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

Bryan Chan

研究動機

有時候想到一個旋律覺得會紅但是不知道怎麼做出來？

- 傳統音樂創作需要演奏技巧或專業工具，創作門檻較高
- EEG 提供全新的人機互動方式，讓使用者能直接透過腦波參與創作
- 有望協助肢體障礙者進行音樂創作與數位內容表達
- BCI 長期面臨跨受試者泛化能力不足的挑戰
- 期望建立更具泛化能力的 BCI 控制訊號

研究難題

- ## EEG 低訊雜比
  Low SNR

  腦波訊號本身微弱，易被雜訊淹沒，難以穩定萃取與音樂相關的有效資訊。

- ## 個體差異大
  Subject Variability

  不同受試者的腦波表徵差異顯著，使模型難以跨受試者泛化。

- ## 音樂結構複雜
  Complex Structure

  音樂本身具有複雜結構，包含節奏、旋律、和聲、音色及多樂器交互關係。

研究背景與想法

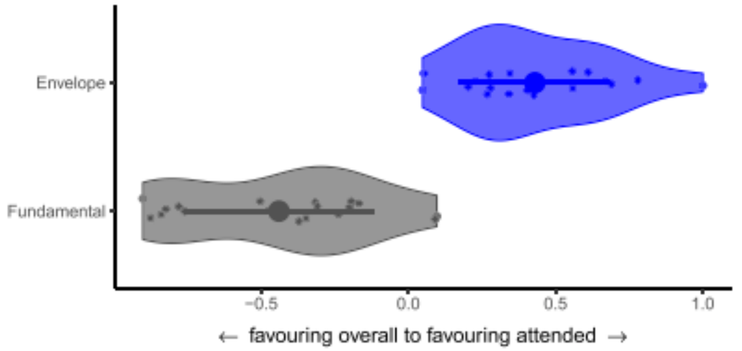
觀察：

Attention Creates Distinct EEG Representations

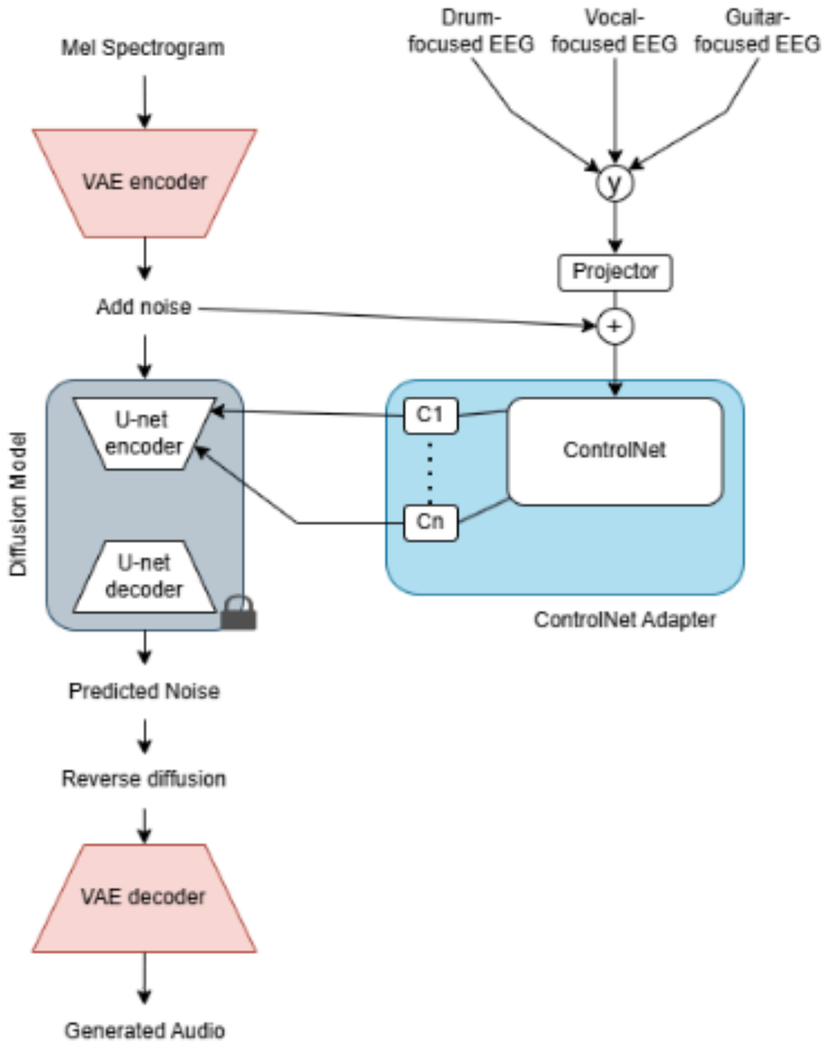
- 聆聽相同音樂時，注意力焦點不同，EEG 表徵也會不同
- 專注 Drum / Guitar / Vocal 可能提供不同音樂相關資訊

核心想法：

- 整合不同注意力狀態下的 EEG 表徵，提供更豐富的音樂相關資訊，進而提升音樂生成品質。
- 同時也探索BCI 泛化性更高的另一種控制方式



Selective Attention Enhances Neural Representations
Adapted from Greenlaw et al., 2020



模型架構

為什麼不用 **Encoder-Decoder**？

- 音樂具有高維複雜結構：rhythm / melody / harmony / timbre
- EEG 搜集成本高使得資料量有限，直接學習 EEG → Music 容易 overfit
- Encoder–Decoder 需同時學習 EEG representation 與 audio generation distribution

研究方法

- 使用 pretrained AudioLDM2 作為 music prior
- Latent diffusion 特性以Latent space 降低生成難度，讓 EEG 引導高層次音樂表徵，而不是直接預測 waveform 細節。
- 以 ControlNet adapter 將 EEG 作為 conditioning signal
- 不直接從 EEG 生成音樂，而是讓 EEG guide diffusion generation

EEG Preprocessing

為什麼要做 **minimal preprocessing**？

- 保留接近 raw EEG 的資訊，避免過度人工濾波或手動 channel selection。
- 我們只做 data-agnostic 的 minimal preprocessing，處理尺度、baseline 與極端值問題。
- 這樣可以減少對特定資料集、受試者或頻段假設的依賴，讓模型更有機會學到可泛化的 EEG → music latent representation。

Step	意義	為什麼選這個
Robust scaling	(x – median) / IQR，統一不同 channel / subject 的尺度	EEG 容易有 outlier，median/IQR 比 mean/std 更不怕尖峰雜訊
Baseline centering	扣掉前 1000 samples 平均值	移除初始 DC offset / 慢速漂移，避免模型學到電極偏移
Optional std clamping	將極端值限制在 ±kσ	降低瞬間 artifact 對訓練穩定性的影響

註：EEG 訊號常會因電極接觸、設備狀態或慢速漂移而產生整體偏移，這些偏移不一定與音樂解碼有關。因此扣除前 1000 個 samples 的平均值，讓模型更關注後續 EEG 的相對變化。相較於使用整段平均值，這種做法較不容易抵消音樂刺激後產生的反應。

Experimental Setup

Dataset

- 4 Subjects
- 128-channel EEG
- Drum, Vocal, Guitar Attention
- 1kHz

Model

- AudioLDM2 + ControlNet
- EEG Projector + Subject Adapter

Training

- 4 Training Songs
- 1 OOD Song
- 100 Epochs
- LR = 1e-4

Evaluation

- OOD Test
- CLAP Similarity

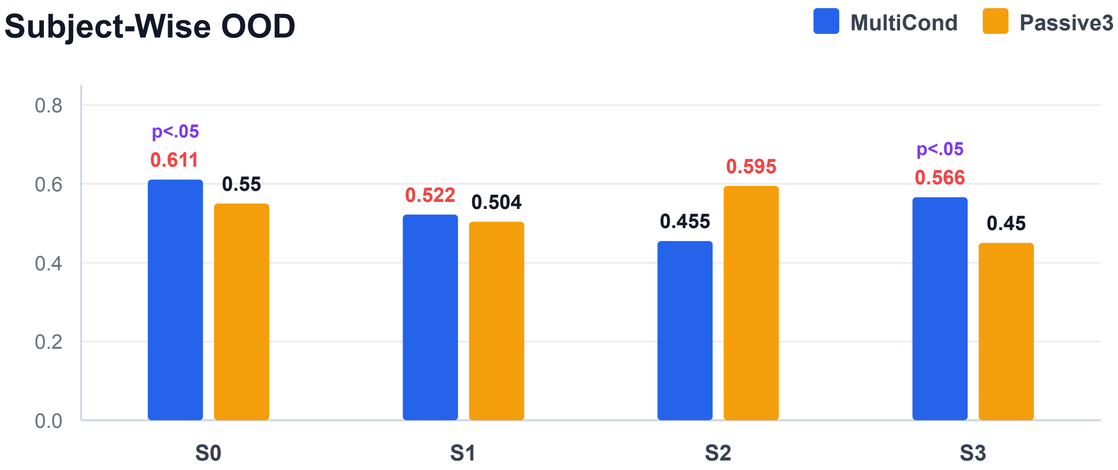
Main Results

主要觀察:
Train-All MultiCond 明顯提升
EEG 音樂生成的語意一致性

Target **MultiCond** **Passive**

主要觀察:
Subject-wise 結果中，MultiCond
在多數受試者優於 Passive3

Note:
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。



Stem-Level Analysis

Stem-Level Analysis (Subject 0)

Generated from \ Target stem	Drum Stem	Guitar 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.

主要觀察

- 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。
- Drum Stem 的 attention effect 最明顯：$p < 0.05$
- Guitar / Vocal Stem 的提升較弱，控制效果不穩定
- 結果顯示模型可部分利用受試者注意力資訊，尤其對節奏明顯的聲部較有效

Generalization

主要觀察

- S0：MultiCond 明顯優於 Passive3（$p < 0.05$）
- S1：兩者表現接近
- S3：Passive3 略高於 MultiCond
- 整體而言，MultiCond 在 LOSO 設定下呈現小幅平均提升，但仍不是非常穩定。

Target MultiCond Passive

跨受試者泛化下，MultiCond 雖然不一定精準重建 target，但比 Passive3 更能生成相近的結果。

結論與未來展望

Conclusion

1. MultiCond 整體優於 Passive3
多數受試者中，MultiCond 皆能提升生成音樂與目標音訊的語意一致性。

2. Train-All 顯示多重注意力整合的潛力
Train-All OOD 結果顯示，多受試者資料整合後，MultiCond 仍能維持較佳表現。

3. Stem-Level 結果支持 attention effect
Drum Stem 呈現最明顯提升，代表模型可部分利用注意力導向 EEG 資訊。

Future Work

1. 擴增受試者與資料規模
目前仍屬 pilot study，受試者數量有限，LOSO 結果容易受個體差異影響。

2. 提升跨受試者泛化穩定性
未來需建立更具代表性的訓練與測試資料，以改善 EEG subject variability 問題。

3. 探索泛化性更高的 EEG attention representation
若能穩定跨受試者表現，attention-guided EEG 有機會成為更可靠且泛用的 BCI 控制方式。