http://bse.spe.ntnu.edu.tw/upload/journal/prog/6O6_21SL_209R_35CM518.pdf