Melodic Machines: A Dual-Model Approach to Artist-Conditioned Music Generation

Man-Fang Liang Qien Lin Alzahra Fayie Zetao Wu

Background

Music creation presents a unique challenge at the intersection of art and science. Crafting authentic lyrics and reproducing an artist's unique sonic signature often demands extensive manual effort and deep domain expertise. Recent breakthroughs in deep generative modeling offer new tools for creative automation such as transformer-based text models and diffusion models. However, end-to-end, artist-conditioned music generation remains under-explored.

Our melodic machines introduces a two-stage pipeline that:

- A transformer-based lyric generator that produces novel lyrics in a target artist's voice and genre
- A stable diffusion model that synthesizes audio spectrograms mimicking an artist's sonic characteristics and converting back to waveform

By coupling text and audio generation, our framework aims to streamline the music creation workflow while preserving individual artistic identity.

Overview

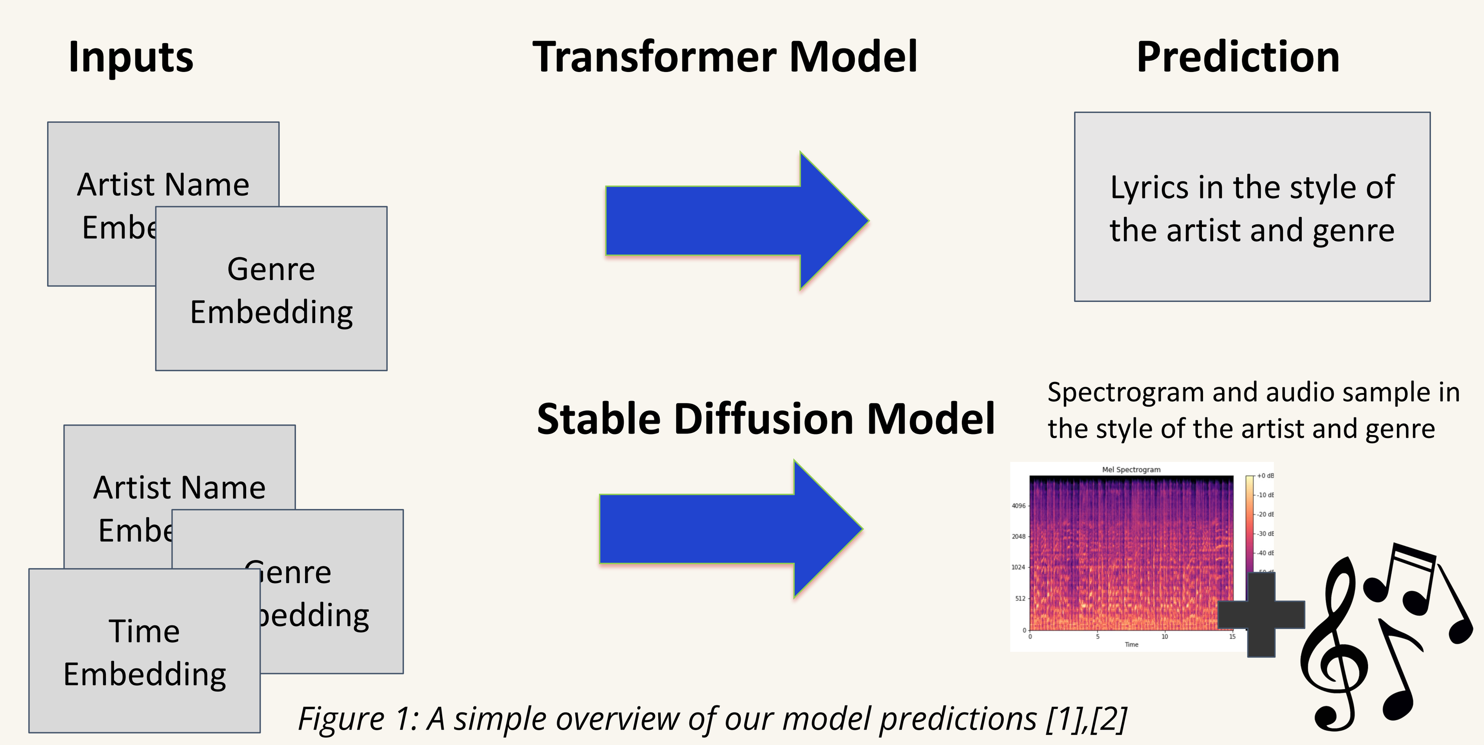

Figure 1: A simple overview of our model predictions [1],[2]

Motivation

- Reduce barriers to professional-quality songwriting and streamline production workflows
  - Manual lyric writing and audio production demand high domain expertise, time, and resources.
  - An end-to-end AI assistant can help artists prototype ideas faster and at a lower cost.
- Democratize creative workflowsGenre Embedding
  - Make artistic production more accessible to those without extensive musical training.
  - Enable rapid exploration of styles, genres, and lyrical concepts.
- Explore AI creativity
  - Investigate how deep learning can capture the essence of an artist's style across multiple modalities.
- Technical innovation
  - Combine transformer-based text generation with diffusion-based audio synthesis and present the unique technical challenges and opportunities for advancement in multimodal AI systems.

Data

Given our two-model architecture, the data used for model training had to undergo significant preprocessing.

- For our stable diffusion model, we used the FreeMusicArchive (FMA) Audio Dataset and converted the .mp3 files into spectrogram images associated to an artist-id and genre, which became the conditional embeddings for our model.
- For our transformer model, we used a famous artists' song dataset manually inputted by a Kaggle user. The data had human-error so we cleaned the data, refined it, and split it into training and testing files.

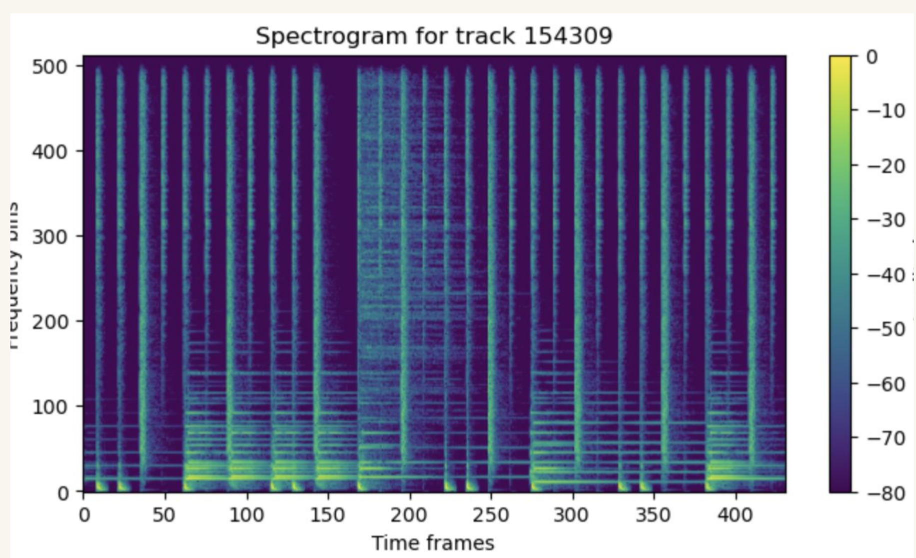

Figure 2: A sample spectrogram we converted from the FreeMusicArchive (FMA) Audio dataset.

Model Architecture

Transformer Model Architecture

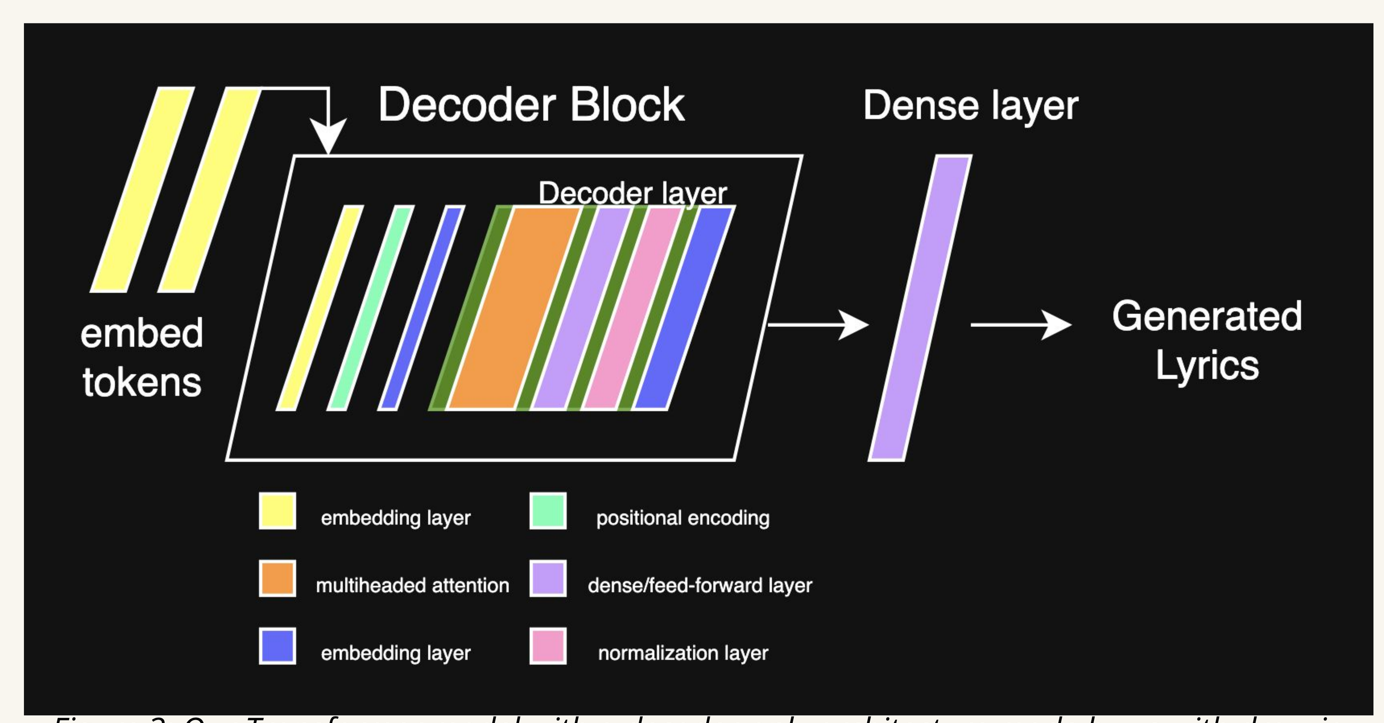

Figure 3: Our Transformer model with a decoder-only architecture made by us with draw.io

Stable Diffusion Model Architecture

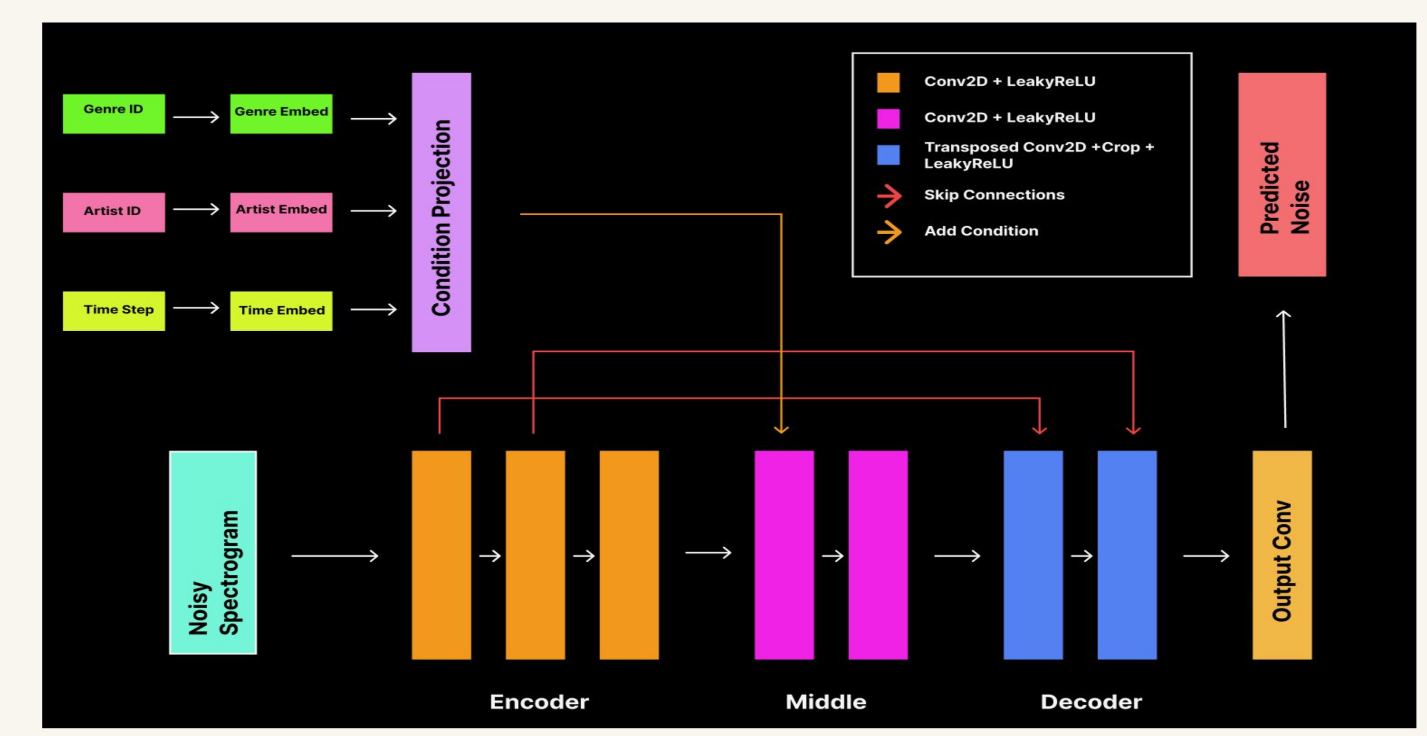

Figure 4: Our Stable Diffusion model architecture made by us with Figma

Results

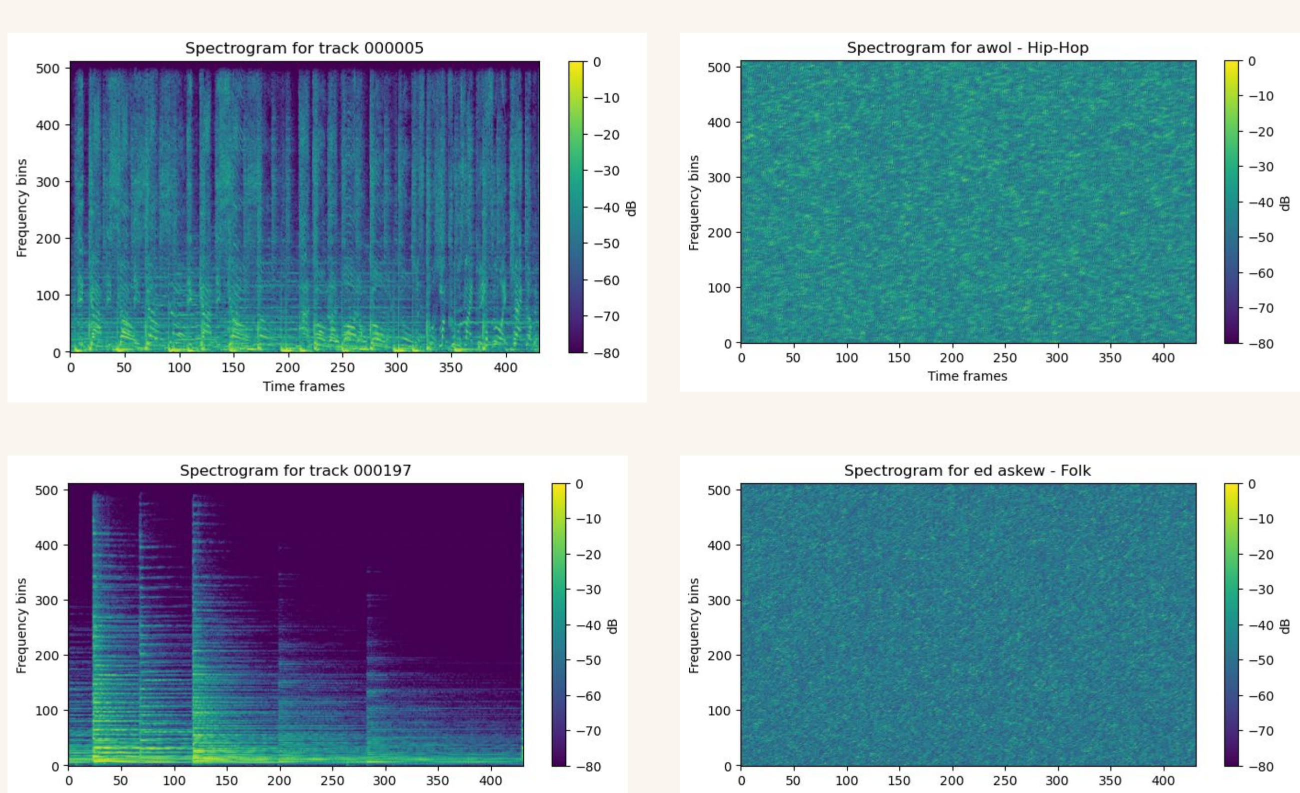

Figure 4: A set of original audio spectrograms and our model's generated spectrograms based on the same artist name and genre

```
Enter artist name (e.g., 'dua lipa'): dua lipa
Enter genre (e.g., 'pop'): pop
Enter optional starting text (or press Enter to skip):
Enter temperature (0.5-1.5, higher = more random): 1

Generating lyrics in the style of dua lipa (pop)...

==================================================
ipod galaxy daybed qualities static lalala rob figures size
lead center crystal lionel accusations public subway ever six
bit control prize pick choose advance twodoor tickin' seasoning
'96 waitress actin' miss well settling barefoot high incline
package belonged badder stripper saturn toptop slacks fools again
hardest travelled here's brrt hear world dodged poolside scratch
bake kirkpatrick bored invitations flippity moment ooooh stereo plant
must've mormons touring daylight lack 's heheadshot change ironin'
guts jingle necessarily toma nowi'm twirl comments distorted pick
these rapper follie's not've nyu shits pigeons kinda split
satisfactory rosaleem nonlyrical gettin' my shock kush spot origami
xo massoccur woah minime filthy lies ohohohohohohoh collins turnin'
economic mercy hahahaha masterpièce curl outlaw dogg headtop swallow
revolutionaries sleazy lakim detractors marvin leaving genes hattie cure
halleluhallehallelujah jobs christ wanted heyhey act here himeros bullets
room's monsters twitter accounts cannot youknowwho elevators donjae wet
sign powder applause spoken fixed buy
==================================================
```

Figure 5: A glimpse into our interactive terminal where you can provide an artist name and genre to receive a lyric generation

Discussion & Ethics

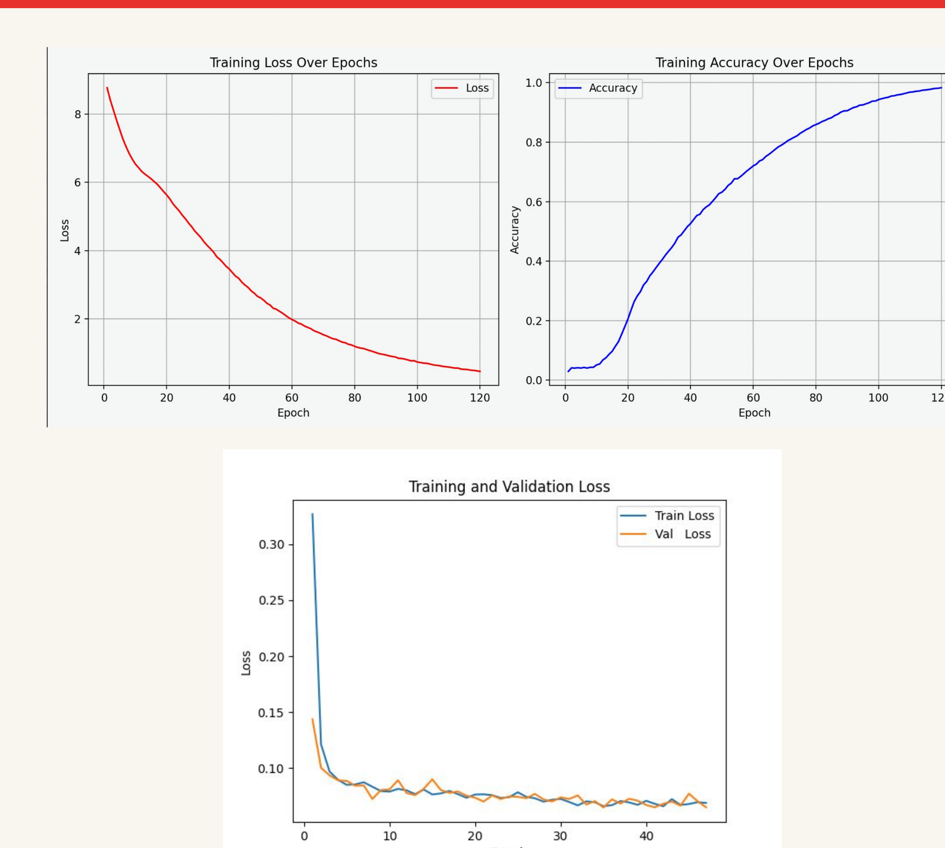

Figure 6: First row: Accuracy and loss metrics for our transformer model. Second row: loss metrics for our stable diffusion model

Technical limitations: Despite achieving promising training metrics (loss and accuracy) after 100 epochs, our lyric generator still produces outputs that lack coherence and artist-specific style. This shows a challenge that small training datasets will severely limit the model's ability to capture nuanced writing patterns and we need more training data. As for our stable diffusion model, our model struggles with denoising, however we do notice a learning pattern when playing their respective audio samples.

Ethical Considerations:

- Artist attribution and consent: Our approach raises questions about creating content that mimics specific artists without explicit permission.
- Creativity authenticity: AI-generated music blurs the line between human and machine creativity. It might devalue human artistic labor potentially.
- Misrepresentation risks: Potential reputational damage if generated content is offensive or low-quality. It might misrepresent an artist's style or generate inappropriate content.

References

[1] Roberts, L. (2024, January 17). Understanding the MEL Spectrogram - Analytics Vidhya - Medium. *Medium.* https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[2] https://pngimg.com/image/22011

Acknowledgements

We would like to thank Professor Eric Ewing for his instruction and guidance throughout the course. We are also very grateful to the TA team and our mentor, Navya Sahay, for her support and suggestion on using stable diffusion. Special thanks to TAs Armaan Patankar and David Lubowski for their advice and dedication :)