Text-to-Music Generation with Emotional Nuance: Multi-dimensional Emotion Space Modeling

1 Motivation & Significance

Recent advances in text-to-music generation, particularly Transformer and Diffusion models, have achieved remarkable success in audio fidelity and stylistic control[2]. Models like MusicGen and MusicLM represent the current state-of-the-art[1]. However, these models struggle to interpret complex or nuanced emotional descriptions in text[3]. For example, generating a "sad but hopeful piano piece" versus a "purely tragic piano piece" often results in musically similar outputs, with little distinction in emotional nuance.

Russell's Circumplex Model of Affect (VA Model) is a psychologically-grounded framework that transforms subjective emotional experiences into a continuous and quantifiable two-dimensional space, in which **Valence** describes the pleasantness of the emotion (positive vs. negative) and **Arousal** indicates its intensity (calm vs. excited)[4].

Research Hypothesis: Incorporating VA-based emotion conditioning into MusicGen model will significantly improve its ability to generate musically distinct outputs for nuanced emotional descriptions, compared to text-only conditioning.

2 Approach and Methodology

Building on empirical evidence from symbolic music generation, we propose enhancing MusicGen with continuous-concatenated VA conditioning through the following methodology:

Step 1 Dataset (2-4 weeks): Standardize existing emotional music datasets (MusicCaps, DEAM, EMOPIA, PMEmo, Lakh-Spotify, MuSe) into a consistent VA framework.

Step 2 Emotion Conditioning (1-2 months): *Encoder:* Map VA vectors into MusicGen's text embedding space (768 dimensions).
Cross-Attention: Concatenate text and emotion embeddings for joint conditioning.

Step 3 Fine-tuning (2 months): Apply Low-Rank Adaptation (LoRA)-based parameter-efficient fine-tuning on paired text–VA data.

Step 4 Evaluation (1 month): *Interactive Demo:* build a prototype that allows users to generate music by inputting both text and emotional cues.
Human-in-the-Loop Evaluation: including A/B Testing – users compare clips from different generative models and decide which one expresses better on a specific emotion; and VA Value Ratings – comparing the score of perceived emotion and target values on a VA scale.

Step 5 Iteration (1 month): Refine using evaluation feedback.

Tools & Resources: PyTorch, HuggingFace Transformers (MusicGen), Gradio (demo), LoRA, ISP computing resources, Colab, Kaggle.

3 Preliminary Results

Figure 1 and Table 1 present Sulun et al.'s comparative evaluation of emotion conditioning approaches[5].

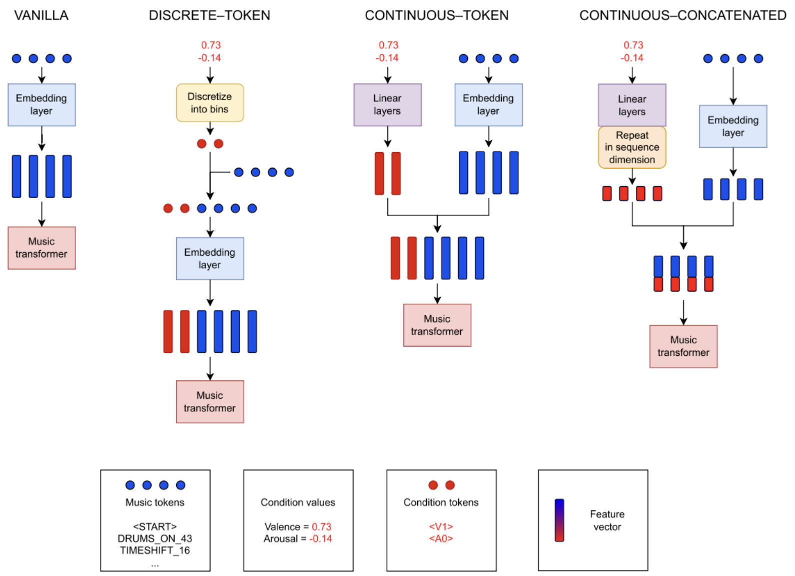

Figure 1: Emotion conditioning approaches[5]

Table 1: Performance of the models during evaluation. NLL refers to negative log-likelihood, where lower is better. Top-1 and Top-5 refer to the accuracy, where higher is better.[5]

Model	NLL	Top-1	Top-5
Vanilla	0.7445	0.7784	0.9513
Discrete-token	0.7375	0.7885	0.9536
Continuous-token	0.7122	0.7895	0.9545
Continuous-concatenated	**0.7075**	**0.7913**	**0.9548**

The continuous-concatenated method achieves optimal performance across all metrics (NLL: 0.7075, Top-1: 0.7913), providing empirical foundation for our approach.

4 Expected Impact & Proposed Work

Contributions: A VA-conditioned text-to-music framework enabling finer emotional control.

Deliverables: Micropublication, open-source code, interactive demo.

Impact: Advances emotion-aware AI music generation for personalized media, therapy, and creative tools.

References

[1] Ma et al. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*, 2024.

[2] R. Mitra and I. Zualkernan. Music generation using deep learning and generative ai: A systematic review. *IEEE Access*, 13:18079–18106, 2025.

[3] J. Mycka and J. Mańdziuk. Artificial intelligence in music: recent trends and challenges. *Neural Comput & Applic*, 37(2):801–839, Jan. 2025.

[4] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[5] S. Sulun et al. Symbolic music generation conditioned on continuous-valued emotions. *IEEE Access*, 10:44617–44626, 2022.