研究計畫書：Clef

題目：Hearing as Seeing: Polyphonic Audio-to-Score Transcription as a Vision-Language Task

（聽覺即視覺：將複音音樂轉譜視為視覺-語言任務之研究）

【 關於專案名稱：Clef 】

Clef，中文譯為「譜號」，是定義樂譜音高系統的關鍵符號。其最深刻的象徵意涵在於**不變性（Invariance）**：當樂曲因更換譜號而被移調時，儘管所有音符的絕對頻率皆已改變，其核心的和聲關係與旋律輪廓卻恆定不變。

本專案以此為名，旨在開發一個能穿透聲學表象（Acoustic Appearance），直接感知音樂內在幾何結構（Internal Geometry）的智慧模型，實現真正的「聽覺理解」。

> The *key* to reading sound

---

1. 研究背景與核心動機 （Background & Motivation）

1.1 核心問題：從「訊號紀錄」到「符號認知」的缺口

對音樂欣賞者來說，「錄音（MP3）」或「數位訊號（MIDI）」似乎已是保存音樂的完整解決方案。然而，儘管上述格式確實能保存物理**訊號（Signal）**，但它們在很大程度上遺失了音樂的**符號結構（Symbolic Structure）**。對於人類認知而言，僅僅「聽到聲音 （Hearing）」與真正「理解音樂 （Understanding）」是兩個完全不同的層次。這個從訊號到符號的轉換鴻溝，正是音樂資訊檢索（MIR）領域中，將錄音（Audio）轉化為樂譜（Score）被視為「聖杯」級難題的核心原因。

- **錄音 （Audio）**：是物理現象的紀錄（聲波的震動）。它稍縱即逝，無法被結構化地閱讀或分析。
- **樂譜 （Sheet Music/Score）**：是人類意圖（Intent）的符號化表達。它不僅記錄了音高，更記錄了作曲家的邏輯結構：調性（Tonality）、節拍（Meter）與樂句（Phrasing）。樂譜就像是人類語言的「文字」，是深度分析、教學與傳承的基礎。

本研究的終極目標，就是讓 AI 具備將「口語（音樂）」轉化為「書面文字（樂譜）」的認知能力。

1.2 為什麼 MIDI 不夠？訊號與符號的語義鴻溝

目前主流的 AI 音樂技術（如 Google Magenta MT3）大多停留在輸出 MIDI 格式。然而，MIDI 在心理學與實用性上存在巨大的缺陷：

- **MIDI （Musical Instrument Digital Interface）**：這是供**機器**讀取的指令，記錄的是物理動作：「在 3.141 秒按下琴鍵，力度 82，持續 0.523 秒」。它是**連續的物理訊號**。
- **樂譜 （人類意圖）**：記錄的是音樂邏輯（如：第三拍，四分音符）。它是**離散的符號結構**。

根據心理學的「範疇知覺（Categorical Perception）」，人類大腦在處理資訊時會自動過濾雜訊並進行分類。例如，演奏者彈出 0.523 秒的音符，人腦會將其認知為「四分音符」（邏輯概念），而非物理時間。MIDI 忠實記錄了 0.523 秒，當轉換為樂譜時，會變成一堆無法閱讀的碎片化音符，這就是「量化災難 （Quantization Artifacts）」。

結論：MIDI 只是物理訊號的「中間表徵」，丟失了音樂最核心的「語義資訊」。要解決這個問題，必須跳過 MIDI，直接讓 AI 學習「端對端（End-to-End）」的翻譯。

1.3 核心假設：音樂即「結構化的數列」

本研究結合認知心理學觀點，提出音樂處理應類比於**自然語言處理 （NLP）**。神經科學家 Patel （2003） 的 **SSIRH 假說**指出，大腦處理音樂語法與語言文法使用重疊的神經資源。更進一步，神經科學研究（如 Sur et al. 的雪貂實驗）顯示，大腦皮層具有高度的可塑性，聽覺皮層處理聲音頻率（Tonotopy）的方式，與視覺皮層處理空間影像（Retinotopy）在數學特徵上具有高度的同構性（Isomorphism）—兩者都在尋找「邊緣」、「紋理」與「模式」。

因此，本研究假設：音樂轉譜不應被視為單純的訊號處理（Signal Processing），而應被重構為一個**「視覺-語言翻譯（Vision-Language Translation）」**任務。我們試圖讓 AI 像人類閱讀圖形一樣，從聲音的頻譜圖中「看」出音樂的幾何結構。

---

2. 文獻回顧與缺口分析 （Literature Review & Gap Analysis）

本研究分析了 2021-2026 年的關鍵文獻，發現雖然學界在各個子領域（音訊分類、語音辨識、光學樂譜識別）皆有突破，但現有方法存在兩大根本性缺陷：（1）**架構侷限**：CNN 的局部感受野與 Whisper 的語音偏見，（2）**資料偏見**：過度依賴單一音色（通常是鋼琴）的訓練資料，導致跨音色泛化能力差。

2.1 傳統路徑：混合架構的挑戰

- **代表文獻：Zeng et al. （IJCAI 2024）**
    - **方法**：採用 **CNN （ConvStack）** 進行前端特徵提取，後端搭配 **GRU （Hierarchical Decoder）** 來建模樂譜的層級結構（如小節、拍號）。此研究旨在解決先前模型難以處理真實演奏資料中複雜結構的問題。
    - **限制與結果**：該研究在鋼琴轉譜任務上取得了顯著進展，在真實演奏資料集（ASAP）上實現了 MV2H 綜合指標 74.2%。值得注意的是，作者採用了多 SoundFont 增強策略（4 種鋼琴音色隨機選擇）並在真實錄音上進行微調，展現了對泛化能力的重視。然而，仔細拆解其 MV2H 子指標可以發現關鍵的**架構侷限**：其分數主要來自對「局部特徵」敏感的聲部（$F_{voi} = 88.4\%$）與時值（$F_{val} = 90.7\%$），但在最核心的**音高準確率（F_p）上僅為 63.3%，和聲準確率（F_{harm}）僅為 54.5%**。
      這個結果決定性地揭示了 CNN 架構的根本缺陷：其局部感受野能有效捕捉音符的起始點（onset detection），但無法「看見」構成和弦的、在頻譜圖上長距離分佈的音高幾何關係。儘管該方法在鋼琴領域內已實現 SOTA，其架構限制也指出了一個關鍵研究方向：需要一個能捕捉全域上下文的新架構來突破和聲理解的瓶頸。

2.2 語音模型路徑的啟示與侷限

- **代表文獻：Zhang & Sun （arXiv 2024）**
    - **方法**：該研究嘗試利用基於 Whisper 的 Transformer 架構，並設計了一種名為「Orpheus' Score」的記譜法，旨在同時轉錄旋律與和弦。儘管作者測試了微調 Whisper 預訓練權重的方案，但其最終推薦的模型是從頭開始訓練的。
    - **限制與結果**：在實驗中，該模型的詞錯誤率（WER）約為 46%。儘管作者視其為成功的嘗試，從實

用角度看，意味著轉錄結果含有大量錯誤。我們認為，此結果恰好揭示了「語音優先」架構的內在偏見。具體而言，為語音識別（ASR）設計的 Transformer 架構，其歸納偏置天然地對「時間軸」極度敏感（為了聽懂說話內容），但對「頻率軸」的解析度不敏感（因為語音不講究精確音高）。這種偏見解釋了為何即便是從頭訓練，模型在音樂轉譜任務上依然表現不佳。因此，這條路徑雖然有其探索價值，但也證明了要達到高保真度的音樂轉錄，需要一個專為音樂「頻譜幾何結構」而非「語音時序」設計的模型。

2.3 關鍵基石：視覺 Transformer 應用於音訊 （ViT for Audio）

- **代表文獻：AST （Audio Spectrogram Transformer） - Gong et al. （2021）**
  - **突破**：這篇 seminal paper 首度證明了 **「不需改變架構，直接將標準 Vision Transformer 應用於頻譜圖」**，在 AudioSet （聲音分類任務） 上取得了 SOTA 成績，擊敗了所有傳統 CNN 模型。
  - **關鍵發現**：研究證實，在 **ImageNet** （真實物體圖片） 上預訓練的 ViT 權重，可以極其有效地遷移（Transfer） 到音訊頻譜圖上。這為本研究提出的「凍結視覺編碼器 （Frozen ViT Encoder）」策略提供了堅實的實證基礎。
  - **缺口**：AST 僅被用於 **「分類任務 （Classification）」**（如：這是狗叫聲還是鋼琴聲），尚未被應用於需要極高結構精度的 **「序列生成任務 （Sequence Generation/Transcription）」**。

2.4 評估指標的反思：從序列到結構 （Rethinking Metrics: From Sequence to Structure）

- **代表文獻：Mayer et al. （2024）**, "Practical End-to-End Optical Music Recognition for Pianoform Music"
  - **對傳統指標的抨擊**：該研究深刻反思了在音樂轉錄任務中過度依賴序列錯誤率（SER/WER）的問題。作者指出，這類基於 Levenshtein 距離的指標存在嚴重缺陷：
    1. **無法跨模型比較**：結果高度依賴於模型內部使用的特定編碼格式（如 Kern、LilyPond），無法公平比較。
    2. **缺乏語意權重**：將所有符號錯誤等同對待，無法區分「錯一個音高」與「錯一個裝飾音」在音樂語意上的巨大差異。
  - **TEDn 指標的引入**：為了取代 SER/WER，該論文採用並實作了 **TEDn （Tree Edit Distance for MusicXML）**。此舉的關鍵在於，MusicXML 是當今所有主流打譜軟體（如 MuseScore, Finale, Sibelius）通用的交換格式。TEDn 基於 MusicXML 的樹狀結構，並針對音樂語意定義了特殊的替換成本，使其評估的結果能更好地反映人類在真實打譜軟體中修正樂譜所需的力氣（Effort-to-correct）。
  - **結論**：WER/SER 只適合在訓練中監控收斂，而 **TEDn 才是衡量系統對使用者是否有用的真實指標**，因為它提供了一個統一且具語意權重的比較標準。本研究將採納此觀點，以 TEDn 作為核心評估指標之一。

2.5 輸出格式的選擇：為何使用 Kern 記譜法？ （Choice of Output Format: Why Kern Notation?）

- **代表文獻：Alfaro-Contreras et al. （2024）, Román et al. (2019)** on polyphonic music transcription
  - **核心論點**：這些研究在處理多軌音樂（弦樂四重奏、合唱）時，選擇了 **Kern 格式，基於其天生的多聲部支援能力與音樂學分析的嚴謹性。
  - **對 MusicXML 的反思**：MusicXML 作為樹狀結構，其格式非常冗長（Verbose），對於序列模型來

說學習成本高昂，且其設計包含大量為軟體交換設計的排版細節，對模型是一種負擔。

- **對 ABC 的反思**：雖然 ABC 記譜法簡潔高效，但其設計初衷是民謠單旋律記譜，多軌支援能力有限（需透過 `V:` 標籤切換聲部），時間對齊依賴隱性的時值計算，增加模型學習複雜度。
- **`**Kern` 記譜法的優勢**：
    1. **多軌支援成熟 （Native Multi-track Support）**：使用 Tab 分隔的 Spine（列）結構，垂直對齊表示同時發生的音符，如同樂譜總譜（Score）般直觀。
    2. **時間對齊明確 （Explicit Temporal Alignment）**：同一行代表同一時間點，模型無需「數拍子」即可學習和聲關係，降低學習難度。
    3. **序列化簡單 （Easy Serialization）**：引入 `<coc>` (Change of Column) Token 即可將多軌並行結構攤平為單一序列，適合 Autoregressive Transformer（Alfaro-Contreras 2024 已驗證）。
    4. **音樂學標準 （Musicological Standard）**：Humdrum Toolkit 的核心格式，適合嚴謹的音樂分析，且已有大量高品質資料集（KernScores）。
    5. **適合分詞 （Tokenizer-friendly）**：其結構同樣適合應用 BPE 等 NLP 分詞技術，模型能學習複合音樂概念。
- **結論**：**Kern 記譜法在多軌支援、時間對齊明確性與音樂學嚴謹性上優於 ABC，且已有成熟的序列化方案（`<coc>` Token）可供參考。因此，本研究選擇** Kern 作為主要的輸出目標格式，最後再轉換成 MusicXML 以利用 TEDn 進行實用性驗證。

---

3. 研究方法 （Methodology）

本研究提出的 **Clef** 模型，本質上是一個**專為音樂理解設計的 Vision Language Model （VLM）**，將傳統的音訊處理問題重新框架為視覺與語言之間的轉譯。該模型是一個基於認知仿生設計（Bio-inspired Architecture）的端對端轉譜系統，其核心架構直接受到 **AST （Audio Spectrogram Transformer）** 論文（Gong et al., 2021）的啟發，該論文證明了純 Transformer 架構在音訊任務上的優越性。我們將 AST 的理念從「分類」任務延伸至「轉錄生成」任務。

3.1.1 輸入處理：將聲音視為圖像

如同 AST，我們將聲音的二維 Log-Mel 頻譜圖視為一張圖像。此頻譜圖會被切分成一系列 16x16 的 Patches（ImageNet 上使用 16×16 的 patch size 訓練），並展平為一個嵌入向量序列。透過加入可學習的位置編碼，模型能理解每個區塊在原始頻譜圖中的時頻位置。

3.1.2 編碼器 （Encoder）：遷移 ImageNet 知識的純視覺 Transformer

本研究的核心假設是，辨識頻譜圖中的音樂結構（如和聲、節奏）與辨識一般圖像中的幾何結構，在本質上是相通的。因此，我們的編碼器是一個**完全不使用卷積、純粹基於注意力機制的 Vision Transformer （ViT）**。

- **關鍵策略（跨模態遷移學習）**：為了讓模型具備強大的通用視覺特徵提取能力，我們直接採用在 **ImageNet** 上預訓練好的 ViT 權重。正如 AST 所證明的，這種跨模態遷移能極大提升模型對頻譜圖的理解能力，使模型如同經驗豐富的指揮家，能綜觀全局，瞬間捕捉到頻譜圖上長距離的諧波關係（和聲），而非像傳統 CNN 一樣僅用放大鏡檢視局部。
- **凍結權重 （Frozen Weights）**：我們將凍結 ViT 編碼器的權重。此策略不僅利用了 ImageNet 預訓練帶來的強大先驗知識，更是一種有效的正規化手段，強迫後端解碼器學會如何「解讀」這些通用的視覺特徵，而非讓編碼器去過擬合訓練資料中的特定音色，從而提升模型的泛化能力。至於是否需要在後期進行微調

（Fine-tuning），將作為消融實驗的一部分進行探索，以量化凍結策略對模型泛化能力的影響。

3.1.3 解碼器 （Decoder）：自回歸的音樂語言生成器

- **任務**：解碼器是一個標準的自回歸 Autoregressive Transformer，其任務是將編碼器提取出的視覺特徵序列，「翻譯」成符合音樂語法的 `**Kern 記譜法` 文字序列。
- **多軌序列化策略**：採用 Alfaro-Contreras et al. (2024) 提出的 `<coc>` (Change of Column) Token 機制，將多軌並行的 **Kern Spine 結構攤平為單一序列。例如：`4c <coc> 2C <coc> 4e <coc> .` 代表上聲部四分音符 C，下聲部二分音符 C，然後上聲部四分音符 E，下聲部延續前音。
- **核心機制：隱性量化 （Implicit Quantization）**：模型並非計算絕對的物理時間，而是基於音樂上下文學習預測下一個最可能的音樂符號（如在 4/4 拍的強拍上，模型會傾向於生成小節線或時值較長的音符）。這使其能自動修正演奏中的微小節奏誤差（Rubato，指演奏者為表達情感而做的彈性速度處理，不嚴格遵循拍點），生成結構工整的樂譜。

3.2 實驗設計 （Experimental Design）

本研究採用 **「Synthetic-to-Real」 （合成到真實）** 的訓練策略。由於真實世界缺乏完美的「音訊-樂譜」對齊資料，我們構建了一套自動化的資料生成與增強管線，用以訓練模型。

3.2.1 資料來源與分層 （Data Sources & Tiering）

我們將資料分為三個層級，分別用於預訓練、微調與最終測試，以反映從合成資料到真實世界的學習路徑：

資料集用途	資料集名稱	來源/描述	預期數量	角色
預訓練 (Pre-training)	**PDMX Dataset**	Public Domain MusicXML (來自 MuseScore)。為 FluidSynth 合成最佳化，具備商業友善授權。	254K+ 首	**結構主力**：提供海量、多樣的樂譜語法。
預訓練 (Pre-training)	**KernScores**	由音樂學家編碼的 `**kern` 格式樂譜，準確率極高。涵蓋巴洛克到古典時期。	108K+ 首	**品質保證**：用高品質資料對齊古典樂理。
真實測試 (Piano)	**ASAP Dataset**	核心真實資料集。包含真實鋼琴錄音 (Yamaha Disklavier) 與 MusicXML 對齊。	~50 小時	**鋼琴考官**：驗證在真實鋼琴演奏上的泛化能力。
真實測試 (Piano)	**Vienna 4x22**	小規模但高精度的真實鋼琴演奏，提供 MusicXML 對齊。	88 段演奏	**精度標竿**：補充 ASAP，用於精確度評估。
真實測試 (Multi-instrument)	**URMP**	唯一的具備獨立音軌的真實多樂器（室內樂）資料集，提供 MIDI/PDF 樂譜。	44 首作品	**泛化挑戰**：測試模型對非鋼琴音色的適應性。

3.2.2 資料增強金字塔：從訊號到音色的漸進式隨機化 （Hierarchical Augmentation Pyramid）

為了讓模型學會 **「Structure Invariance」 （結構不變性）**，我們設計了三層增強機制，有層次地剝離模型對表面特徵的依賴，最終透過 **Timbre Domain Randomization (TDR)** 強迫模型學習音色不變的音高幾何結構。

Level 1: Signal-Level Augmentation（訊號層增強）

> **目的**：模擬不同的錄音環境與設備，解決 「**Channel Robustness**」 問題。

- **聲學環境模擬 (Acoustic Simulation)**：
  - **Impulse Response (IR) Convolution**：隨機掛載「大教堂」、「浴室」、「錄音室」、「小房間」的空間殘響 (Reverb)。
- **訊號劣化 (Signal Degradation)**：
  - **Additive Noise**：加入白噪音、粉紅噪音、街道環境音、錄音帶底噪。
  - **Frequency Cutoff**：隨機 Low-pass / High-pass filter（模擬爛麥克風或手機錄音）。
  - **Compression**：動態壓縮，模擬現代流行音樂的響度戰爭。

Level 2: Source-Level Augmentation（音源層增強）

> **目的**：如同 Zeng et al. 的做法，解決 「**Intra-class Variance**」 **（類內變異）**，讓模型適應同一種樂器的不同狀態。

- **多樣化採樣 (Multi-Sampling)**：
  - 針對同一種樂器（如鋼琴），隨機切換不同的 SoundFont (如 Steinway, Yamaha, Upright, Honky-tonk)。
- **物理缺陷模擬 (Physical Imperfections)**：
  - **Detuning**：隨機對音高進行微小偏移 (± 10 cents)，模擬沒調準的樂器。
  - **Timing Jitter**：在生成 Audio 時加入微小的時間抖動，模擬人類演奏的不完美。

Level 3: Timbre Domain Randomization（TDR，音色領域隨機化）

> **目的**：解決 「**Timbre Invariance**」 **（音色不變性）**，強迫模型學習音色不變的音高表徵。

受機器人學（Robotics）中 Domain Randomization（Tobin et al., 2017）啟發，我們提出 **Timbre Domain Randomization (TDR)**，這是首次將領域隨機化原則應用於音樂轉譜任務的嘗試。

核心洞見：音樂的本質是**音高關係（Pitch Relations）**，而非**音色紋理（Timbre Texture）**。傳統模型容易過擬合於訓練資料的特定音色（如 Steinway 鋼琴），導致在未見過的音色上表現崩潰。

方法：

- **極端音色隨機化**：同一份樂譜，隨機選擇截然不同的音色生成音訊：
  - **聲學樂器**：Piano (Steinway, Yamaha), Violin, Flute, Pizzicato Strings
  - **合成音色**：Sawtooth Wave, Square Wave, FM Synthesis
  - **復古音色**：8-bit Chiptune (NES), Triangle Wave, PSG Sound

理論依據：
當音色（諧波結構、包絡、噪音特性）劇烈變化時，模型若要正確轉譜，唯一的策略是學習頻譜圖上的**幾何不變量（Geometric Invariants）**：

- 音程（Interval）：兩音在頻率軸上的相對距離
- 和聲（Harmony）：多音在垂直方向上的對齊模式

- 節奏（Rhythm）：音符在時間軸上的起始點分佈

這種方法實現了 **Timbre-Pitch Disentanglement（音色-音高解耦）**，使 ViT 編碼器聚焦於音高的幾何結構，而非音色的波形紋理。

3.2.3 實作技術堆疊 （Implementation Tech Stack）

音訊渲染（含 TDR 實作）：

- **聲學樂器**: `FluidSynth` 搭配多樣化開源 SoundFonts（Piano: Salamander, Steinway, Yamaha; Strings: Versilian, SSO; Winds: Sonatina）
- **合成音色**: `pyo` 或 `csound`（Sawtooth, Square, FM）
- **復古音色**: `Chip32` 或自製 8-bit 合成器（NES Triangle, PSG）

訊號處理與增強：

- **Level 1-2 增強**: `Spotify Pedalboard` 或 `TorchAudio`（Reverb, EQ, Noise）
- **頻譜轉換**: `nnAudio`（GPU 加速的 Log-Mel Spectrogram）

資料管線：

- **TDR 隨機選擇邏輯**: 每個 MusicXML 在訓練時隨機採樣一種音色（均勻分佈），確保模型不會偏向任何特定音色

3.2.4 評估與驗證

- **評估指標（Evaluation Metrics）**：傳統的詞錯誤率（WER）無法真實反映樂譜的可用性，因為它既不懂音樂語意，也不懂結構語法（Mayer et al., 2024）。因此，本研究將採用以下兩個互補的核心指標：
    - **音樂性（Musicality） - MV2H**：該指標由 Zeng et al.（2024） 提出，專門評估轉錄結果在音樂層面的準確性，包含音高、和聲與節奏等多個維度。相較於無法理解多聲部垂直關係的 WER，MV2H 能更真實地反映音樂內容的正確性。
    - **結構性（Structure） - TEDn**：該指標由 Calvo-Zaragoza et al.（2020） 提出並由 Mayer et al.（2024） 確立其重要性，用以評估輸出的 MusicXML 在語法結構上的正確性。相較於無法理解 XML 樹狀結構的 WER，TEDn 能有效衡量樂譜是否「語法正確」以致能被 MuseScore 等打譜軟體順利開啟與編輯，這對應了最終的產品可用性。
- **泛化能力驗證**：在合成資料上訓練後，將在 **ASAP Dataset**（真實人類鋼琴演奏資料）上進行 **Zero-shot** 測試，即模型在**未見過任何真實錄音資料**的情況下直接進行轉譜。此設計旨在驗證資料增強金字塔是否能讓模型具備跨領域泛化能力。測試結果將與 Zeng （CNN-based） 與 Zhang （Whisper-based） 進行對比，以驗證 ViT 架構的優越性。
- **消融實驗（Ablation Study）**：為釐清模型效能提升的來源，本研究將設計一系列對照組。透過比較不同編碼器架構（**ViT vs. CNN vs. Whisper**）與不同資料增強策略（基礎 vs. 強化）的組合表現，量化編碼器架構選擇與資料增強金字塔各自的貢獻，並驗證兩者結合的協同效應。具體而言，實驗將包含：
    - **架構對比**：在相同資料與訓練設定下，比較 ViT（視覺編碼器）、CNN（傳統音訊編碼器）與 Whisper（語音編碼器）在 **MV2H** 與 **TEDn** 指標上的表現差異，特別關注 MV2H 中的音高（F_p）與和聲（F_{harm}）子指標，以驗證 ViT 在捕捉長距離諧波關係上的優勢。
    - **資料增強效應**：針對每種架構，分別測試「基礎增強」（僅 Level 1 訊號層）與「完整增強」（Level 1-3 含 TDR）的效果，以驗證 **Timbre Domain Randomization** 對跨音色泛化能力的貢獻。
- **模型可解釋性分析（Model Interpretability Analysis）**：為「Hearing as Seeing」的核心論點提供直接

證據，我們將進行視覺化分析。具體而言，將繪製 ViT 編碼器的注意力熱圖（Attention Map），檢視模型在生成特定音高或和弦時，其注意力是否確實聚焦於頻譜圖上對應的諧波結構與特徵線，從而驗證其「看見」音樂結構的能力。

---

4. 預期貢獻 （Expected Contributions）

1. **驗證跨模態遷移 （Cross-Modal Transfer）**：首度驗證預訓練的視覺模型（ViT）可以有效處理聽覺頻譜訊號，且在捕捉和聲結構（Global Harmony）上優於傳統的 CNN 與 Speech Transformer 架構。這挑戰了傳統認為「聲音應該用語音模型處理」的直覺。
2. **提出 Timbre Domain Randomization（音色領域隨機化）**：首次將 Domain Randomization 原則從機器人學引入音樂轉譜領域，透過極端音色變異（鋼琴、小提琴、8-bit 合成器）強迫模型學習音色不變的音高表徵。實證端對端模型具備「隱性量化 （Implicit Quantization）」能力，學會的是音樂的「統計規律」而非物理測量，為音樂認知模型提供計算學上的證據。同時，展示透過「凍結視覺編碼器」策略，能在不依賴大量真實標記資料的情況下，訓練出具備跨音色泛化能力的轉譜模型。
3. **重新定義轉譜任務與實用價值**：將自動採譜從「訊號處理」問題重新框架為「視覺-語言翻譯」問題，為未來的音樂 AI 研究提供新的典範。最終，**Clef** 將解決長期以來「AI 轉出的樂譜無法閱讀」的痛點，為音樂教育、創作與保存提供一個真正可用的「聽寫工具」。

附錄：Novelty

本研究的獨特貢獻與創新點

本研究不僅是建構一個轉譜模型，更是一場 **驗證「聽覺即視覺（Hearing as Seeing）」假說的計算神經科學實驗**。其獨特性體現在以下三大創新：

創新點一：架構創新 — 聽覺認知的視覺化重構

（Architectural Innovation: Reframing Auditory Cognition as a Vision-Language Task）

- **痛點（Gap）**：現有 SOTA（如 Zeng et al.）使用 CNN 處理音訊，受限於局部感受野，難以捕捉長距離的和聲結構；或使用 Whisper（如 Zhang & Sun），受限於語音偏見，對音高不敏感。
- **本研究解法（Solution）**：
  - 提出**「Hearing as Seeing」** 理論架構，引用神經科學證據（如 Sur et al. 的雪貂實驗），假設大腦皮層具備通用幾何運算能力。
  - 首創將 **Vision Transformer（ViT）** 應用於複音音樂轉譜，利用 ViT 的 **Global Attention** 機制來模擬人類對音樂結構的「完形感知（Gestalt Perception）」。
- **科學價值（Impact）**：驗證跨模態的皮質可塑性（Cortical Plasticity），證明視覺預訓練模型可以遷移至聽覺任務，實現對音樂幾何結構的深度理解。

創新點二：資料創新 — Timbre Domain Randomization (TDR)

（Data Innovation: Timbre Domain Randomization）

- **痛點（Gap）**：傳統模型的泛化能力差，訓練於單一音色（如 Steinway 鋼琴），遇到不同音色即失效。現有研究（如 Hawthorne et al., 2018; Zeng et al., 2024）僅使用單一 SoundFont 進行訓練，導致模型過擬合於特定音色的波形紋理，無法跨樂器泛化。

- **本研究解法（Solution）**：
  - 首次將機器人學的 **Domain Randomization（Tobin et al., 2017）** 原則應用於音樂轉譜。
  - 提出 **Timbre Domain Randomization (TDR)**：同一樂譜用極端不同的音色生成訓練資料（Piano, Violin, 8-bit Chiptune, Sawtooth Synth），強迫模型學習音色不變的音高幾何特徵。
  - 引入 **音樂製作（Music Production）** 的領域知識，建立三層增強金字塔（訊號層 → 音源層 → **音色層 TDR**），系統性地提升模型的泛化能力。
- **科學價值（Impact）**：此方法不僅是資料增強，更是 **Timbre-Pitch Disentanglement（音色-音高解耦）** 與 **Invariance Learning（不變性學習）** 的實踐。據我們所知，這是首次在音樂轉譜任務中系統性地應用音色隨機化策略，填補了該領域的重要空白。

創新點三：任務創新 —— 隱性量化與生態效度

（Task Innovation: Implicit Quantization & Ecological Validity）

- **痛點（Gap）**：傳統 Audio-to-MIDI 方法依賴物理時間的顯性對齊，無法處理 Rubato（彈性速度），導致產出的樂譜碎片化，缺乏生態效度（人類無法視奏）。
- **本研究解法（Solution）**：
  - 利用 Transformer Decoder 的語言模型特性，實現 **「Implicit Quantization」（隱性量化）**。模型不計算絕對時間，而是預測最合理的「音樂語法」。
  - 重新定義輸出規格，從單純的音符列表升級為結構化樂譜，並使用 **MV2H** 與 **TEDn** 雙指標驗證其生態效度。