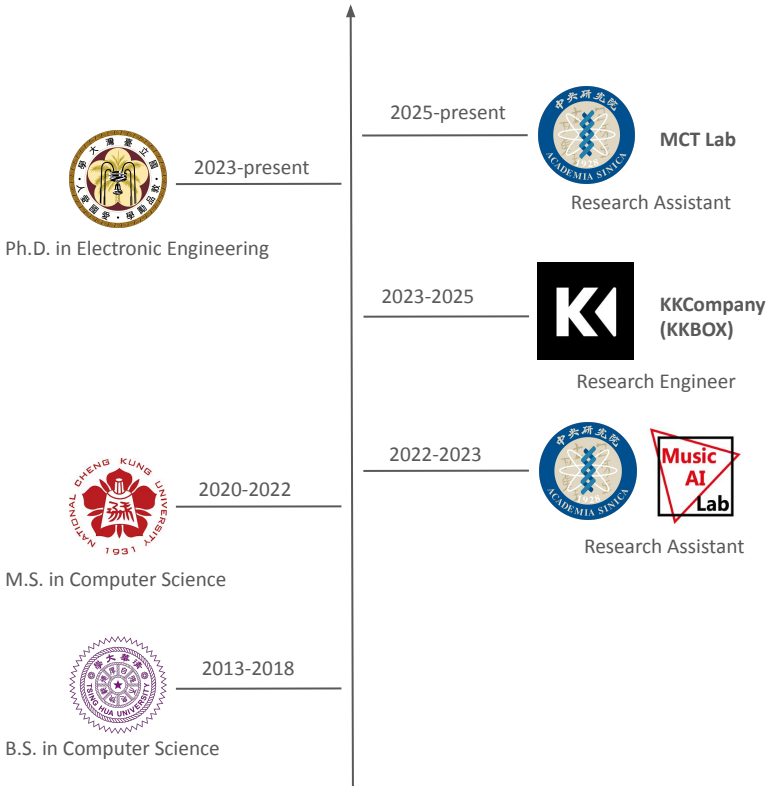
Music Adaptation
Arrangement & Cover Generation

Chih-Pin Tan 譚至斌
National Taiwan University

About me

- Chih-Pin (CP) Tan
- Email: tanchihpin0517@gmail.com
- Research Topic
  - Music generation & arrangement
  - Cover song generation



Introduction

Have you ever heard a song that sounds so familiar to another song?

Example

- papaya - Listen to me <-> 王心凌 - 愛你

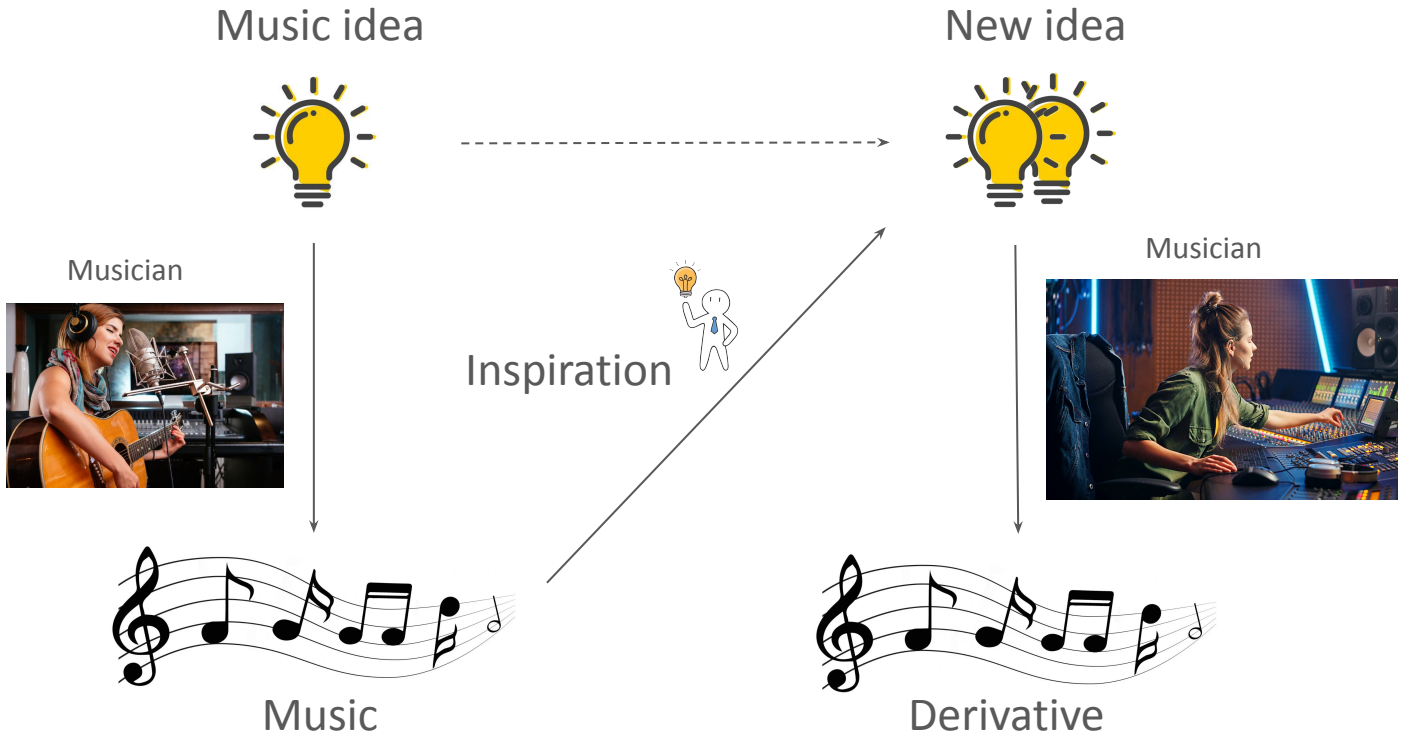
Song adaptation

An adaptation in music, which falls under the category of *Derivative Work*, refers to the process of altering or modifying an existing musical piece to create a new version.

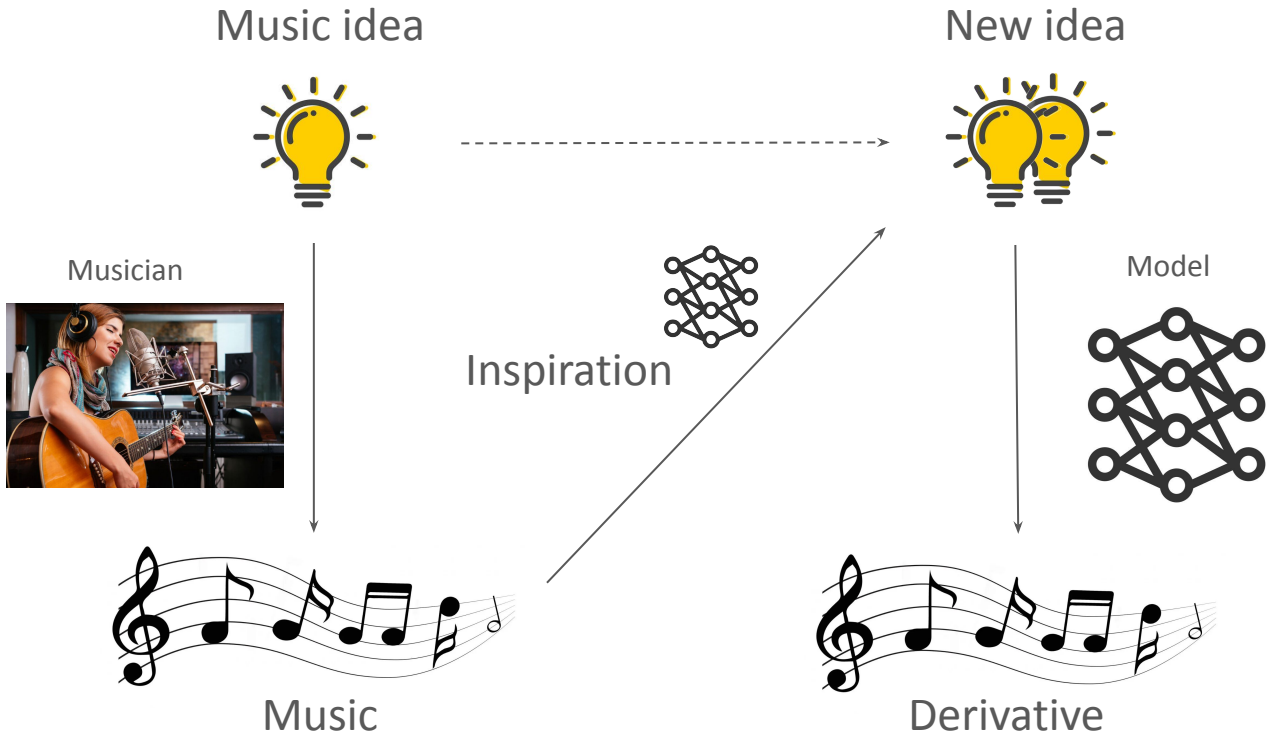
This creative process can take various forms, such as changing the **lyrics**, **melody**, **rhythm**, or **arrangement** of the original composition. It can also include ***translating a song into another language or reshaping it to match a different musical genre or style***.

https://www.aftersunsetmusic.com/resources/glossary/adaptation/

Create a music adaptation



Create a music adaptation with deep learning



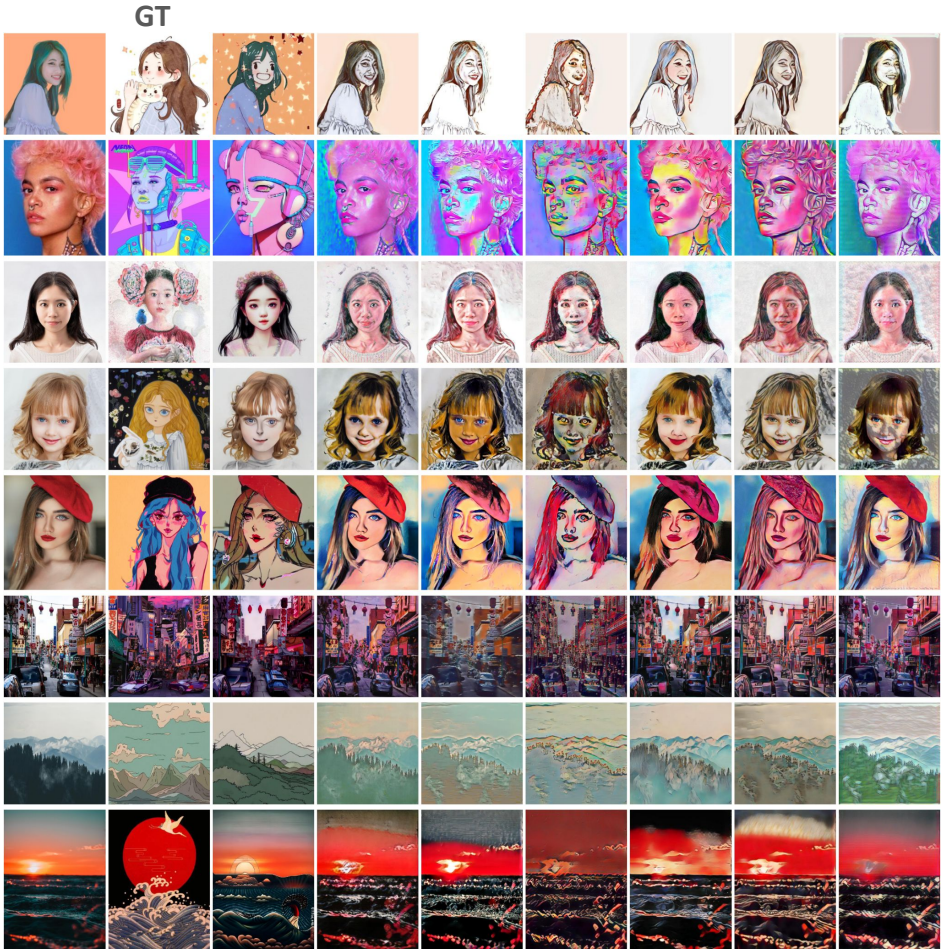
Challenges in cover song generation

- Data acquisition and preparation
  - Difficulty in collecting paired data.
  - Specifically, ensuring proper data alignment.
- Preserving the core musical idea
  - Defining the essential musical idea.
  - Establishing a measurable definition for "sounds similar".

Music Style Transfer

Image style transfer

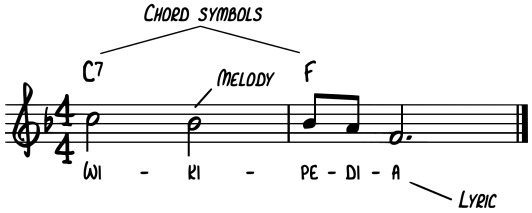
We take the **edges** of the content image and dress them up in the texture of the style image



Zhang et al. "Inversion-Based Style Transfer with Diffusion Models" in Proc. of CVPR 2023

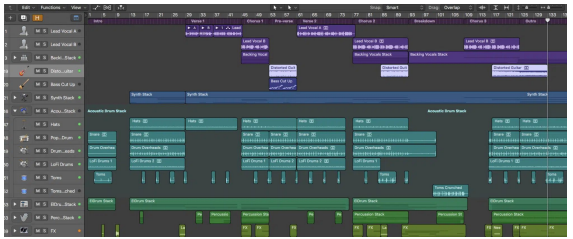
What's content and style of music?

- Content
  - Melody
  - Chord
- Style
  - Everything except content



Genre

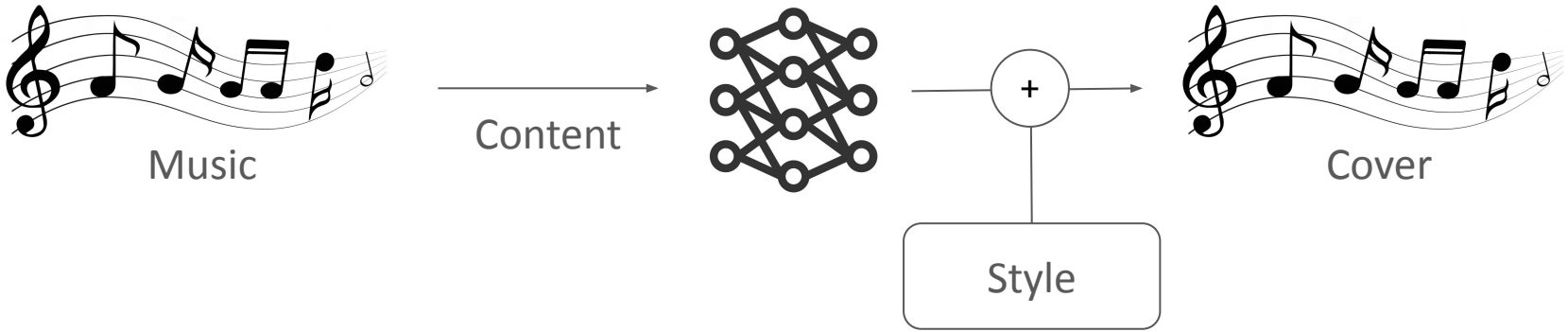
Instrument



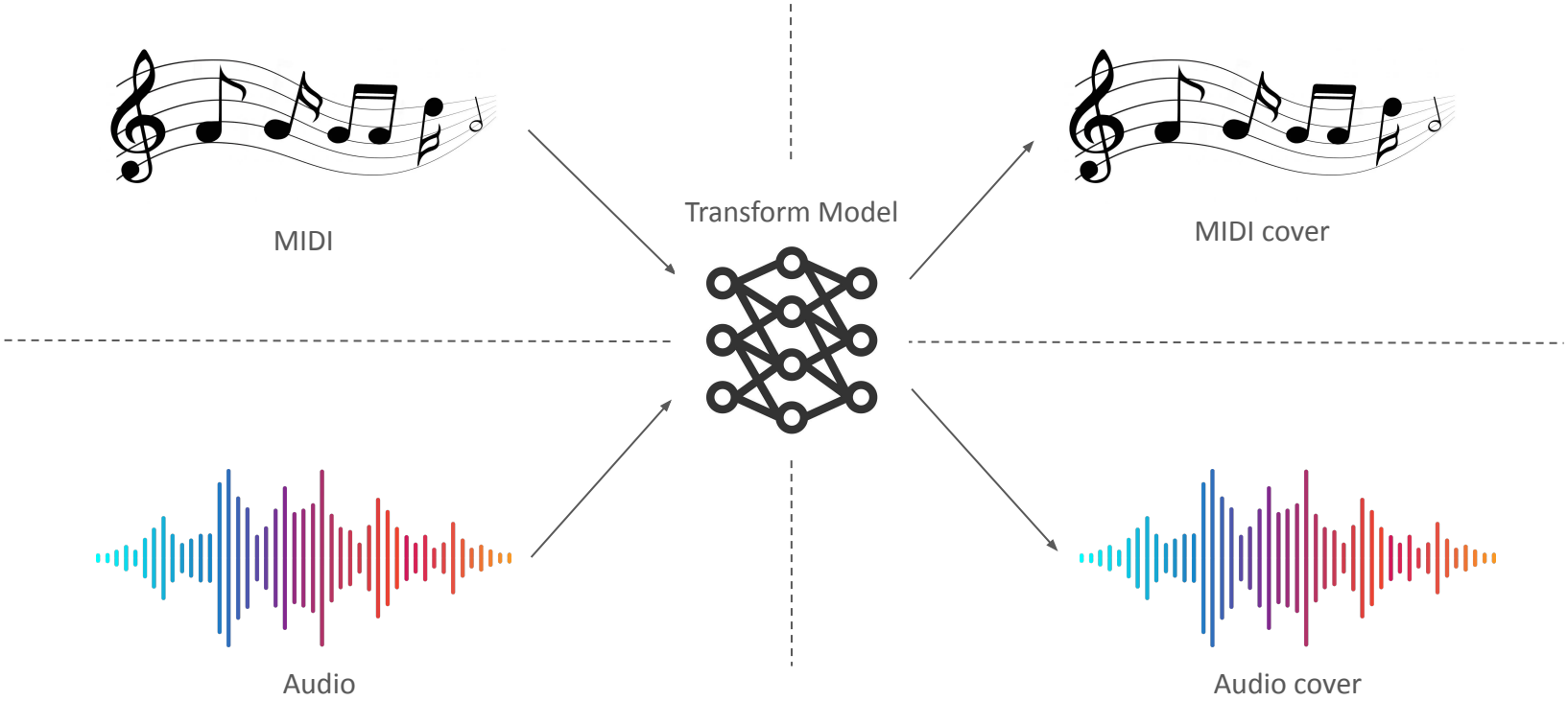
Arrangement

Problem formulation

- Cover Song Generation is formulated as a problem of ***Content-Preserving Style Transfer***
  - Extracting the content (core **melody**, **harmony**, and **structure**) from the original source track.
  - Replacing or re-rendering the expressive and temporal style (rhythm, timbre, articulation, dynamics) of the content with that of a target style (e.g., genre or artist).



Cross-modal music transformation

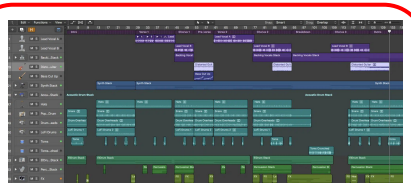


Music Arrangement

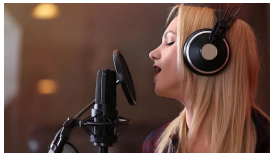
Music production workflow

The process of creating a final musical piece involves a sequence of creative and technical stages, transforming an initial idea into a polished, publishable track.

Compose



Arrange



Record

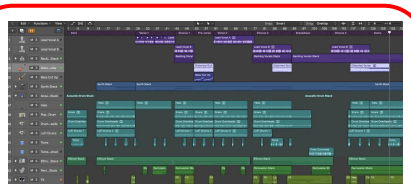
Mix

Publish

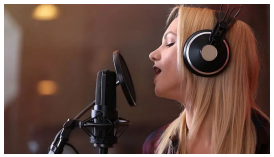
Music adaptation workflow

This workflow represents the creative process of developing a new piece of music ***inspired by or derived from*** existing material, rather than starting from an original composition.

Idea from other music



Arrange

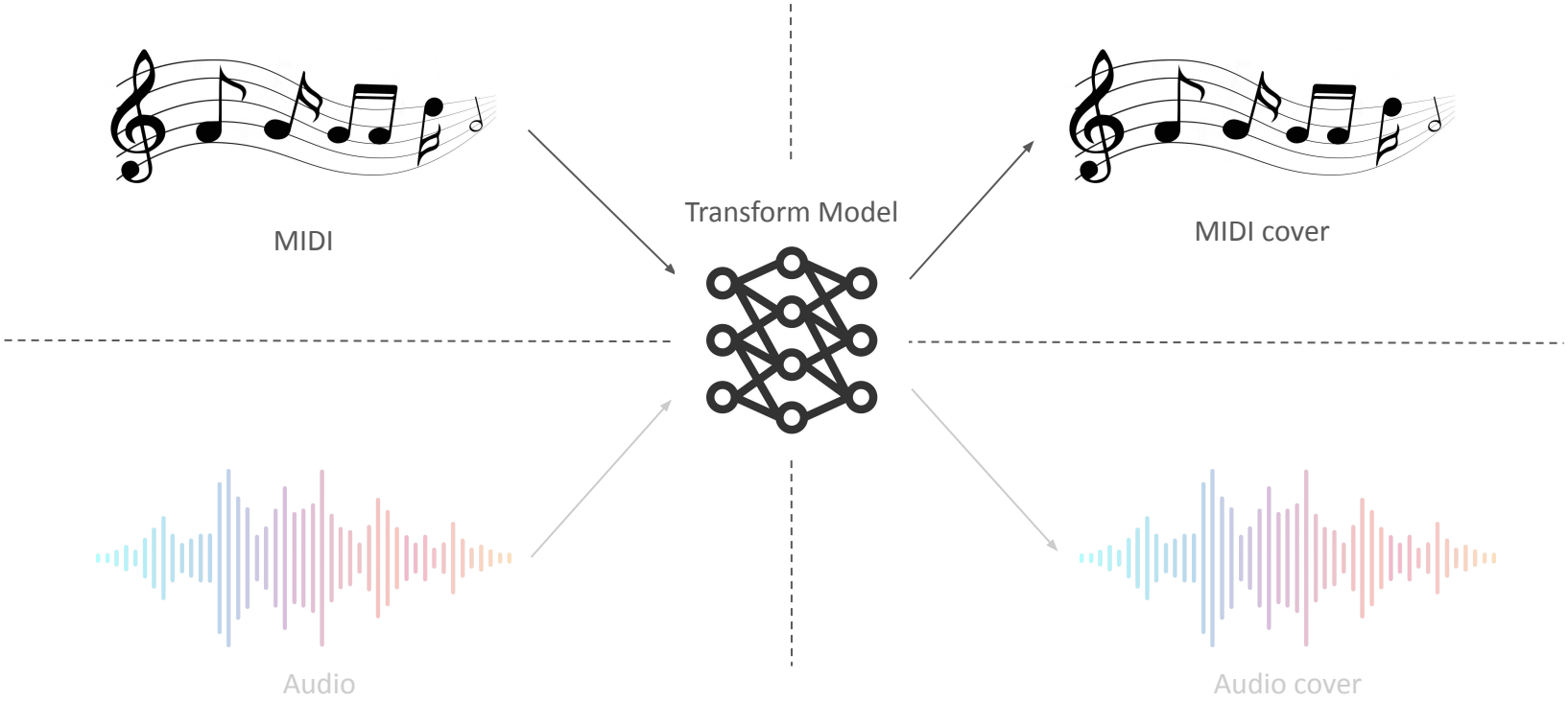


Record

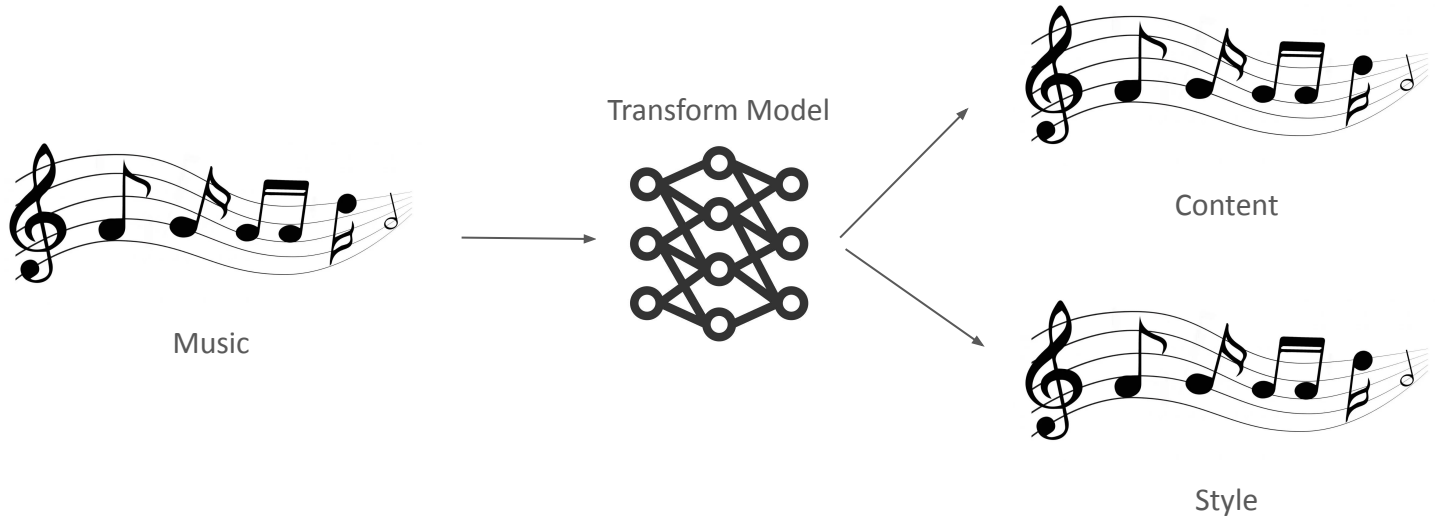
Mix

Publish

MIDI to MIDI

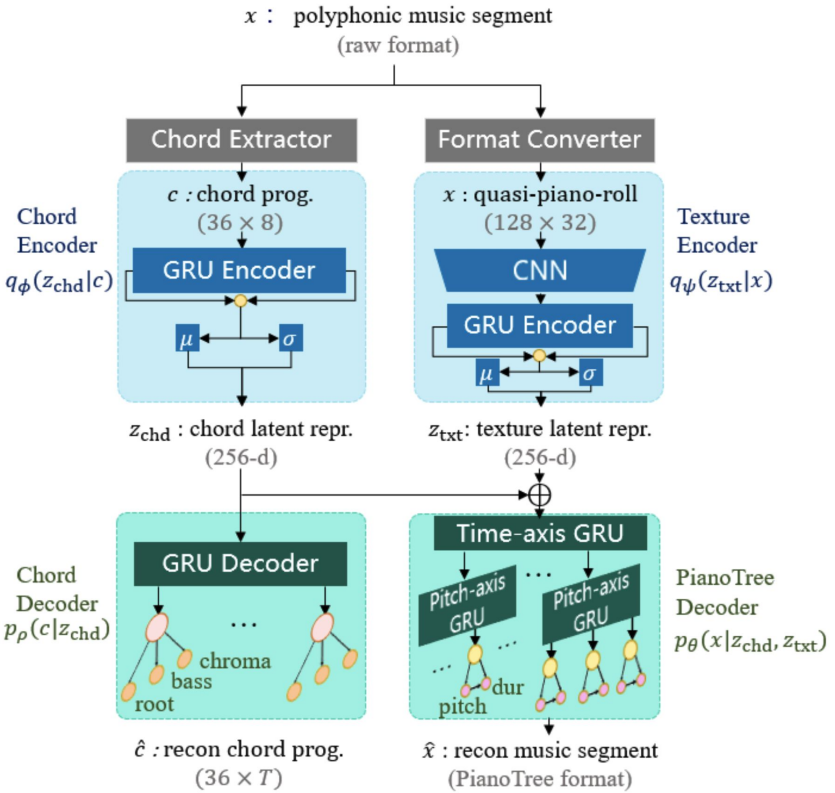


Task: disentangle music content



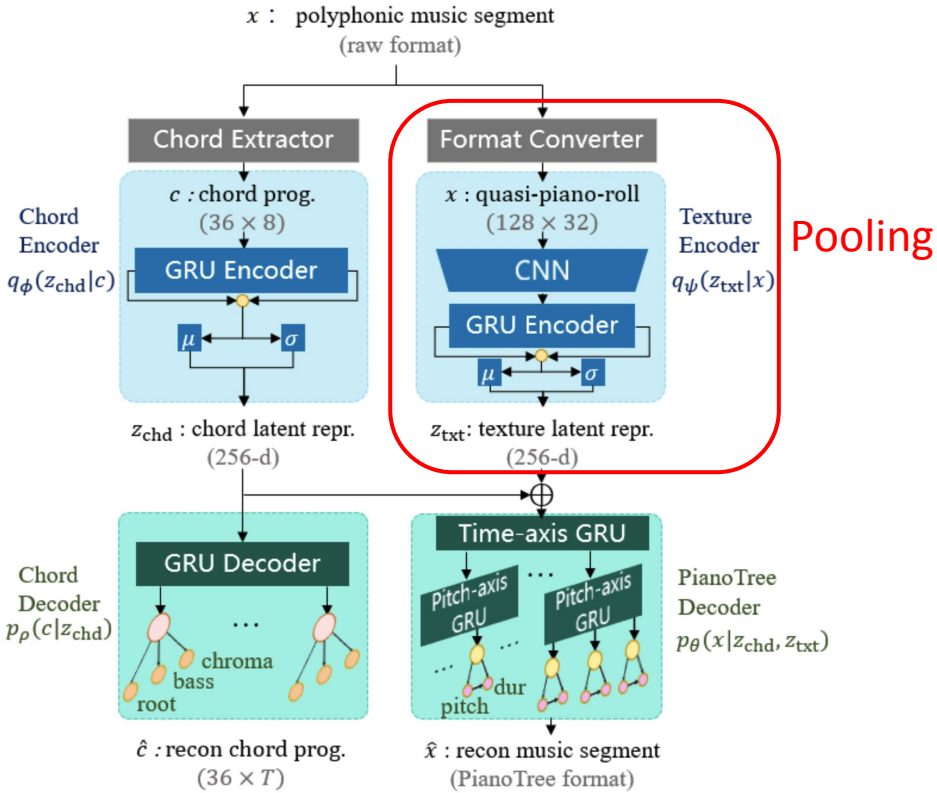
Controllable polyphonic music generation

- Disentangle chord (content) from music
  - Melody is not considered as content



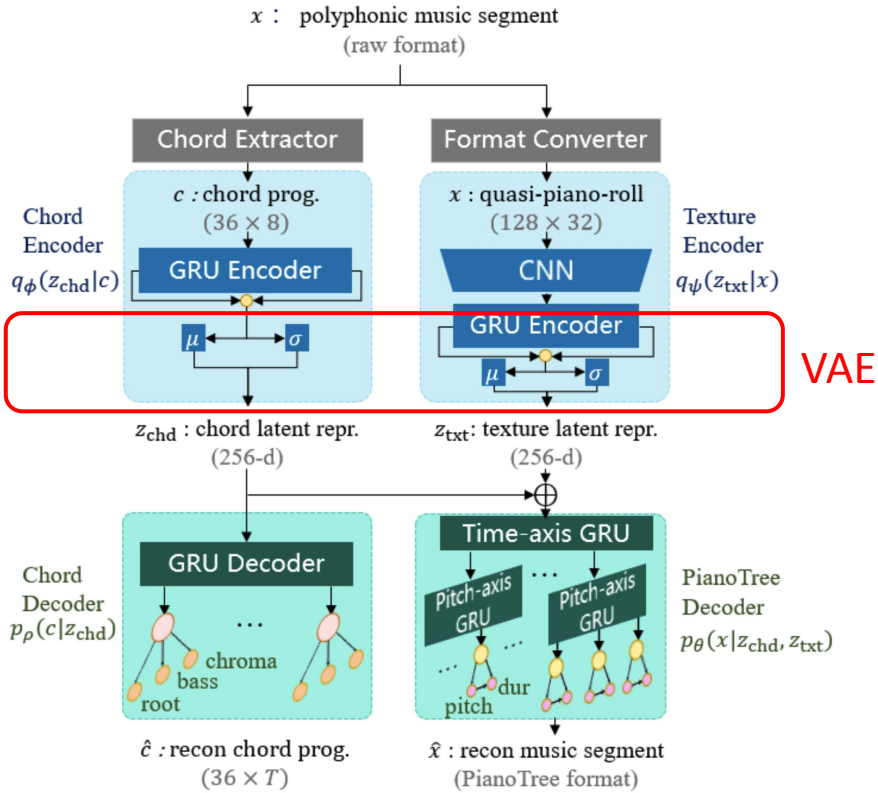
Wang et al. "Learning Interpretable Representation for Controllable Polyphonic Music Generation" in Proc. of ISMIR 2020

Controllable polyphonic music generation



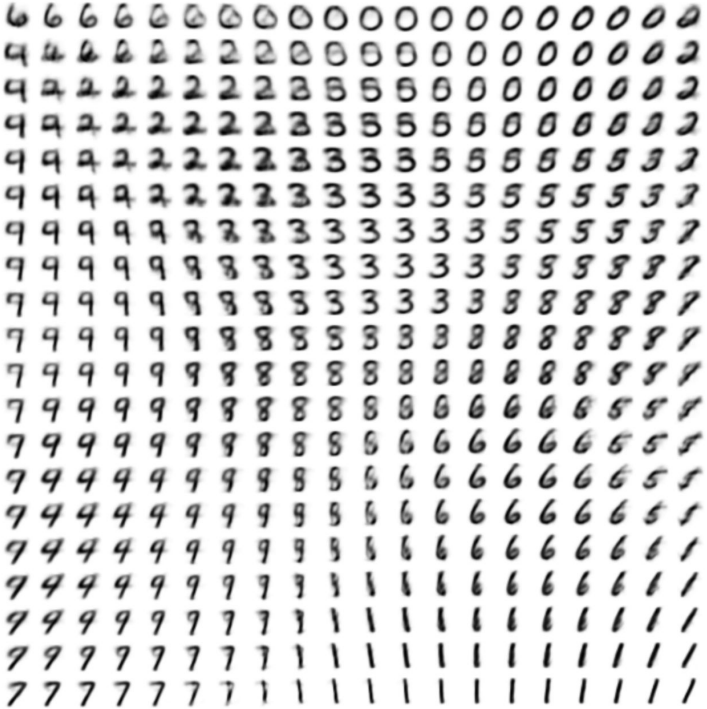
Wang et al. "Learning Interpretable Representation for Controllable Polyphonic Music Generation" in Proc. of ISMIR 2020

Controllable polyphonic music generation



Wang et al. "Learning Interpretable Representation for Controllable Polyphonic Music Generation" in Proc. of ISMIR 2020

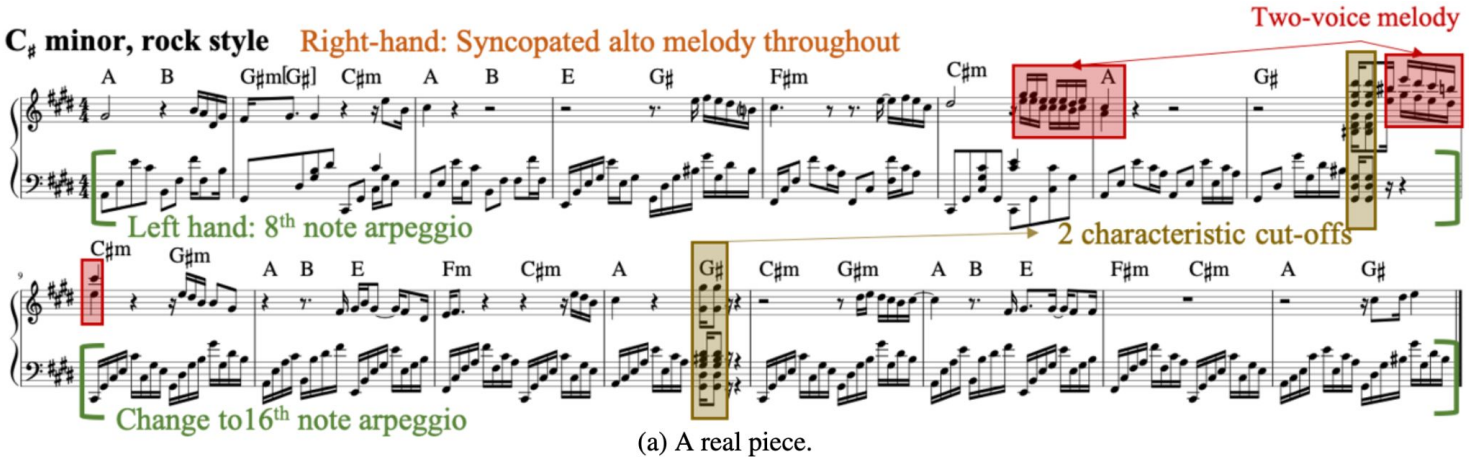
Controllable polyphonic music generation (VAE)



Wang et al. "Learning Interpretable Representation for Controllable Polyphonic Music Generation" in Proc. of ISMIR 2020

Controllable polyphonic music generation

However, the melodic fidelity of the generated output is poor.



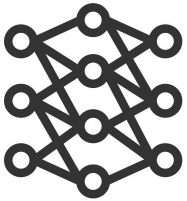
(a) A real piece.

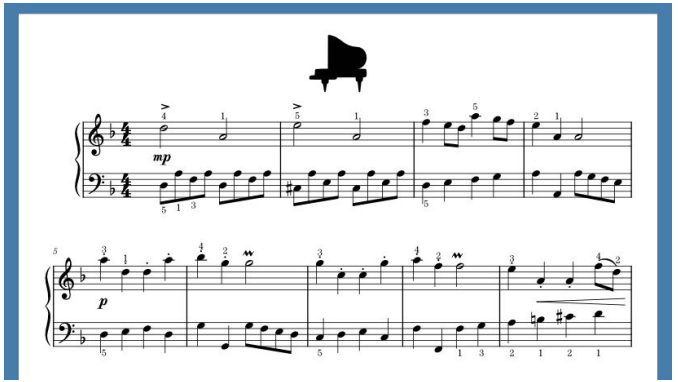
Wang et al. "Learning Interpretable Representation for Controllable Polyphonic Music Generation" in Proc. of ISMIR 2020

Task: piano accompaniment generation

Melody + Chord

Transform Model

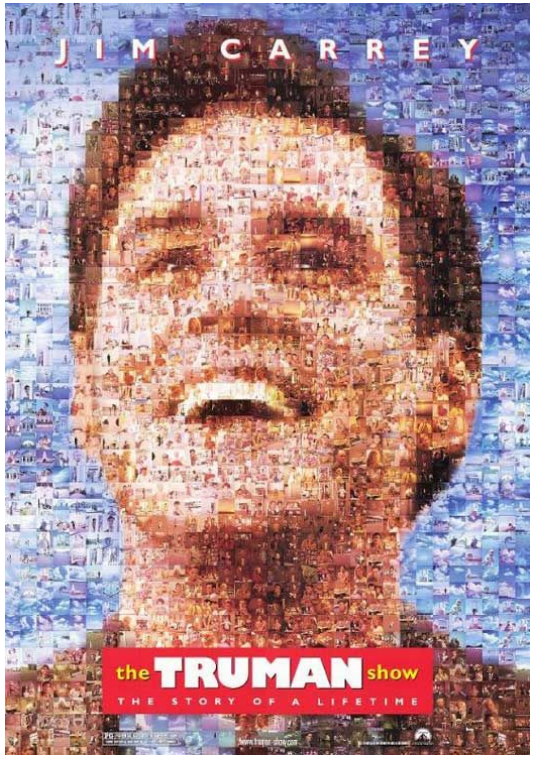




Piano accompaniment

AccoMontage (Accompaniment **Montage)**

"A montage is a film editing technique in which a series of short shots are sequenced to condense space, time, and information." - Wikipedia



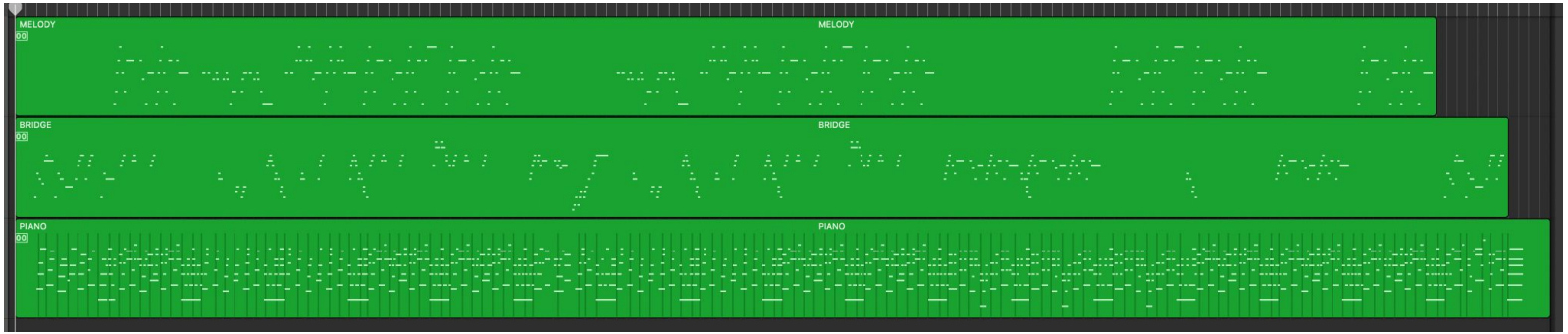
Zhao et al. "AccoMontage: Accompaniment Arrangement via Phrase Selection and Style Transfer" in Proc. of ISMIR 2021

AccoMontage (POP909)

Melody

Melody2

Accompaniment

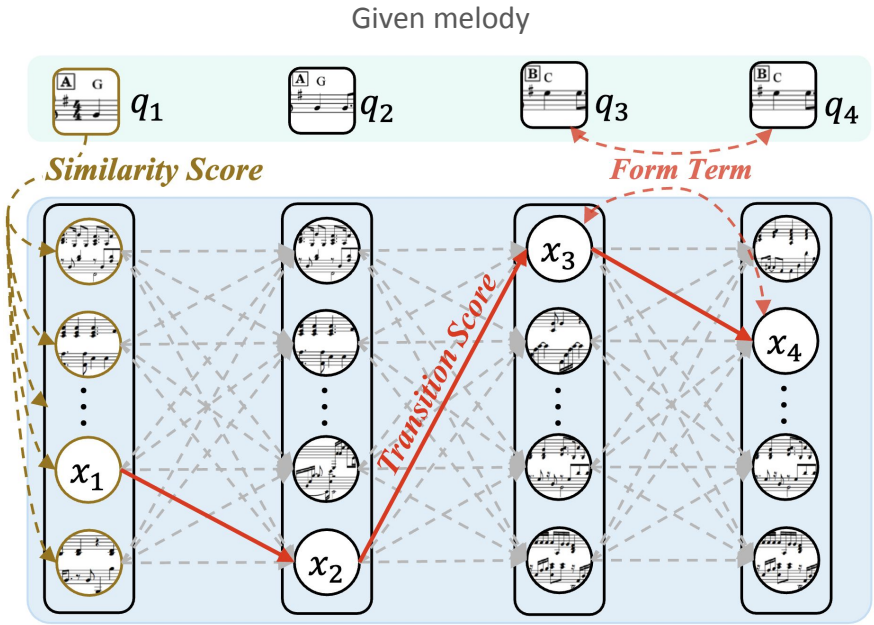


Structure A1 A2 B1 B2 ……

Wang et al. "POP909: A Pop-song Dataset for Music Arrangement Generation" in Proc. of ISMIR 2020
Dai et al. "Automatic Detection of Hierarchical Structure and Influence of Structure on Melody, Harmony and Rhythm in Popular Music" in Proc. of CSMC-MuMe 2020

AccoMontage

- Optimization
  - **Chord/Similarity Score**: Measures the fitness between the target melody query and the candidate accompaniment phrase
  - **Transition Score**: Measures the smoothness and musicality of the connection between two successive accompaniment phrases
  - **Form Term**: Enforces constraints on the overall musical structure and arrangement (e.g., song sections A, B, C).

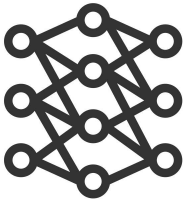


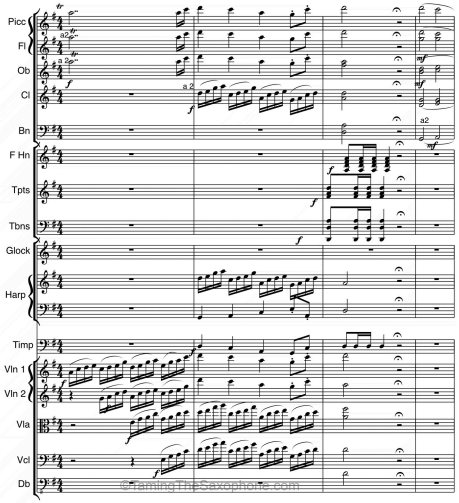
Zhao et al. "AccoMontage: Accompaniment Arrangement via Phrase Selection and Style Transfer" in Proc. of ISMIR 2021

Task: multitrack arragement

Melody + Chord

Transform Model

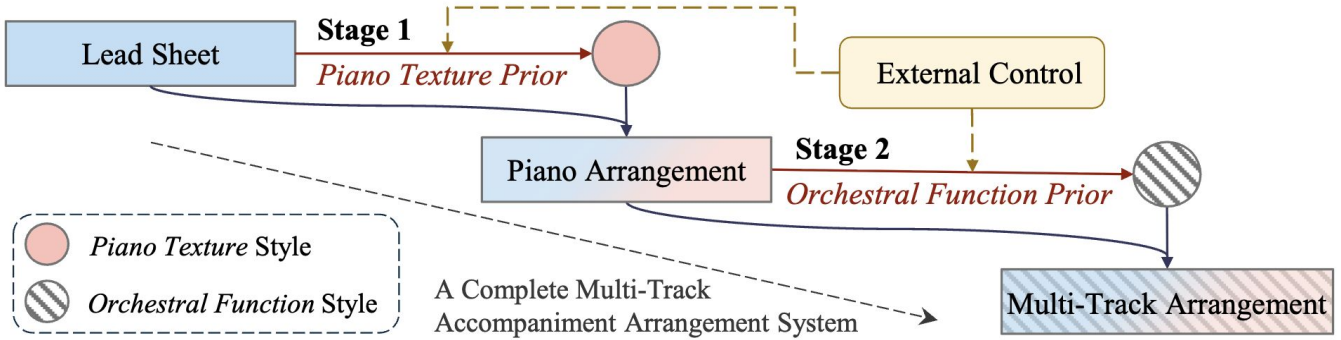




Multitrack arrangement

Multi-Track accompaniment arrangement

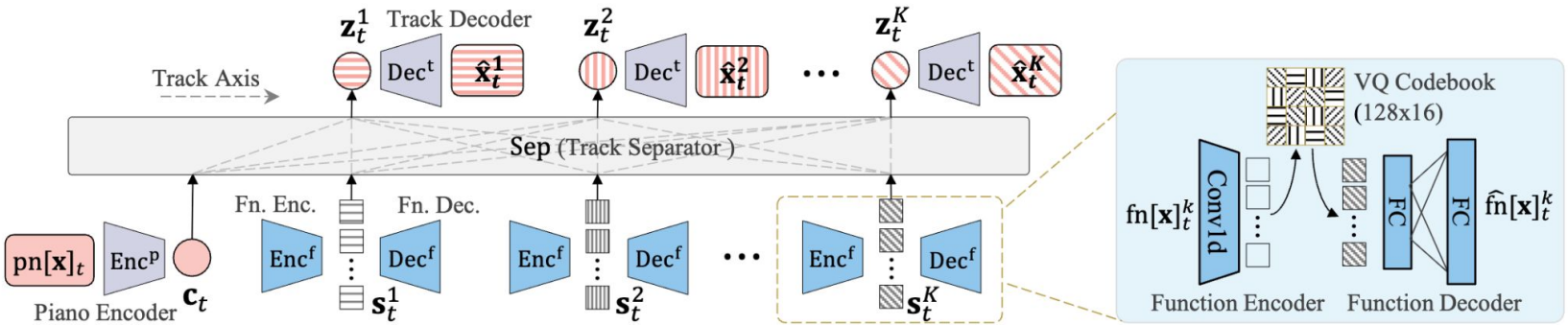
Concept 1: Arrange multi-track accompaniments from a simple lead sheet.



Zhao et al. "Structured Multi-Track Accompaniment Arrangement via Style Prior Modelling" in Proc. of NeurIPS 2024

Multi-Track accompaniment arrangement

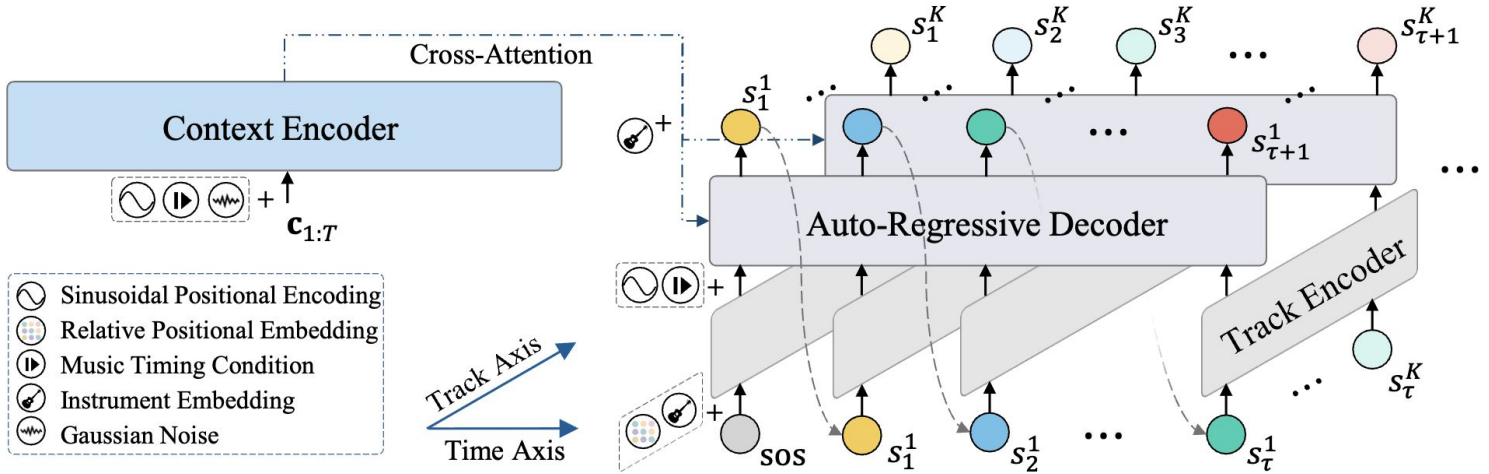
Concept 2: Generate each track by deriving the musical content from the **reduced piano** arrangement and assigning its instrumental role using a **quantized orchestral function**.



Zhao et al. "Structured Multi-Track Accompaniment Arrangement via Style Prior Modelling" in Proc. of NeurIPS 2024

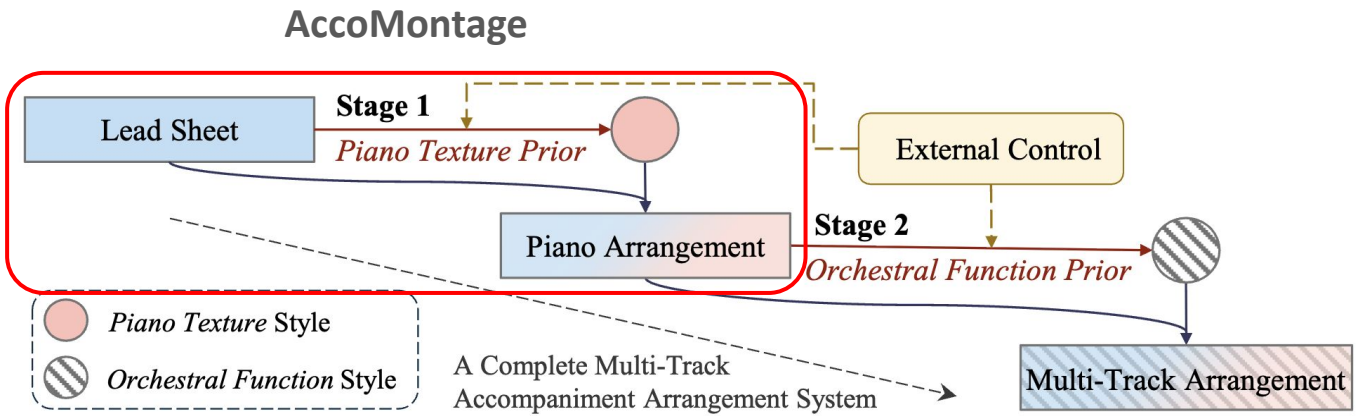
Multi-Track accompaniment arrangement

Model: Generate a complete orchestral score via the automatic orchestration of a piano reduction.



Zhao et al. "Structured Multi-Track Accompaniment Arrangement via Style Prior Modelling" in Proc. of NeurIPS 2024

Multi-Track accompaniment arrangement



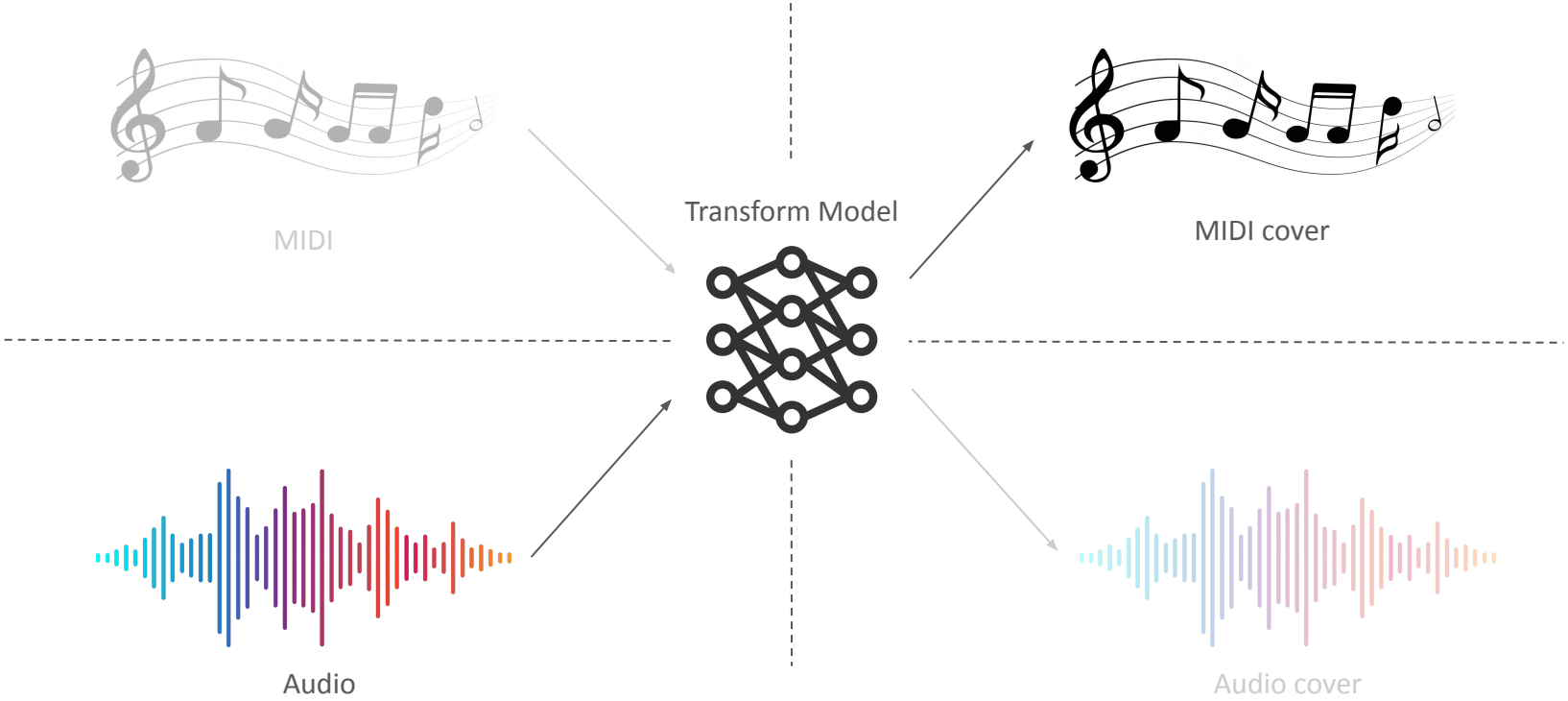
Zhao et al. "Structured Multi-Track Accompaniment Arrangement via Style Prior Modelling" in Proc. of NeurIPS 2024

We didn't mention how to get content …

- Pop909 is a ***perfect*** dataset
  - It may not fully represent real-world musical complexity
- Melody extraction from MIDI is easier than from raw audio.

Cover (Adaptation) Generation

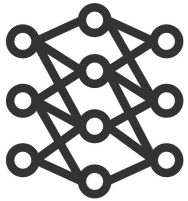
Audio to MIDI

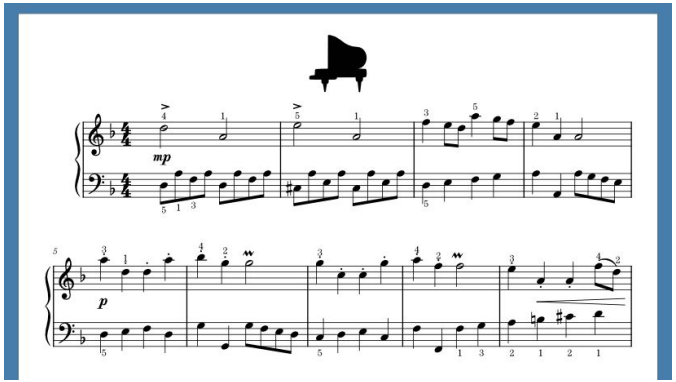


Task: piano cover generation

Song audio

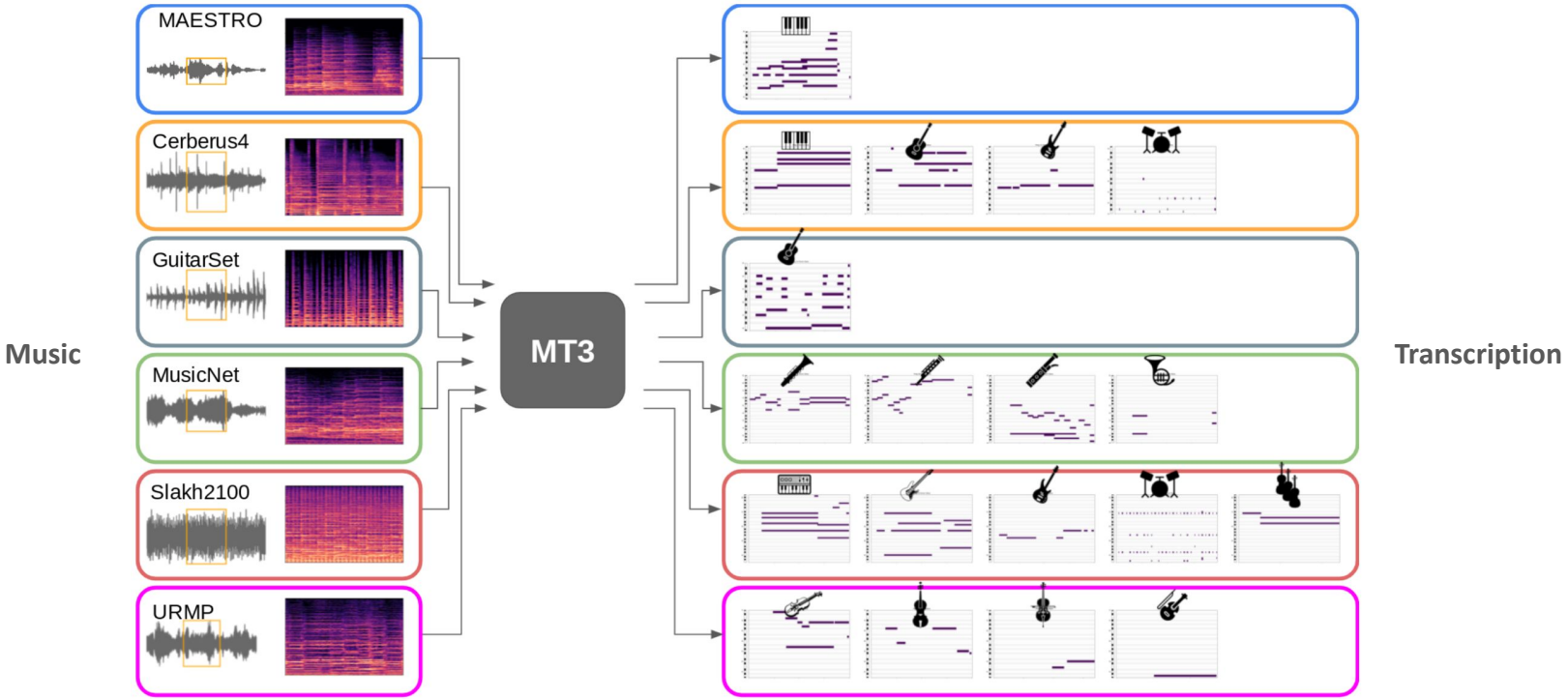
Transform Model





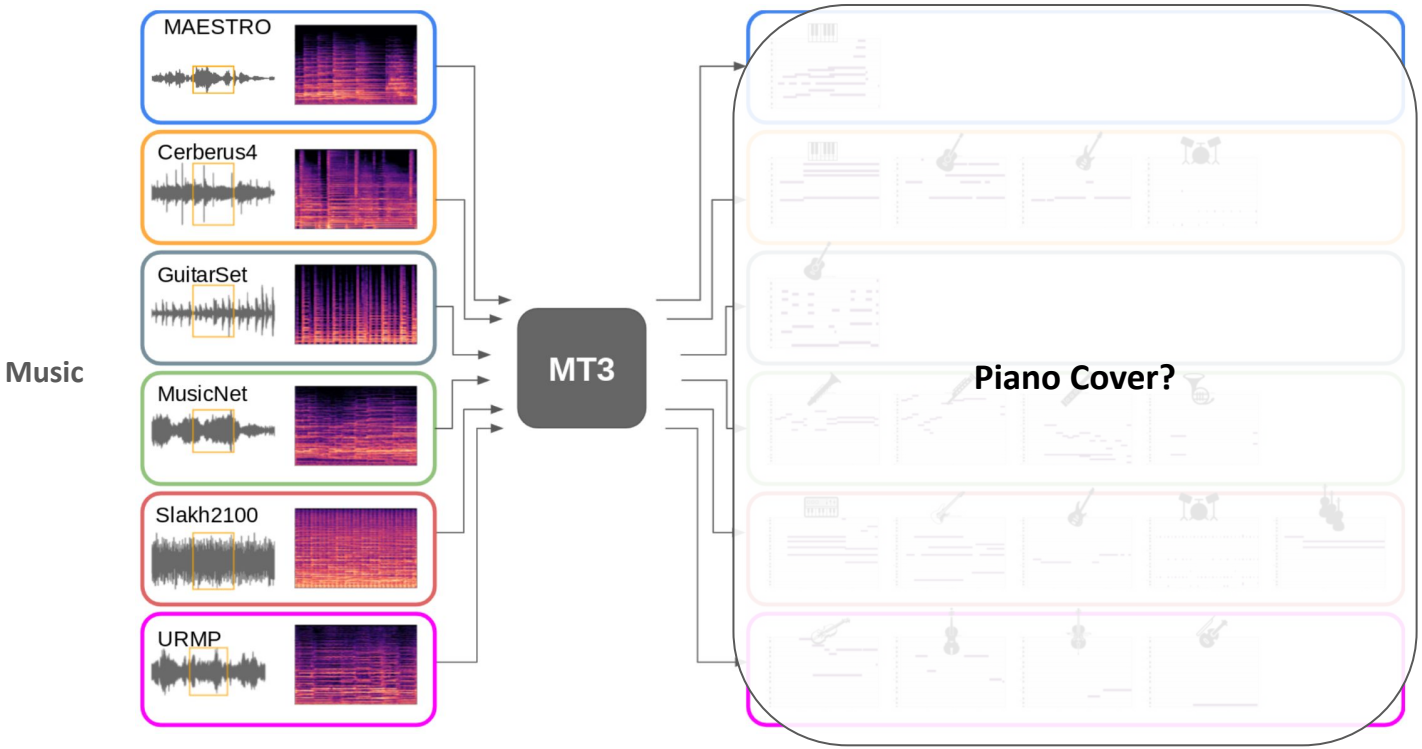
Piano cover

Pop2Piano (MT3)



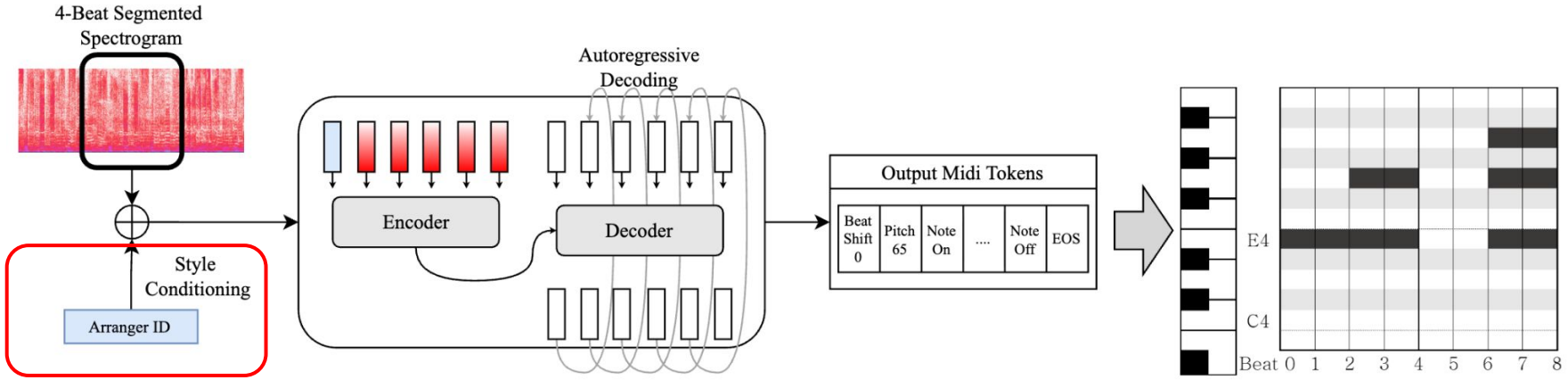
Gardner et al. "MT3: Multi-Task Multitrack Music Transcription" in Proc. of ICLR 2022

Pop2Piano (MT3)



Gardner et al. "MT3: Multi-Task Multitrack Music Transcription" in Proc. of ICLR 2022

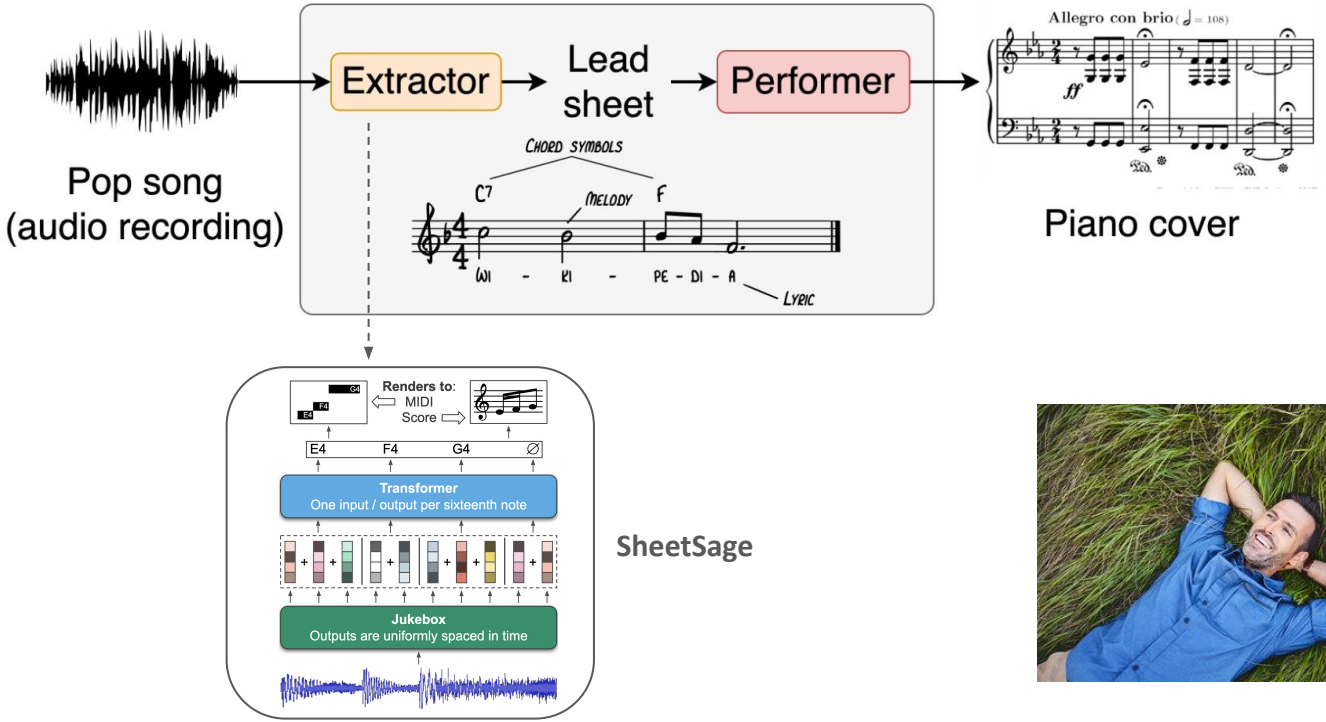
Pop2Piano



Choi et al. "Pop2Piano : Pop Audio-based Piano Cover Generation" in Proc. of ICASSP 2023

Rethink: cover generation = extraction + generation

PiCoGen1



Tan et al. "PiCoGen: Generate Piano Covers with a Two-stage Approach" in Proc. of ICMR 2024

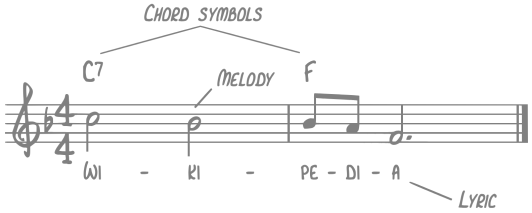
Content preservation

The Question: Should the core elements (melody, harmony, rhythm) be preserved exactly when generating a cover?

- No. We aim for **perceptual similarity**, not strict note-for-note identity. The content should be **adaptable** to the new style.

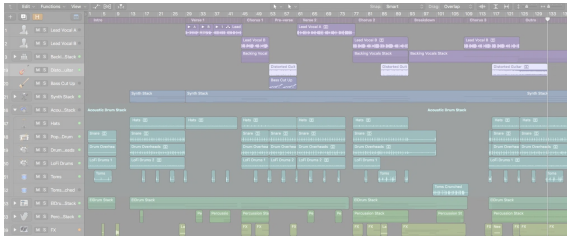
What's content and style of music?

- Content
  - Melody
  - Chord
- Style
  - ***Everything except content***



Genre

Instrument



Arrangement

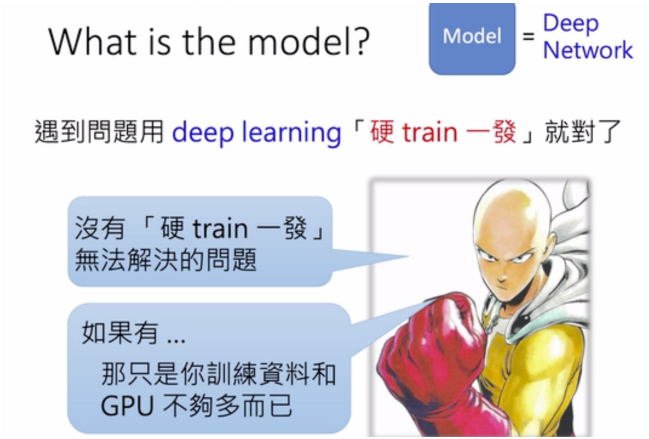
Preservation of content and style

The Challenge of Content: to what extent must the **original melody** be altered (or preserved) to fit a target?

- Specifically: how do we define the "**essential content**" of the melody?
- The detail: which notes are **mandatory** for recognition, and which can be changed for embellishment?

The Challenge of Style: beyond the core melody and harmony, which **other musical attributes** should be preserved in the cover song?

Extended reading: Dai et al. "Music Style Transfer: A Position Paper" in Proc. of MUME 2018



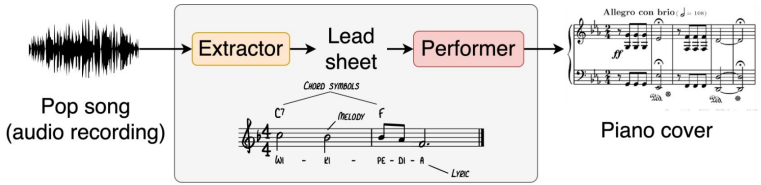
How about data-driven approach?

"..... 不知道什麼是'硬train一發' 的，我試著解釋一下什麼是'硬train一發'。這個，'硬train一發' 啊...它是一個非常神奇的東西：它是一種***信念***，它是一種***夢想***，它是一種***浪漫***，它是***人類最原始的衝動***，它是***恆古以來人類的目標*** ....." - by 李宏毅

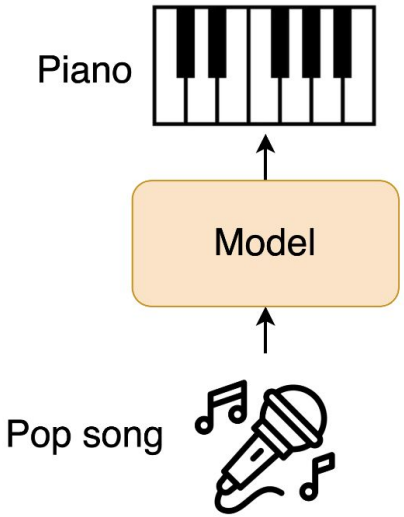
遇到问题就用Deep Learning【硬train一发】就对了: https://www.bilibili.com/video/BV1kE411A7Uf/

PiCoGen2

We aim to build a complete, end-to-end system based on the PiCoGen1.



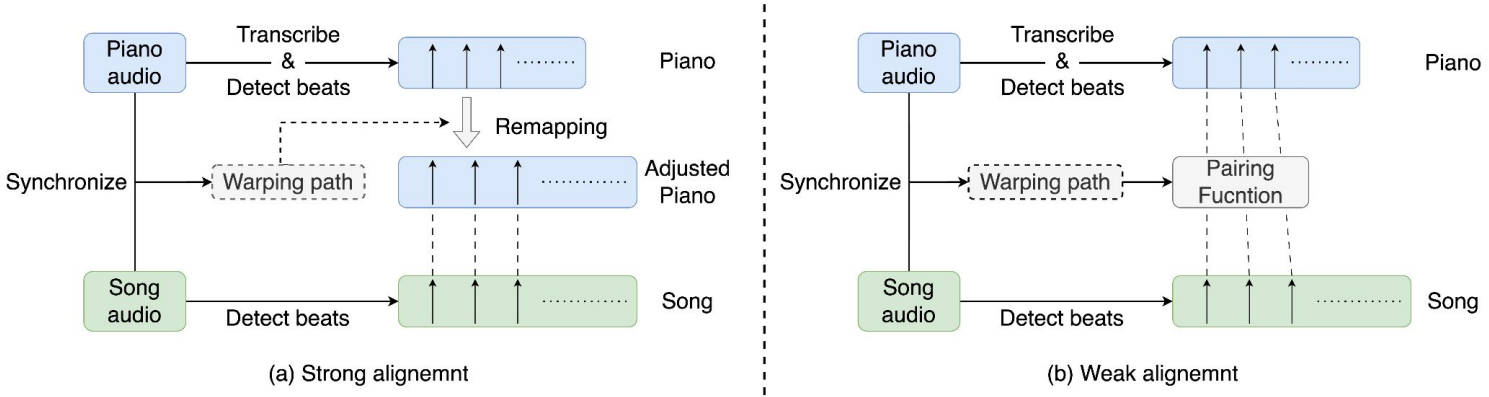
PiCoGen1



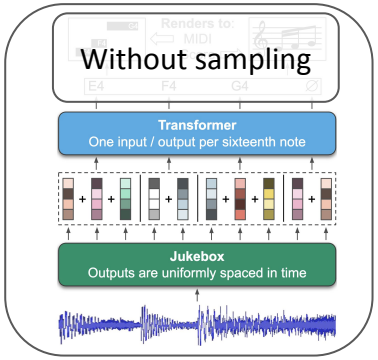
PiCoGen2

PiCoGen2 - data

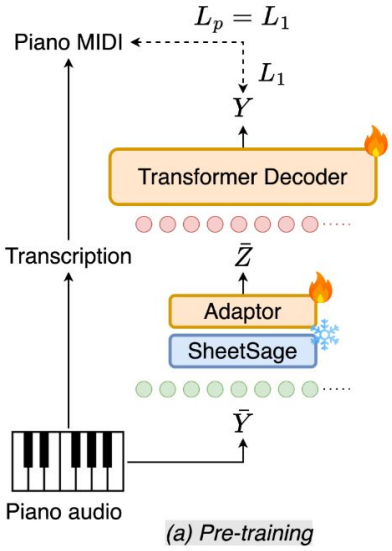
Training data is synchronized based on the musical **beat** (quantized time), rather than raw **time** values.



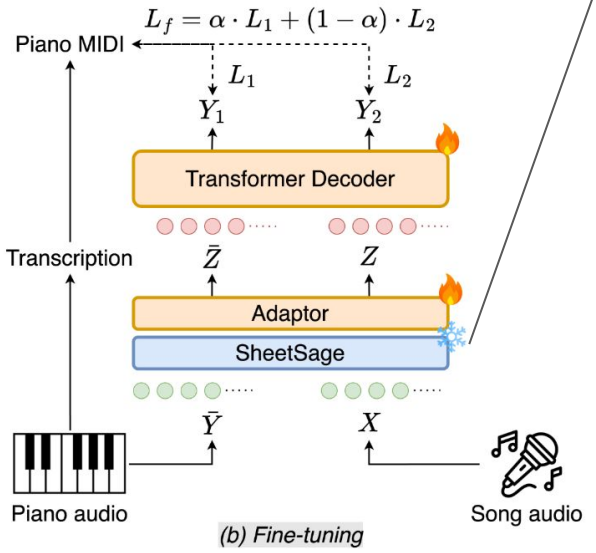
PiCoGen2 - training & inference



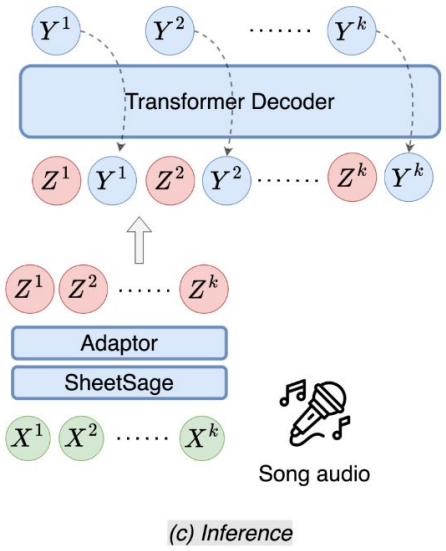
SheetSage



(a) Pre-training



(b) Fine-tuning



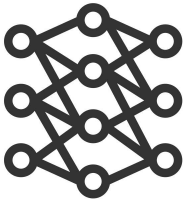
(c) Inference

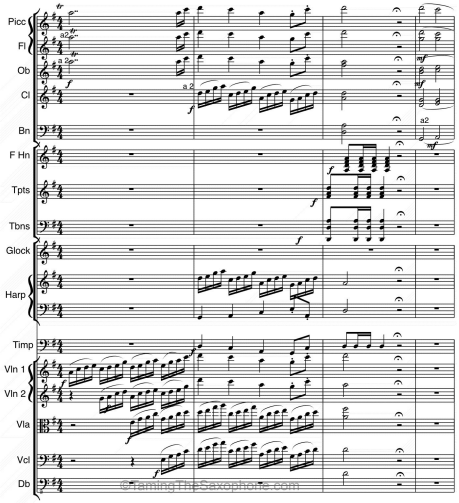
PiCoGen2 - demo

Task: multitrack cover

Song audio

Transform Model

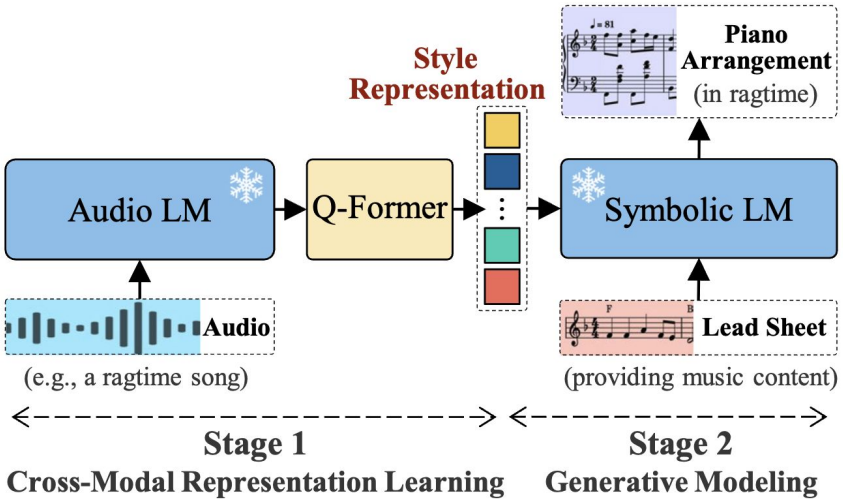




Multitrack cover

BOSSA

Learn a **style representation** (e.g., "ragtime") from an audio example and then use that compact style vector to condition a **symbolic music generation model**

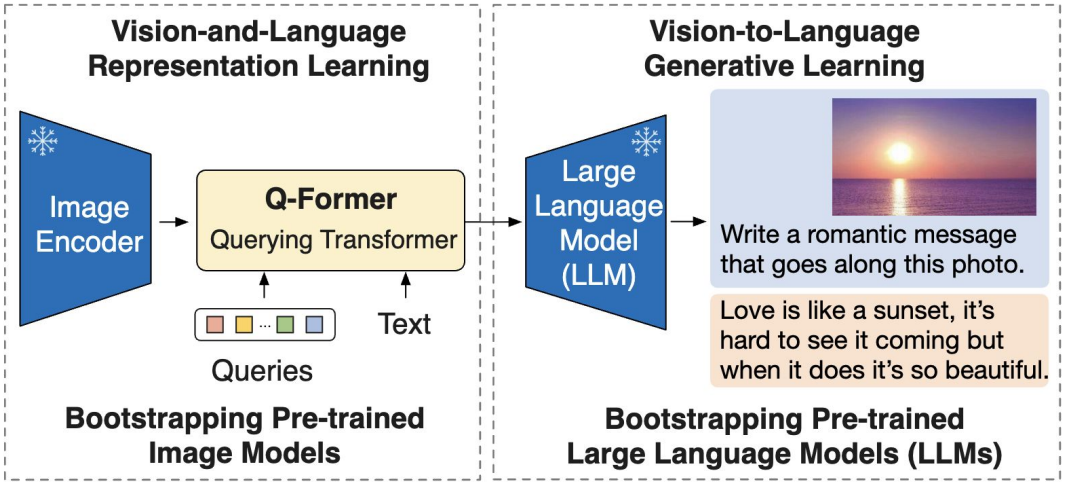


Zhao et al. "BOSSA: Learning Music Style Through Cross-Modal Bootstrapping" in NeurIPS 2025 AI4Music Workshop

BOSSA (BLIP-2)

BLIP-2 bootstraps vision-language pre-training from off-the-shelf frozen pretrained image encoders and frozen large language models.

BLIP-2 bridges the modality gap with a lightweight **Querying Transformer**, which is pretrained in two stages.

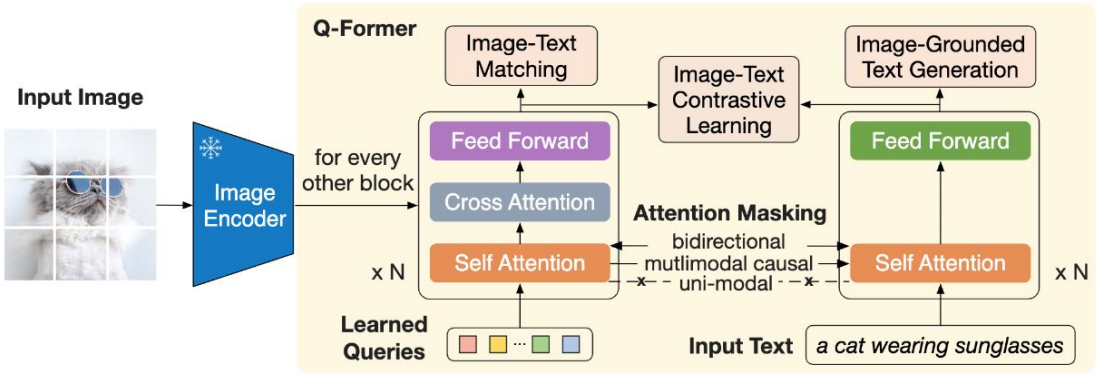


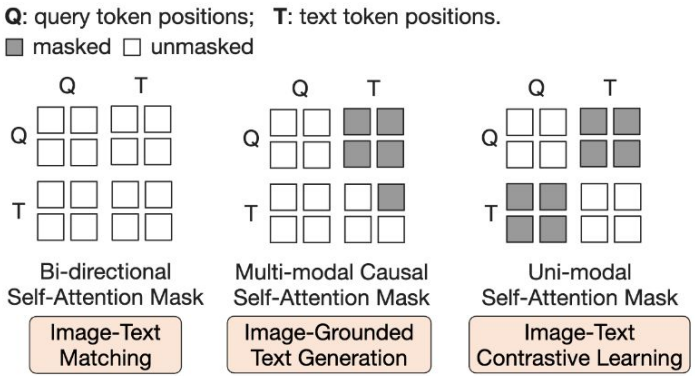
Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" in Proc. of ICML 2023

BOSSA (BLIP-2)

Optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text.

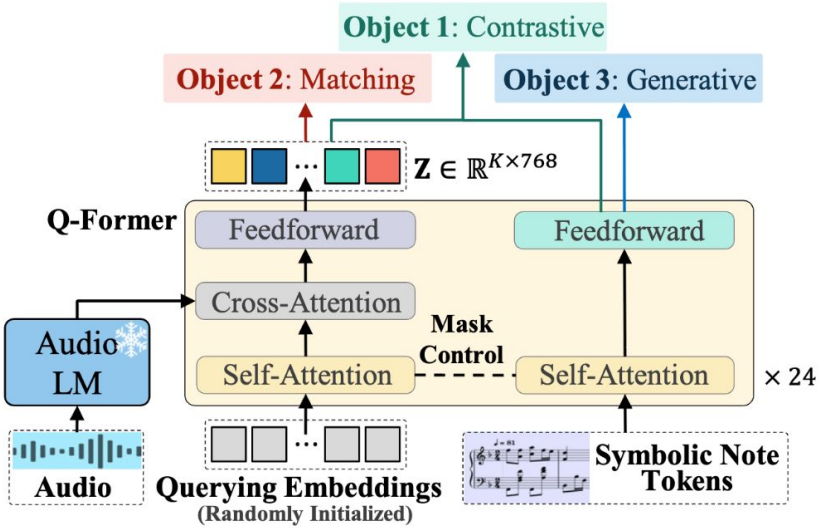
The self-attention masking strategy for each objective to control query-text interaction.



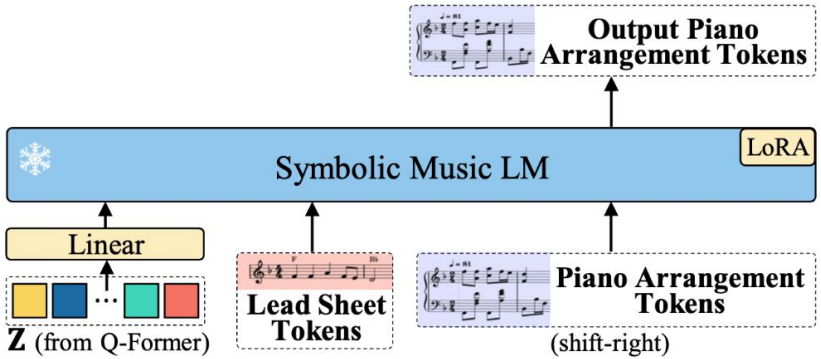


Li et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" in Proc. of ICML 2023

BOSSA



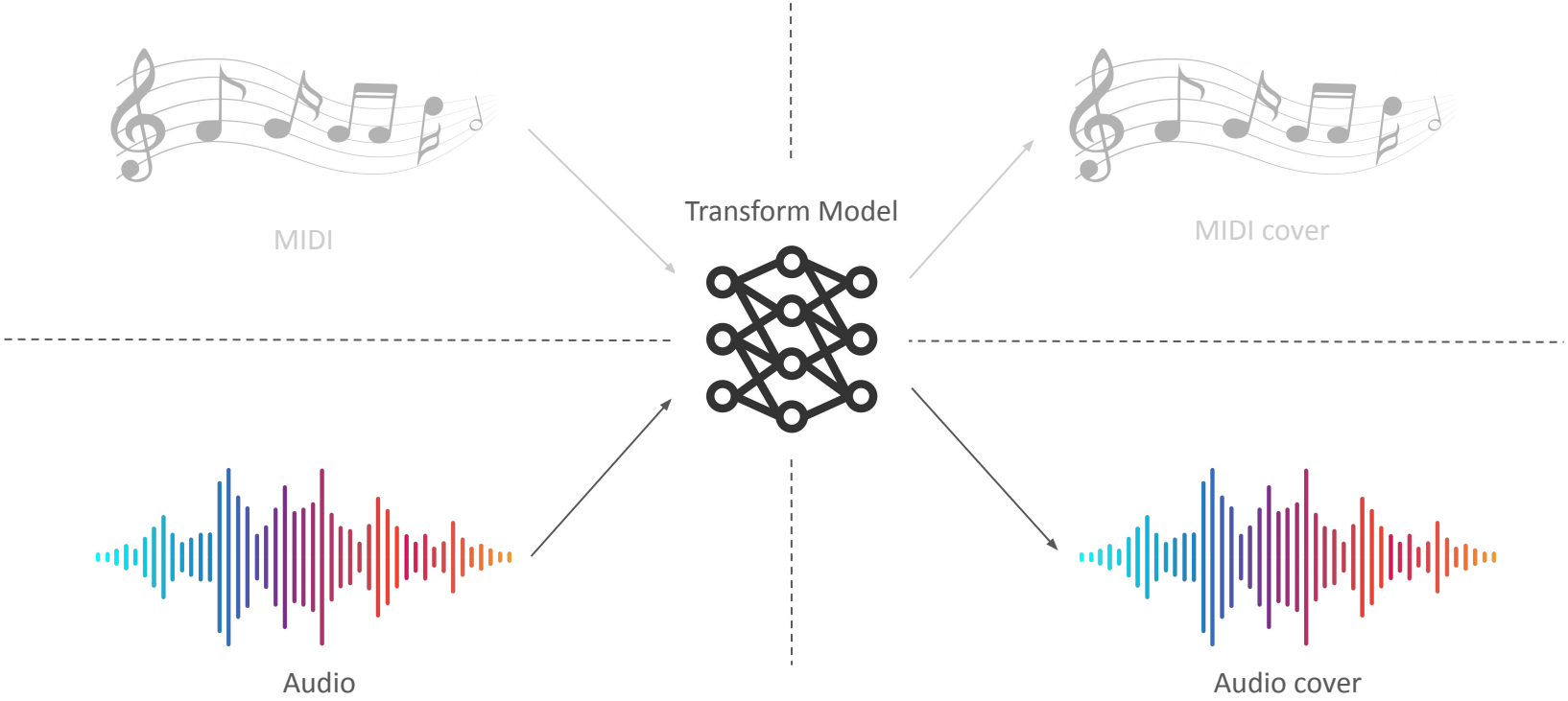
(a) Stage 1: Cross-modal learning with Q-Former.



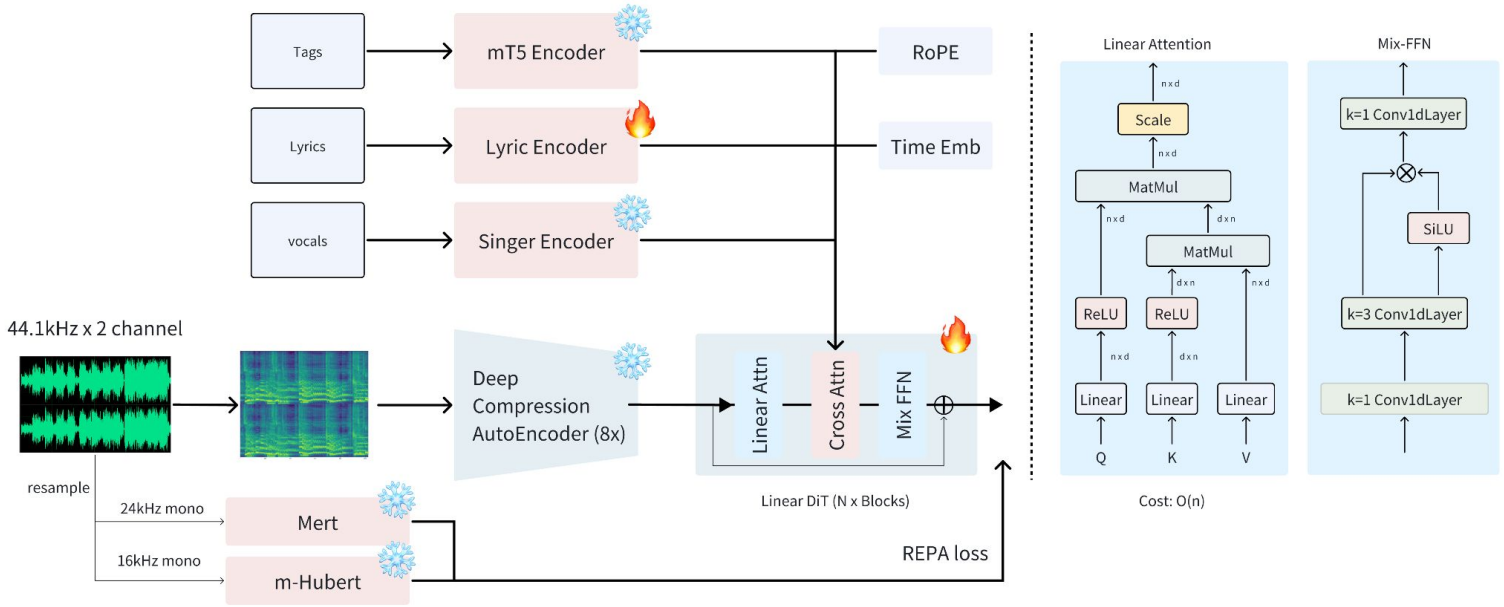
(b) Stage 2: Audio-to-symbolic arrangement.

Zhao et al. "BOSSA: Learning Music Style Through Cross-Modal Bootstrapping" in NeurIPS 2025 AI4Music Workshop

Audio to Audio

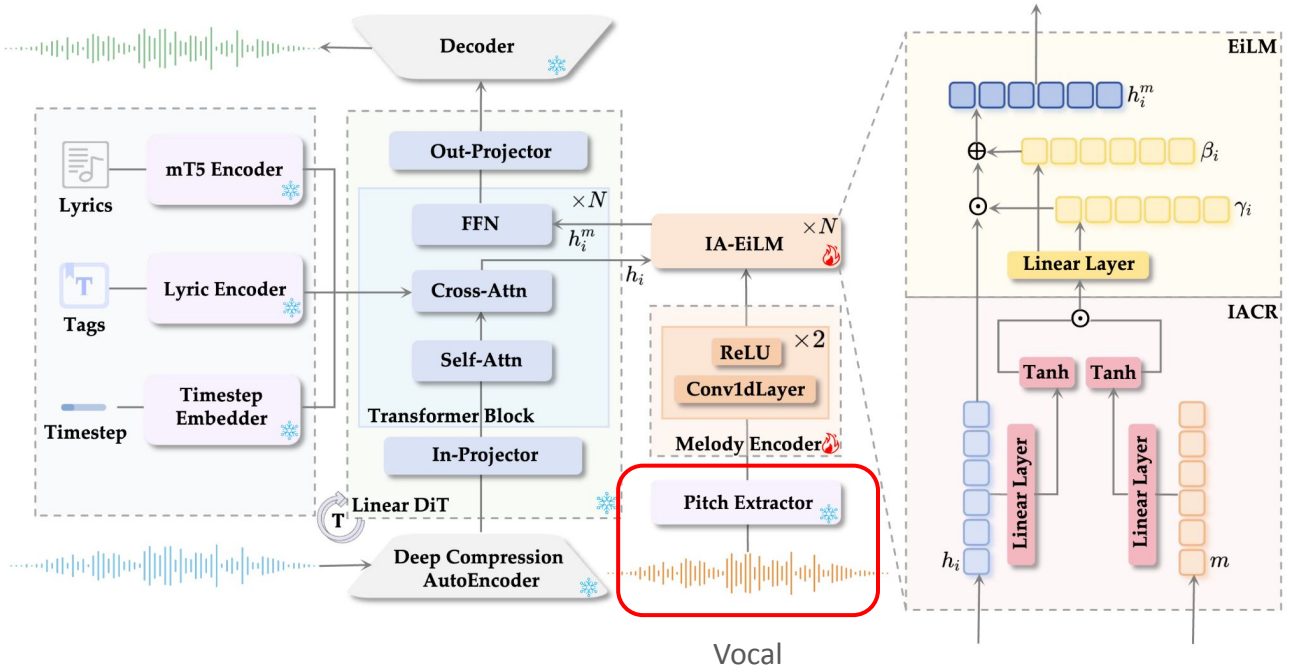


SongEcho (Ace-Step)



"SongEcho: Cover Song Generation via Instance-Adaptive Element-wise Linear Modulation" in Arxiv

SongEcho



"SongEcho: Cover Song Generation via Instance-Adaptive Element-wise Linear Modulation" in Arxiv

Evaluation - what a mess

What happened?

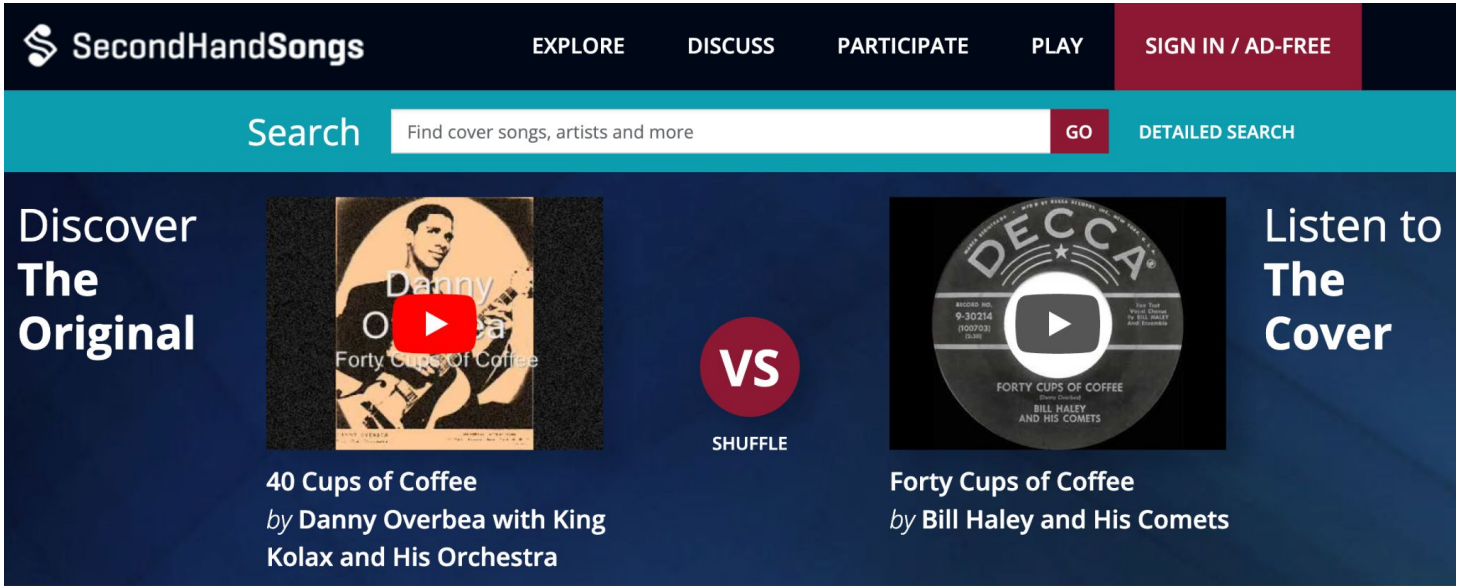
- The musical concept of "style" is not well-defined, leading to inconsistent model objectives.
- Evaluation metrics suffer from strong **personal biases**.
  - A high score on technical similarity is **not equivalent** to a good musical cover (a cover needs interpretation, not cloning).
  - Similarly, high overall quality does **not guarantee** a good cover (it needs to sound like the original and the new style).
- We currently rely on metrics borrowed from general music generation, which are fundamentally **unsuited for cover generation**.

Evaluation of cover song generation

There is still much room for improvement ……

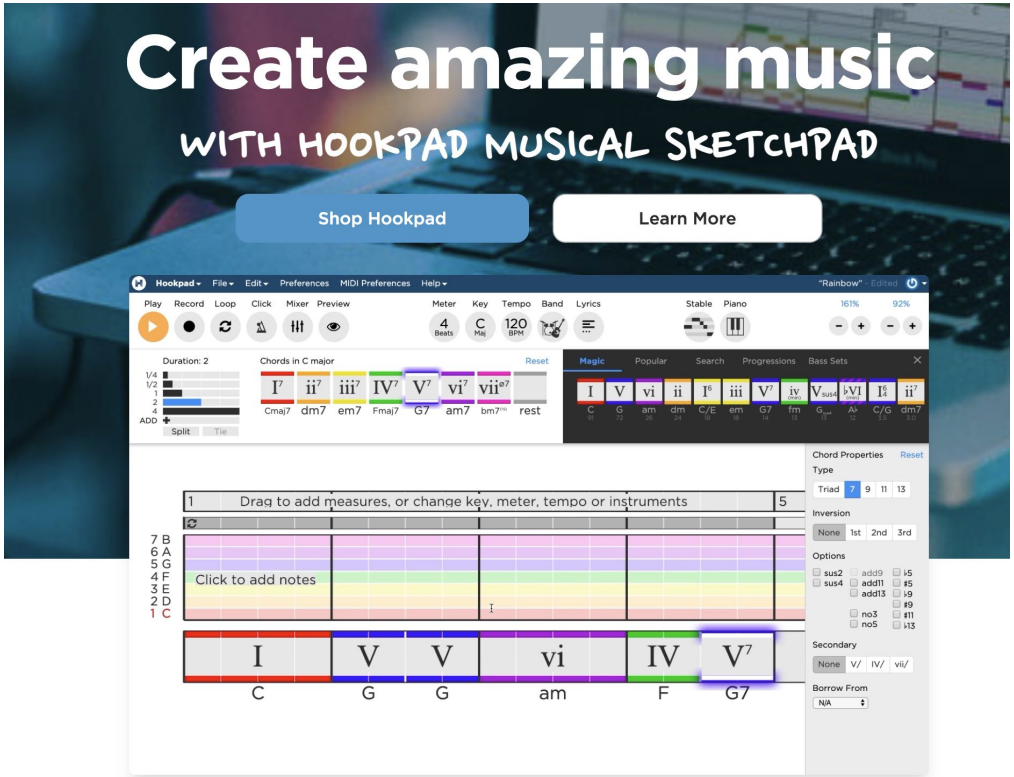
Resource

SecondHandSongs



https://secondhandsongs.com/

HookTheory



https://www.hooktheory.com/

Dataset

- POP909
- Live Orchestral Piano
- LargeSHS
- Discog-VI
- Pop2Piano

If you are interested in doing anything related to cover song generation, feel free to contact me