基於潛在擴散模型之注意力導向腦波音樂生成研究

Attention-Guided EEG Music Generation via Latent Diffusion Models

指導教授：梁勝富 教授　**作者：**詹博堯

專題動機

近年來 Brain-Computer Interface (BCI) 與生成式 AI 快速發展，使得利用腦波訊號（EEG）直接生成音樂成為可能。然而，由於 EEG 訊號具有**低訊雜比（Low SNR）**、**高個體差異**以及**時間解析度限制**等問題，使得目前 EEG-to-Music 系統仍難以生成高品質且具語意一致性的音樂內容。

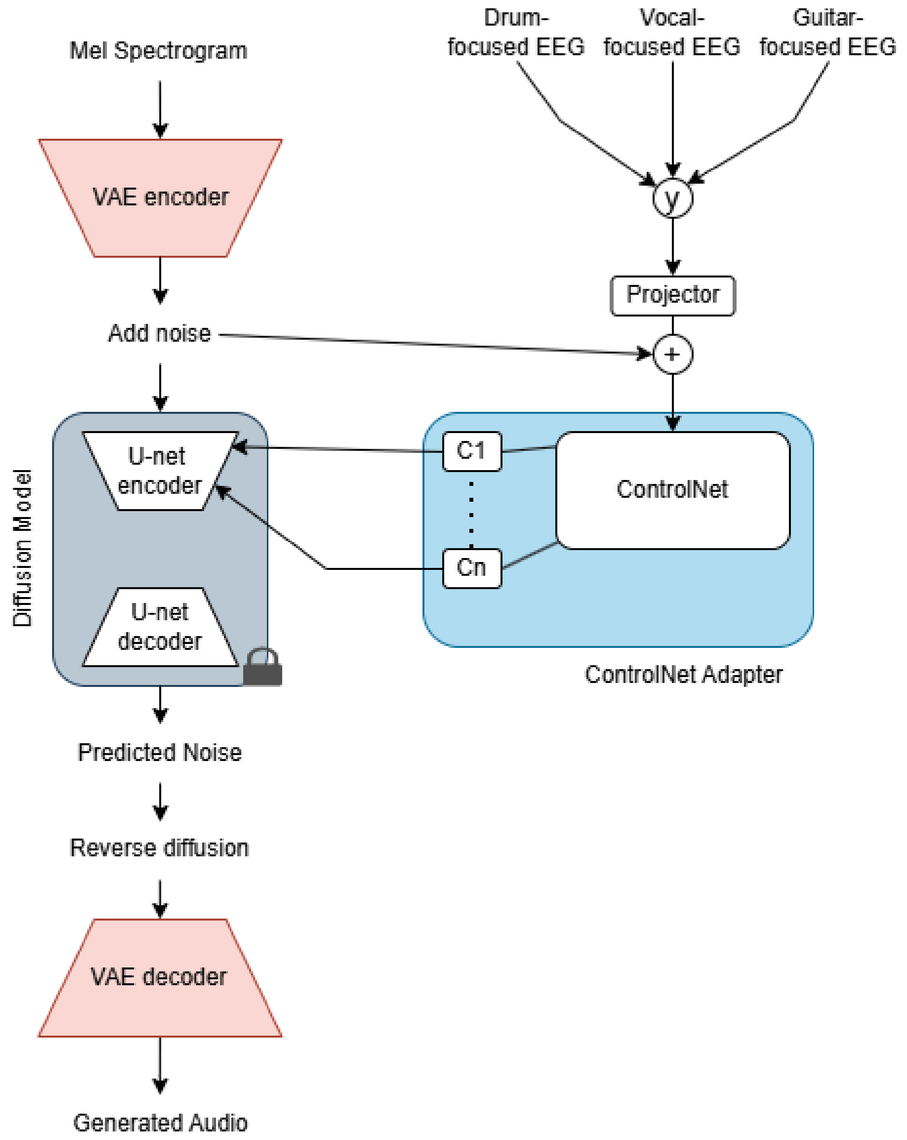
神經科學研究指出，即使在相同音樂刺激下，當使用者將**注意力**集中於不同樂器聲部（例如 **Drum、Guitar、Vocal**）時，大腦仍會產生**可區分的神經表徵**。因此本研究提出**多重注意力整合（Multi-Attention Integration）**架構，希望利用不同注意力狀態下所產生的 EEG 表徵，提升生成模型所能獲取的音樂資訊量，進而**改善生成音樂品質**。

模型架構

本研究以**潛在擴散模型（Latent Diffusion Model）**為基礎，結合 **ControlNet** 建立 **EEG 條件化音樂生成**架構。過去部分 EEG-to-Music 研究採用 Encoder-Decoder 架構，直接將 EEG 映射至音訊或音訊特徵。然而，音樂本身具有極高維度與複雜結構，包含節奏、旋律、和聲、音色及多樂器交互關係，因此若完全依賴 Encoder-Decoder 從頭學習 EEG 與音樂之間的映射關係，通常需要極為龐大的資料量才能獲得穩定的生成效果。相較之下，腦波資料取得成本高昂，受試者數量與錄製時間皆受到限制，使得直接訓練大型生成模型在實務上相當困難。

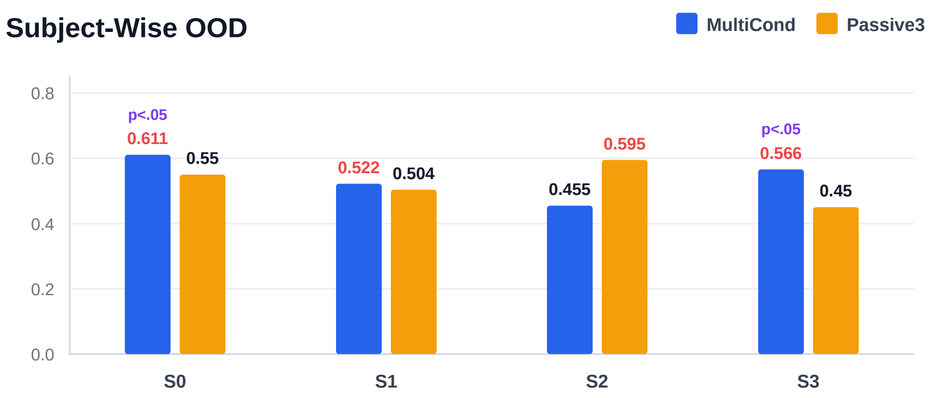
因此，本研究採用預先在大規模音樂資料上訓練完成的 **AudioLDM2** 作為基礎生成模型，利用其已學習到的音樂結構、音色分布與聲部關係，避免從零開始學習音樂生成能力。EEG 不再負責直接生成音樂，而是透過 ControlNet 作為條件訊號，引導擴散模型朝向與受試者神經活動相符的音樂方向生成。此方法將問題由「EEG 直接生成音樂」轉換為「**EEG 引導預訓練音樂生成模型**」，大幅降低資料需求並提升模型訓練穩定性。

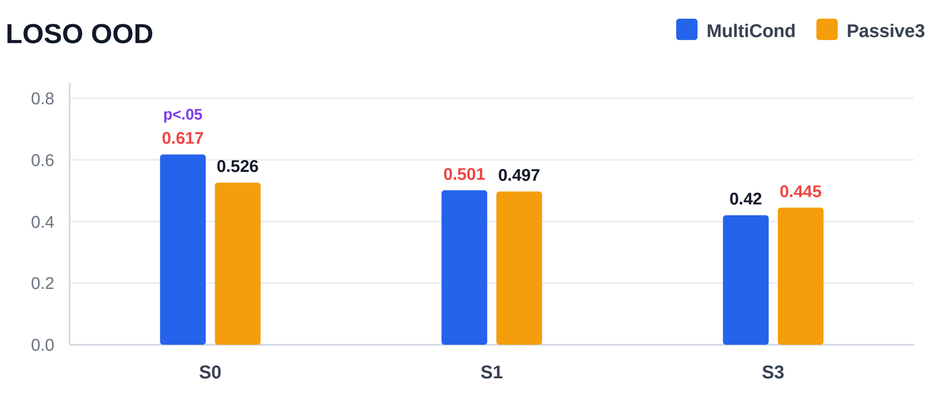
接著，受試者分別專注於 Drum、Guitar 與 Vocal 聲部進行聆聽，並記錄對應 EEG 訊號。三組注意力導向的神經表徵會於通道維度進行整合，形成**多重注意力條件（Multi-Condition EEG）**，再透過 **Projector** 映射至與音訊潛在空間相同的表示形式，並利用 ControlNet 注入至 AudioLDM2 的 U-Net 中進行條件控制。最終模型從隨機高斯雜訊開始，在 EEG 條件引導下逐步進行反向擴散與去噪，生成與受試者腦波表徵相對應的自然音樂內容



圖一 、模型架構

成果指標 (CLAP on OOD Track)





Stem-Level Analysis (Subject 0)

Generated from	Target stem: Drum Stem	Target stem: Guitar Stem	Target stem: Vocal Stem
Drum Attention	p<.05 0.797	0.574	0.458
Guitar Attention	0.748	0.566	0.468
Vocal Attention	0.677	0.59	0.469
Passive	0.609	0.527	0.466

Red numbers mark each target-stem column maximum; p labels appear only for p < .05.

結論與未來展望

本研究結果顯示，多數受試者的 **MultiCond 表現優於 Passive3**，Train-All 實驗亦驗證多重注意力整合能**提升 EEG 音樂生成的語意一致性**。Stem-Level Analysis 顯示，相較於 Passive 條件，注意力導向 EEG 普遍能提升生成音樂與目標 Stem 的相似度。其中 **Drum Stem** 呈現最**明顯的 Attention Effect**，顯示模型已能部分利用受試者的注意力資訊進行音樂生成。LOSO 實驗結果顯示，MultiCond 在跨受試者情境下仍具有一定優勢，其平均 CLAP 分數（0.513）高於 Passive3（0.489）。其中 Subject 0 呈現最明顯的提升效果，而 Subject 1 兩者表現接近；Subject 3 則出現 Passive3 略優於 MultiCond 的情況。整體而言，多重注意力整合有助於**提升模型泛化能力**，但跨受試者表現仍受到個體差異影響。值得注意的是，Subject 0 具有專業音樂製作背景，在各實驗中表現較符合預期；相反地，Subject 2 表示難以持續專注於指定樂器，容易受到其他聲部干擾，因此未納入 Train-All 與 LOSO 實驗。我們推測音樂能力以及注意力穩定度會直接影響 EEG 中注意力相關表徵的品質，進而影響模型學習效果。

本研究屬於 **Pilot Study**，目前僅包含三位受試者，因此 LOSO 結果容易受到個體差異影響，尚不足以穩定評估模型的**跨受試者泛化**能力。未來將持續擴增受試者數量與資料規模，以建立更具代表性的訓練與測試資料集，進一步提升模型的泛化能力與結果穩定性。跨受試者泛化能力一直是 **BCI 領域的重要挑戰**。本研究的潛在價值不僅在於 EEG 音樂生成，更在於探索具有較**高泛化能力的注意力導向 EEG 表徵**。未來若能驗證其跨受試者穩定性，將有機會成為更**可靠且泛用的 BCI 控制方式**，並應用於各類人機互動系統。