Melody Diffuser: Interactive Symbolic Music Generation via Discrete Diffusion and Self-Supervised Gesture Control

Duncan Larzelere
East Lansing High School
duncan.larzelere@gmail.com

Abstract

Text-to-music generation models have significantly advanced through Diffusion-Transformer architectures [1, 2]. However, these systems primarily rely on text-based conditioning, accurately generating stylistic elements such as genre, yet enabling very little control over melodic and rhythmic structure.

I developed **Melody Diffuser**, a controllable AI music generation model built to enable human-AI collaboration through gesture-based input. Melody Diffuser was trained using classifier-free guidance on synthetic tokenized hand gesture sequences. I addressed the lack of existing data by creating a custom self-supervised pipeline. The *4 Bars Monophonic Melodies Dataset*, containing over 10 million symbolic monophonic melodies, was processed through a rule-based interval analysis to generate conditioning sequences. This approach converted pitch intervals into 8 discrete gesture tokens (small up, medium up, large up, small down, medium down, large down, hold, repeat).

I trained a 32M-parameter Discrete Diffusion Transformer to iteratively reconstruct melodies at any time point in a categorical token corruption process. Tokens are randomly replaced according to a 64-step schedule. The model uses classifier-free guidance (CFG) to enable conditioning sequences to be fed to the model through cross-attention mechanisms in each transformer block. Based on prior success [3,4], I integrated **RMSNorm** and **SwiGLU** activations to stabilize training.

At inference-time, index finger coordinates are tracked using MediaPipe [5], a computer vision model, and tokenized through a rule-based system into 8 discrete gesture tokens aligning with the synthetic dataset. I deployed Melody Diffuser to a browser-based composition system where live webcam footage conditions the diffusion process on a cloud GPU (NVIDIA T4), enabling intuitive human-centered melodic control.

The demo is available at dl8ai.com. Note that there is approximately a three-minute delay to wake the GPU cluster upon the first generation due to free-tier hosting constraints. The model successfully generates melodies that align with the video-captured gesture sequences, creating a full human-AI collaboration pipeline. I am currently working to expand the model to polyphonic music generation by training a variational auto-encoder on Bach chorales, which will allow latent diffusion to generate 4-part harmony.

References

[1] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," *ICCV*, 2023.

[2] J. Engel et al., "Long-form music generation with latent diffusion," *arXiv:2404.10301*, 2024.

[3] B. Zhang and R. Sennrich, "Root Mean Square Layer Normalization," *NeurIPS*, 2019.

[4] N. Shazeer, "GLU Variants Improve Transformer," *arXiv:2002.05202*, 2020.

[5] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," *arXiv:1906.08172*, 2019.