基於語音與文本多模態特徵之條件式潛在擴散模型：應用於阿茲海默症PET影像生成
Multi-modal Latent Diffusion Models Conditioned on Speech and Text Features for Alzheimer’s Diease PET Synthesis

學生：黃禹璇　　指導教授：蔡佩璇

在當代阿茲海默症 (Alzheimer's Disease, AD) 的早期診斷中，結合多模態異質數據的人工智慧輔助系統已成為精準醫療的重要趨勢。雖然正子斷層造影 (PET) 能提供大腦代謝異常與病理蛋白沉積的關鍵分子層級資訊，但其高昂的成本與設備門檻導致數據極度匱乏。相較之下，語音訊號具備非侵入性與易取得性，且能有效反映患者的神經認知狀態。為解決高價值醫療影像短缺的痛點，本研究提出一項創新的跨模態生成框架，利用患者在圖片描述任務（Cookie Theft Task）中的語音與文本數據，結合人口統計變數（年齡與性別），合成其對應的 2D 腦部 PET 影像。

首先，在特徵萃取階段，分類器將輸入資料拆分為三個獨立路徑：語音模態透過 wav2vec 2.0 提取聲學特徵，並搭配區塊平均降採樣以解決與文本長度落差過大的問題；文本模態利用 Whisper 轉錄後，經 BERT 模型萃取深層語意與上下文特徵；圖文關聯模態（Graph）則藉由視覺語言模型（如 BLIP-2）建立二分圖（Bipartite Graph）來計算圖文相似度，並透過 3 層圖卷積神經網路（GCN）提取結構化表徵。綜合上述模態與臨床量表，模型進一步萃取出具備高疾病鑑別度（AD vs. Cognitively Normal）的 128 維潛在空間特徵。隨後，針對 ADNI 資料庫之 3D PET 影像進行空間標準化，並於特定 Z 軸高度（如側腦室與海馬迴區域）進行 2D 切片前處理。在影像生成階段，本研究以潛在擴散模型（Latent Diffusion Model, LDM）為骨幹，於潛在空間中執行逐步去噪；為確保生成的器官大腦結構合理且一致，模型導入了專門處理「結構二值遮罩（Structural Binary Mask）」的輔助去噪網路。實作上，除了利用 Canny 邊緣檢測器從原圖擷取輪廓遮罩外，該輔助網路更複製了主 LDM 架構中編碼器與中間區塊的參數作為處理條件的基礎，並在將特徵注入主網路解碼器前，加入了初始化為零的卷積層（Zero convolution）。