



How Could We Add Emotional Nuances to AI-Generated Music?

Moaksh Kakkar

mypublishedpaper@gmail.com

Bennett University, Uttar Pradesh

ABSTRACT

Artificial intelligence (AI) has come a long way in recent years in generating music through architectures like RNNs, Transformers, GANs, VAEs, diffusion models, and large language models. Although these models are capable of generating structurally coherent and stylistically accurate music, they tend to lack the subtle emotional nuance and depth of human music. This paper examines the idea of emotional nuance—the ability of AI-generated music to express subtle variations, mixed affects, changing affective trajectories, and selective emotional impact. Combining theories from music psychology, affective computing, and computational creativity, I translate musical features like tempo, mode, harmony, dynamics, articulation, timbre, and melodic contour into their perceived emotional counterparts. I survey and compare methods of emotional control, ranging from conditional generation and reinforcement learning with affective rewards to employing music theory and hybrid symbolic–neural methods. Key challenges are presented, such as the subjective nature of emotional perception, dataset limitations, cultural variability, and the challenge of quantifying nuanced affect. I also outline directions for future work around more robust datasets, culturally adaptive models, cognitively inspired emotion representations, interpretable control mechanisms, and sound evaluation frameworks. By refining these strategies, AI music systems can move closer to being not just pattern generators, but creative collaborators able to express genuine emotion.

KEYWORDS: AI music generation, emotional nuance, affective computing, music psychology, valence-arousal model, reinforcement learning, large language models, generative models, music theory integration, computational creativity.

INTRODUCTION

Music is often described as "universal language" [1], transcending cultural and temporal boundaries to express the vast spectrum of human emotions and creativity [1]. Its significance lies in its aesthetic appeal and its powerful influence on listeners' cognitive and emotional states [2]. Music can evoke joy, sorrow, tension, and peace, often with remarkable consistency among listeners. This capacity to communicate and induce emotion underpins many societal roles, from entertainment and social bonding to therapeutic applications like mood regulation and stress reduction [1]. The emotional impact of music is deeply intertwined with its structure and performance, making emotional expressivity a core component of musical artistry [3].

In recent years, artificial intelligence (AI) has witnessed remarkable advancements in music generation [4]. Fueled by developments in deep learning, AI models can now compose and synthesize music with increasing complexity and stylistic fidelity [5]. Techniques based on Recurrent Neural Networks (RNNs), Transformers, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models have demonstrated capabilities ranging from melody generation and harmonization to full-track audio synthesis [1]. However, despite these technical achievements, a significant gap often remains between the music generated by AI and the emotionally resonant music created by humans. Current AI systems frequently struggle to imbue their creations with genuine emotional depth, nuance, expressiveness, and warmth [6]. While models might generate technically proficient or stylistically coherent pieces, they often lack the subtle emotional variations and convincing affective arcs that characterize human musical expression, sometimes resulting in music perceived as emotionally flat or generic.

Addressing the "emotional gap" requires moving beyond simplistic representations of emotion. "Emotional nuance" in this context refers to the ability to generate music that expresses more than just essential, discrete emotional categories (e.g., happy, sad, angry) or broad positions within a dimensional space (e.g., high arousal, negative valence). It encompasses the capacity to convey:

Subtle Variations: Differentiating between closely related emotions (e.g., contentment vs. joy, pensiveness vs. sadness).

Mixed Emotions: Expressing complex states involving multiple, potentially conflicting, feelings simultaneously (e.g., bittersweetness, nostalgia) [7].

Evolving Emotional Arcs: Creating music where the emotional tone shifts and develops over time in a coherent and meaningful way, including building tension and providing release [3].

Specific Effects: Targeting distinct emotional experiences like nostalgia, awe, suspense, or solemnity, which have unique phenomenological qualities and musical associations.

This contrasts sharply with many current AI approaches, often limited to generating music based on predefined, broad categories or quadrants within the valence-arousal (VA) circumplex model [6]. Achieving proper emotional nuance requires AI systems to understand and manipulate musical elements with greater sophistication and sensitivity.

FOUNDATIONS FOR EMOTION-AWARE MUSIC GENERATION

Mapping Musical Features to Emotional Perception

A significant body of research in music psychology and MIR has established correlations between specific, quantifiable musical features and the emotions perceived by listeners, particularly within the context of Western tonal music [8]. These mappings provide a crucial starting point for controlling emotion in AI music generation. Key features include:

Tempo (Note Density): Often considered one of the most influential cues [8], faster tempos (higher note density) are strongly associated with high arousal emotions like excitement, joy, or anger, while slower tempos (lower note density) correlate with low arousal states such as sadness, calmness, or peace [8].

Mode (Pitch Histogram/Tonality): The distinction between major and minor modes is a powerful indicator of valence [8]. Primary keys are predominantly linked with positive valence (happiness, tenderness), whereas minor keys are associated with negative valence (sadness, anger) [8]. Pitch histograms can quantify the prevalence of scale degrees characteristic of major or minor modes [8].

Harmony: The degree of consonance (harmonious intervals) versus dissonance (clashing intervals) significantly impacts perceived tension and valence, although context plays a crucial role. Harmonic complexity and the specific progression of chords also contribute to the emotional narrative [9].

Dynamics (Loudness): Loudness levels generally correlate with energy and intensity. Increased loudness can suggest power, anger, or joy, while softer dynamics are often linked to tenderness, sadness, or fear [10]. Variations in loudness (dynamic range and variability) also contribute to expressiveness [10].

Articulation: How notes are played—staccato (short, detached) versus legato (smooth, connected)—influences perceived energy, agitation, or smoothness [10].

Timbre: The unique sound quality of instruments or voices (spectral characteristics) plays a significant role in emotional coloring, potentially influencing valence more subtly than mode or tempo [9]. Different instruments carry different emotional connotations [9].

Melodic Contour/Pitch Height: The overall shape of a melody and its average pitch level can contribute to emotional expression. Ascending contours or higher pitch registers might be associated with increasing excitement or happiness, while descending contours or lower registers may suggest sadness or seriousness.

Rhythm: Rhythmic features, such as complexity, regularity, and syncopation, impact perceived stability, energy, and arousal.

While these established feature-emotion correlations offer a vital foundation, particularly for rule-based systems [10] and feature-driven approaches like EmotionBox [8], they represent a simplified view of a complex reality. The emotional impact of music arises not just from individual features in isolation, but from their intricate interactions, the surrounding musical context, the listener's cultural background, personal experiences, and expectations [8]. Relying solely on these basic, often generalized mappings limits the potential for AI to generate music with genuine emotional depth and nuance. More sophisticated models are needed that can learn these complex interactions and contextual dependencies directly from data or through more advanced modeling paradigms.

Neural Architectures for Affective Expression

Modern AI music systems, such as transformer-based models like MusicLM and MuseNet, primarily focus on structural coherence and genre fidelity [11]. However, their latent spaces often disregard explicit emotional descriptors. Emerging approaches integrate emotion embeddings directly into neural architectures, enabling fine-grained control over practical outputs [12]. For instance, hybrid variational autoencoder (VAE) frameworks now incorporate valence-arousal dimensions as conditional inputs, allowing users to navigate emotional trajectories through latent space interpolations [13]. These systems leverage datasets like EMOPIA, which tags MIDI compositions with crowd-sourced emotional labels, to establish correlations between musical features and perceived affect [14].

The challenge lies in decoding emotion-specific features from the broader stylistic patterns. Techniques like attention masking in transformer models have shown promise in isolating emotionally salient components, such as melodic contour or harmonic tension [7]. Models can amplify these features during generation by applying differential attention weights to pitch intervals and rhythmic patterns associated with specific emotions [15]. For instance, generating “joyful” melodies might prioritize ascending fourth intervals and syncopated rhythms based on their strong correlation with positive valence in perceptual studies [6].

Reinforcement Learning with Human Feedback

Purely supervised learning approaches face limitations due to the subjectivity of emotional perception. Reinforcement learning (RL) frameworks incorporating real-time human feedback create adaptive loops for emotional calibration [16]. Human evaluators rate generated samples across emotional dimensions in these systems, creating reward signals that update the generator's policy. The Polyphonic Emotion RL Environment (PERE) demonstrated how this approach can achieve greater emotional recognition accuracy than static models.

Crucially, these systems must account for cultural and individual variability in emotional responses. Bayesian optimization techniques enable personalization by modeling user-specific emotional associations with musical features [17].

For example, Western listeners associate minor keys with sadness, but this correlation drops in the East Asian population, necessitating culturally adaptive reward functions [18].

Psychological Models of Musical Emotion

The circumplex model of affect positions emotions along valence (positive-negative) and arousal (calm-excited) axes and provides a robust framework for computational implementation [19]. AI systems can map acoustic features to these dimensions: tempo and spectral flux strongly predict arousal, while mode and harmony complexity correlate with valence [20]. Deep neural networks trained on psychophysiology data, such as galvanic skin response during music listening, can generate music that elicits target arousal levels [21].

Critically, dimensional models enable continuous emotional trajectories rather than static labels; recurrent architectures with emotional memory cells can maintain coherence across affective shifts, mimicking the narrative arc of human-composed pieces [22]. This capability proves essential for applications like film scoring, where music must synchronize with dynamic emotional narratives.

While dimensional models offer granularity, categorical emotions (e.g., joy, anger, nostalgia) provide intuitive control parameters. Transfer learning from music emotion recognition (MER) systems allows generative models to inherit robust emotion classifiers [23]. The EmoGen system uses a pre-trained MER CNN to compute against gradient signals that steer the generation toward target emotion clusters, achieving 41% higher categorical accuracy than baseline models [24].

Cross-modal training enhances categorical specificity. Systems trained on paired music-lyric datasets learn nuanced associations between lyrical themes and musical devices, for instance, linking metaphors of “flower” with lower registers and slower tempos [25]. This approach mirrors human composer practices where textual and musical elements collaboratively convey emotion [26].

Data Acquisition and Curation Strategies

The EMOPIA dataset exemplifies crowd-sourced emotion labeling, capturing 10000 MIDI clips annotated with categorical and valence-arousal ratings [27]. However, label noise remains a challenge. Inter-annotator agreement peaks at 68% for broad categories but drops to 49% for nuanced states like “bittersweet”. Semi-supervised techniques that combine sparse human labels with pseudo-labels from acoustic models can mitigate this issue, expanding usable training data [28].

Active learning frameworks prioritize annotation efforts on musically ambiguous samples where prediction confidence falls below the threshold [29]. This strategy reduces annotation costs while maintaining model performance, crucial for scaling emotional datasets.

ARCHITECTURES FOR AI MUSIC GENERATION AND EMOTION

Foundational Generative Models

Recurrent Neural Networks (RNNs) & Long Short-Term Memory (LSTMs): As early innovators in sequence modeling, RNNs and their extension, LSTMs, were two of the earliest deep learning approaches used in music generation [8]. Their capacity for sequential processing makes them well adapted to modeling musical structure through time. They are good at capturing local dependencies but tend to lose coherence when dealing with longer sequences [5]. Early efforts to generate conditional emotions employed such architectures, commonly by providing emotion labels as side inputs [8].

Transformers: This structure has become prevalent in sequence modeling tasks, such as music generation, thanks mainly to its self-attention mechanism [5]. Attention enables the model to assign the relative importance of various elements within the input sequence when producing the output, allowing it to extract long-range dependencies better than RNNs [30]. Transformers have now become standard in state-of-the-art systems for conditional music generation, such as those guided by emotional prompts, attributes, or cross-modal inputs [5].

Generative Adversarial Networks (GANs): GANs are composed of two networks, a generator and a discriminator, which are trained in competition [6]. The generator produces fake data (e.g., music), and the discriminator attempts to identify actual versus generated data. This process of adversarially iterating between the two networks can create very realistic outputs [6]. The applications of GANs have been used in audio and symbolic music generation [1], occasionally in conditional configurations to regulate features such as style or emotion [6]. But GAN training is unstable.

Variational Autoencoders (VAEs): VAEs learn a probabilistic compressed latent representation of the input data. Through sampling from or transforming this latent space, VAEs can generate new data. This architecture is especially beneficial for controllable generation, since various dimensions of the latent space can map to different data attributes [1]. Conditional VAEs (CVAEs) directly involve conditioning variables (such as emotion labels) in the encoding and decoding. Extensions such as Gaussian Mixture VAEs (GMVAEs) try to organize the latent space into clusters mapping to various categories or attributes, enabling disentanglement and control [1].

Diffusion Models: Diffusion models have recently become state-of-the-art for producing high-fidelity data, particularly images and audio [1]. They operate by progressively adding noise to data while training and then learning to invert this process, beginning from noise to produce clean data. Diffusion models form the basis of several top-performing text-to-music audio generation systems [1] and hold promise for high-quality conditional generation. However, their control mechanisms remain an active area of research.

The Role of Large Language Models (LLMs) in Text-to-Music and Control

Large Language Models (LLMs) now dominate the field, especially in text-to-music, as they utilize their strong natural language comprehension abilities. Their use follows many trends:

Classic Models: Use traditional sequence models (RNNs, Transformers, etc.) without LLMs, usually with more straightforward conditioning methods [1].

Hybrid LLM-Augmented Frameworks: Use a large language model (LLM) as a sophisticated “front-end” to examine text prompts.

The LLM may detect musical features like tempo, genre, and mood, generate lyrics, or paraphrase descriptions, then pass these as structured input to a standalone music generation model (e.g., a Transformer or diffusion model) [1].

End-to-End LLM-Centric Models: Think of music generation as a task similar to language modeling, where musical events (e.g., notes, timing, and audio tokens) are represented as tokens identical to words. The large language model (LLM) processes the text prompt as is and produces the corresponding stream of music tokens [1]. Examples include MusicLM, MusicGen, and SongComposer.

LLMs perform better on text understanding of implicit descriptions, for instance, of emotion, style, and instrumentation, and consequently improve controllability and expressiveness in text-to-music systems. Yet, regardless of LLMs, the challenges of scarce training data, token representation shortages of intricate music structures, and long-term coherence preservation remain present [1].

Symbolic vs. Audio Generation: Implications for Emotional Control

AI music generation operates in two primary domains: symbolic and audio, each with distinct implications for emotional control [1].

Symbolic Generation: Creates music represented as discrete events, typically in formats like MIDI or musical scores. This representation includes information about pitch, duration, onset time, and velocity (loudness) [1].

- i. **Pros for Emotion:** Allows for direct manipulation of musical elements (e.g., changing mode, adjusting tempo, altering note density) that have established links to music theory and psychological emotion perception [1]. This facilitates interpretable control and the integration of music theory knowledge. Emotional control often focuses on modifying these structural parameters [8].
- ii. **Cons for Emotion:** Lacks the richness of timbre, articulation nuances, and expressive performance dynamics (micro-timing, subtle loudness variations) that are crucial components of emotional expression in performed music [6]. The generated output can sound mechanical or lack expressive depth.

Audio Generation: Directly synthesizes the sound waveform or its spectral representation (like a spectrogram) [1].

- i. **Pros for Emotion:** Captures the full sonic richness of music, including timbre, performance nuances, and acoustic effects, which significantly contribute to the perceived emotional quality and realism [6].
- ii. **Cons for Emotion:** Offers less direct control over specific, theory-based musical elements like harmony or mode. Control is often achieved indirectly by conditioning the entire generation process on high-level descriptions (e.g., text prompts) or global parameters. Ensuring long-term structural coherence can also be more challenging than symbolic methods [1].

Hybrid approaches combine both strengths, using symbolic representations for planning and structure while employing audio synthesis for expressive rendering [4].

Table 1: Comparison of AI Music Generation Architectures for Emotional Control

Architecture Type	Key Mechanism	Typical Representation	Strengths for Music/Emotion	Weaknesses/Challenges for Music/Emotion	Example Systems/Refs.
RNN / LSTM	Recurrence, Sequential Processing	Symbolic, Audio	Sequence modeling, Local dependencies	Long-range dependencies, Vanishing gradients	EmotionBox [8], Ferreira et al [8], LHVAE [31]
Transformer	Self-Attention	Symbolic, Audio	Long-range dependencies, Parallel processing, Contextual understanding	Computationally intensive, Data hungry	Music Transformer[30], EmoMusicTV[31], XMUSIC[30]
GAN	Adversarial Training	Symbolic, Audio	Realistic output, Implicit distribution learning	Training instability (mode collapse), Difficult evaluation	MuseGAN [32], C-RNN-GAN [32], MeloHarmony [7]
VAE / CVAE	Latent Space Learning, Inference	Symbolic, Audio	Controllable generation, Disentanglement, Data compression	Blurry outputs (sometimes), Latent space interpretation	Emotion-CVAE [33], CDGMVAE [34], Music FaderNets[34]
Diffusion Models	Noise Addition & Reversal	Audio	High-fidelity generation, State-of-the-art audio quality	Slow sampling, Complex control mechanisms	Riffusion [1], MusicLM [1], Noise2Music ⁶⁸
LLMs (End-to-End)	Language Modeling, Attention	Symbolic Tokens, Audio Tokens	Natural language understanding, Controllability via prompts	Data scarcity, Representation limitations, Long coherence	MusicLM [1], MusicGen [1], SongComposer [1]

METHODOLOGIES FOR CULTIVATING EMOTIONAL NUANCE

Conditional Generation Strategies

Conditional generation, where the model's output is guided by auxiliary inputs or "conditions," is a primary approach for emotional control [35]. Various types of conditions are used:

Emotion Labels/Categories: The most direct method involves training models to generate music conditioned on discrete emotion labels (e.g., 'happy', 'sad', 'angry', 'relaxed') or quadrant labels from the VA model (Q1, Q2, Q3, Q4) [6].

Feeding label embeddings often implements this as additional input to RNNs or Transformers, or by using Conditional VAEs (CVAEs) [33]. While straightforward, this approach suffers from the inherent subjectivity, ambiguity, and coarseness of emotion labels, potentially confusing the learning process [24].

Continuous Emotion Dimensions (VA): To allow for finer control and the representation of intermediate emotional states, models can be conditioned on continuous values along the Valence and Arousal axes [35]. This enables conditioning on arbitrary points within the circumplex model, potentially capturing more subtlety than discrete categories. Transformer-based architectures have been adapted for this purpose [35]. VA values can also be used as targets in loss functions, for instance, in image-to-music generation aiming for emotional alignment [36].

Musical Attributes: Instead of directly using emotion labels, models can be conditioned on intermediate, musically meaningful attributes that are known correlates of emotion, such as tempo, mode, complexity, consonance, or note density [8]. The EmoGen system proposes using attributes as a bridge between emotion and music to mitigate the subjectivity of direct emotion labels [24]. EmotionBox uses pitch histogram (representing mode) and note density (representing tempo) as explicit control inputs [8]. This approach grounds control in interpretable musical terms.

Cross-Modal Inputs: Music generation can be conditioned on information from other modalities, often carrying implicit or explicit emotional content:

- i. **Text:** Text-to-music systems allow users to specify desired emotions, moods, styles, or instrumentation using natural language prompts. LLMs play a key role in interpreting these prompts [1].
- ii. **Images:** Systems generate music intended to match the atmosphere or emotion conveyed by an image [36]. Emotion acts as the bridge, learned by associating visual features with musical features from image-music pairs labeled with the same emotion. Systems like TransEmote analyze image content (objects, actions) and color palettes to create an emotional profile that guides music generation.

Reinforcement Learning (RL) with Affective Rewards

Reinforcement learning offers a different paradigm for shaping the output of generative models [37]. In this framework, the music generation model acts as an "agent" that learns a "policy" (how to generate music, e.g., note by note) to maximize a cumulative "reward" signal provided by the "environment" [37]. RL is beneficial for optimizing non-differentiable objectives or incorporating complex criteria into the generation process [38].

Reward Design for Emotion: The core challenge lies in designing reward functions that effectively guide the model towards generating music with the desired emotional qualities.⁸¹ Several types of rewards have been explored:

- i. **Rule-Based Rewards:** Assigning positive or negative rewards based on whether the generated music adheres to predefined rules, often derived from music theory (e.g., rewarding harmonic consonance, penalizing excessive repetition or notes outside the key) [38].
- ii. **Model-Based Rewards:** Incorporating the likelihood of a generated sequence under a pre-trained predictive model (e.g., an RNN trained on a large music corpus) as part of the reward. This helps maintain musical coherence and style learned from data while optimizing for other criteria [37].
- iii. **Emotion-Based Rewards:** Directly rewarding the model based on the perceived emotion of its output. This could involve using a separate, pre-trained emotion recognition model to classify the generated music and provide a reward signal based on alignment with the target emotion. For instance, Zhang and Tian used RL with rewards to ensure emotional consistency between generated melodies and input lyrics [5].
- iv. **Human Feedback (RLHF):** A robust approach involves learning a reward model directly from human preferences [38]. Humans compare pairs of generated musical excerpts and indicate which one they prefer based on criteria like musicality, adherence to a prompt, or emotional impact. A reward model is trained on these pairwise comparisons, and RL is then used to fine-tune the generative model to maximize the scores predicted by this learned reward model [38]. The MusicRL system successfully applied this, using automatic metrics (text adherence, acoustic fidelity) and a reward model trained on large-scale user feedback to improve a text-to-music model. User preferences in this context strongly correlate with perceived musicality [38].

RL holds significant potential for optimizing generated music for subjective qualities like emotional impact, which are difficult to capture with standard loss functions. Research linking musical pleasure to the brain's dopaminergic reward system, which also influences RL, provides a neurobiological basis for this connection [39]. However, crafting practical reward functions that reflect nuanced emotional goals remains a key challenge.

Integrating Music Theory Knowledge

While many deep learning approaches are purely data-driven, there is growing interest in integrating explicit knowledge from music theory into AI music generation models [40]. Music theory provides a rich framework of principles governing musical structure, harmony, counterpoint, form, and their relationship to perceived tension, release, and emotional expression in many musical traditions.

Methods for integration include:

- i. **Rule-Based Constraints/Rewards:** Incorporating music theory rules as hard constraints during generation or as soft constraints via reward functions in RL [37].
- ii. **Theory-Informed Representations:** Designing input or output representations that explicitly encode theoretical concepts, such as using Roman numeral analysis for functional harmony [41] or representing hierarchical structures inspired by Schenkerian analysis [42]. The SynTheory dataset was created to probe whether foundation models learn basic theory concepts.
- iii. **Analysis-Synthesis Loop:** Using computational music analysis based on theory (pre-analysis) to extract features or structures that inform the generation process, or using post-analysis to evaluate and refine the generated output according to theoretical principles [43].

iv. Model Architecture Design: Designing network architectures that inherently reflect musical hierarchies or theoretical relationships [42].

Incorporating music theory could enhance the structural coherence, musicality, and potentially the emotional appropriateness (within a given stylistic context) of AI-generated music [42]. It might provide a way to guide the generation towards more complex and meaningful structures than purely statistical learning might discover. However, this integration presents a potential dilemma. Music theory is often specific to particular styles (e.g., Western common practice tonality) and historical periods. Rigid adherence to established rules could stifle creativity, limit the generation of novel styles, or prove ineffective for non-Western or contemporary music genres that operate under different principles [42]. The challenge lies in finding flexible ways to integrate theoretical knowledge—perhaps as biases, priors, or evaluation components—that inform and guide the generative process without unduly restricting its creative potential or cross-cultural applicability.

Hybrid Symbolic-Neural Approaches

Another potential avenue involves creating hybrid systems combining traditional symbolic AI (GOFAI) strengths with modern neural networks. Symbolic AI represents knowledge explicitly through symbols and rules, enabling logical reasoning and providing interpretability [44]. Neural networks excel at learning complex patterns from large datasets and generating creative content [44]. A hybrid approach could potentially leverage symbolic reasoning for aspects requiring explicit knowledge or structure, such as applying music theory rules, ensuring long-range formal coherence, or implementing psychology-based emotion mappings, while using neural networks for tasks involving perceptual feature extraction, stylistic learning, and nuanced generation [44]. This could lead to robust, more transparent, and explainable systems than purely neural approaches. While direct applications to nuanced music emotion generation are still emerging, the concept offers a promising direction for combining rule-based interpretability with data-driven flexibility. Challenges include finding effective ways to integrate these fundamentally different AI paradigms [44].

Table 2: Methodologies for Emotion Control in AI Music Generation

Methodology	Control Mechanism / Signal Used	Key Characteristics	Representative Systems/Refs.
Conditional Gen - Labels	Discrete emotion categories (e.g., happy, sad, Q1-Q4)	Direct control, Simple implementation, Subjectivity/coarseness	Emotion-CVAE [33], LHVAE [31], EMOPIA [27]
Conditional Gen - VA	Continuous Valence-Arousal values	Finer control than labels, represents intermediate states, Maps well to dimensions	Sulun et al. [35], Image-to-Music[36]
Conditional Gen - Attributes	Musical features (tempo, mode, density, complexity)	Interpretable musical control, Bridges emotion & features, Avoids label subjectivity	EmotionBox [8] , EmoGen [24]
Conditional Gen - Text	Natural language prompts describing emotion, mood, and style	High flexibility, Leverages LLM understanding, and User-friendly input	MusicLM [1], MusicGen[1], SongGen [45], MusicRL [38]
Conditional Gen Image/Video	Visual content, Extracted visual features (color, objects, motion)	Cross-modal generation, Emotion as bridge, Dynamic synchronization (video)	Video2Music [46], XMUSIC [30]
Conditional Gen EEG/Physio	Real-time bio-signals (EEG, GSR, etc.)	Personalized generation, Real-time responsiveness, Requires sensor data & mapping	EEG-Transformer [5], EEG-driven system[47]
Feature Manipulation (Psych)	Direct modification of score/performance features via rules	Highly interpretable, based on psychological findings, can be rigid	EmotionBox [8]
RL - Rule/Model-Based Rewards	Rewards from music theory rules or predictive models	Optimizes for structure/musicality, can incorporate prior knowledge	Jaques et al [37]., Jiang et al. [38]
RL - Human Feedback (RLHF)	Rewards learned from human preference comparisons	Optimizes for subjective quality/appeal, Aligns with user judgment	MusicRL[38]
Disentanglement (VAE-based)	Manipulation of specific latent space dimensions	Targeted control over learned factors, Potential for interpretability	Music FaderNets [34], CDGMVAE [34], Two-stage valence/arousal [33]
Music Theory Integration	Incorporating harmony, counterpoint, and form principles	Enhances structure/coherence, Potential for style-specific emotional guidance	Schenkerian-informed models [42], Functional harmony representation [41], Theory rewards [37]
Hybrid Symbolic-Neural	Combining rule-based reasoning with neural generation	Potential for interpretability + flexibility, Complex integration	Conceptual [44]

This table summarizes the diverse strategies employed to instill emotional control in AI music, highlighting their control mechanisms, characteristics, and examples from the literature.

CONCLUSION

The quest to imbue AI-generated music with emotional nuance represents a significant frontier in computational creativity and affective computing. Current research demonstrates considerable progress in developing sophisticated generative models (Transformers, VAEs, GANs, Diffusion Models, LLMs) and exploring various methodologies for emotional control. Techniques range from conditioning on explicit labels or attributes, leveraging cross-modal inputs (text, images, bio-signals), directly manipulating musical features based on psychological principles, and employing reinforcement learning with affective rewards. However, despite these advancements, the generation of music consistently exhibiting deep, subtle, and evolving emotional expression comparable to human artistry remains an unsolved problem.

The primary obstacles hindering progress are multifaceted. The inherent subjectivity and context-dependency of musical emotion make it difficult to universally define, model, and evaluate. Current emotion representations (categories, VA space) often oversimplify the richness of affective experience. Data limitations severely constrain model learning, including scarcity of nuanced datasets, problematic annotations, and cultural biases. Furthermore, evaluating emotional nuance effectively poses a significant methodological challenge, with objective metrics often lacking perceptual relevance and subjective methods being resource-intensive and variable. Finally, achieving fine-grained control over specific effects and ensuring long-range structural coherence under emotional constraints remain open research questions.

Promising avenues for future research lie in addressing these hurdles directly. Creating more affluent, more diverse, and carefully annotated datasets is crucial. Developing more sophisticated emotion models that capture dynamics, complexity, and specific effects, potentially informed by cognitive science, is essential. Enhancing controllability and interpretability through advanced architectures, representation learning, and user interaction paradigms will be key. Establishing robust and meaningful evaluation frameworks targeting compelling nuance is necessary to guide progress. Furthermore, fostering culturally aware AI, deepening the integration of music theory and cognitive principles, and proactively addressing ethical considerations are vital for developing emotionally sophisticated AI music systems.

The journey towards AI that can generate emotionally nuanced music is complex but holds immense potential. Success requires a profoundly interdisciplinary approach, integrating computer science, musicology, psychology, neuroscience, and cultural studies insights. As AI models become more capable of understanding context, learning complex relationships, and interacting with users, the possibility of AI evolving from a mere generator of patterns to a genuine tool for emotional expression—or even a creative partner—becomes increasingly tangible. Achieving this goal would advance AI capabilities and unlock transformative applications in areas as diverse as personalized therapy, adaptive educational tools, immersive entertainment experiences, and assistive technologies for creative expression. Continued research focused explicitly on musical emotions' intricacies, which are essential for realizing AI's artistic and functional potential in music.

REFERENCE

- [1] Y. Zhao *et al.*, “AI-Enabled Text-to-Music Generation: A Comprehensive Review of Methods, Frameworks, and Future Directions,” *Electronics*, vol. 14, no. 6, Mar. 2025, doi: 10.3390/electronics14061197.
- [2] J. Liang, “Harmonizing minds and machines: survey on transformative power of machine learning in music,” *Frontiers in Neurobotics*, vol. 17, Nov. 2023, doi: 10.3389/fnbot.2023.1267561.
- [3] M. Reybrouck and T. Eerola, “Music and Its Inductive Power: A Psychobiological and Evolutionary Approach to Musical Emotions,” *Frontiers in Psychology*, vol. 8, Apr. 2017, doi: 10.3389/fpsyg.2017.00494.
- [4] Y. Chen, L. Huang, and T. Gou, “Applications and Advances of Artificial Intelligence in Music Generation: A Review,” arXiv.org. Accessed: Apr. 10, 2025. [Online]. Available: <https://arxiv.org/abs/2409.03715>
- [5] H. Jiang, Y. Chen, D. Wu, and J. Yan, “EEG-driven automatic generation of emotive music based on transformer,” *Frontiers in Neurobotics*, vol. 18, Aug. 2024, doi: 10.3389/fnbot.2024.1437737.
- [6] X. Gao *et al.*, “AI-Driven Music Generation and Emotion Conversion,” in *AHFE International*, AHFE International, 2024. Accessed: Apr. 10, 2025. [Online]. Available: https://openaccess-api.cms-conferences.org/articles/download/978-1-958651-99-5_9
- [7] T. Adhikari, “MeloHarmony: Exploring Emotion in Crafting AI-Generated Music with Generative Adversarial Network Powered Harmony,” *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4575257.
- [8] K. Zheng *et al.*, “EmotionBox: A music-element-driven emotional music generation system based on music psychology,” *Frontiers in Psychology*, vol. 13, Aug. 2022, doi: 10.3389/fpsyg.2022.841926.
- [9] B. Xie, J. C. Kim, and C. H. Park, “Musical Emotion Recognition with Spectral Feature Extraction Based on a Sinusoidal Model with Model-Based and Deep-Learning Approaches,” *Applied Sciences*, vol. 10, no. 3, Jan. 2020, doi: 10.3390/app10030902.
- [10] S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson, “Changing Musical Emotion: A Computational Rule System for Modifying Score and Performance,” MIT Press. Accessed: Apr. 10, 2025. [Online]. Available: https://www.researchgate.net/publication/49512283_Changing_Musical_Emotion_A_Computational_Rule_System_for_Modifying_Score_and_Performance
- [11] M. Civit, V. Drai-Zerbib, D. Lizcano, and M. J. Escalona, “SunoCaps: A novel dataset of text-prompt based AI-generated music with emotion annotations,” *Data in Brief*, vol. 55, p. 110743, Aug. 2024, doi: 10.1016/j.dib.2024.110743.
- [12] C. Peng and Z. Zhang, “Exploring the use of AI-generated AI-based drawings and music in Bipolar affective disorder interventions,” *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns-2024-3651.
- [13] Y. Sun, M. Kuo, X. Wang, W. Li, and Q. Bai, “Emotion-Conditioned MusicLM: Enhancing Emotional Resonance in Music Generation,” in *2024 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, Jun. 2024, pp. 1–8. Accessed: Apr. 10,

2025. [Online]. Available: <https://doi.org/10.1109/cec60901.2024.10612075>
- [14] L. Mou *et al.*, “MemoMusic 4.0: Personalized Emotion Music Generation Conditioned by Valence and Arousal as Virtual Tokens,” in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, Jul. 2024, pp. 1–6. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/icmew63481.2024.10645420>
- [15] D. Zhang, X. Li, D. Lu, Y. Tie, Y. Gao, and L. Qi, “Multitrack Emotion-Based Music Generation Network Using Continuous Symbolic Features,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2024, pp. 1–6. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/icme57554.2024.10688343>
- [16] H. Li, Y. Zeng, Z. Bai, W. Li, K. Wu, and J. Zhou, “EEG-fNIRS-Based Music Emotion Decoding and Individualized Music Generation,” in *2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, IEEE, Sep. 2024, pp. 394–397. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/ichci63580.2024.10808139>
- [17] C.-H. Liu and C.-K. Ting, “Evolutionary composition using music theory and charts,” in *2013 IEEE Symposium on Computational Intelligence for Creativity and Affective Computing (CICAC)*, IEEE, Apr. 2013, pp. 63–70. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/cicac.2013.6595222>
- [18] N. Imasato, K. Miyazawa, C. Duncan, and T. Nagai, “Using a Language Model to Generate Music in its Symbolic Domain while Controlling its Perceived Emotion,” *IEEE Access*, pp. 1–1, 2023, doi: 10.1109/access.2023.3280603.
- [19] M. A. de Santana, C. L. de Lima, A. S. Torcate, F. S. Fonseca, and W. P. dos Santos, “Affective computing in the context of music therapy: a systematic review,” *Research, Society and Development*, vol. 10, no. 15, p. e392101522844, Nov. 2021, doi: 10.33448/rsd-v10i15.22844.
- [20] Dash and K. Agres, “AI-Based Affective Music Generation Systems: A Review of Methods and Challenges,” *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–34, Jul. 2024, doi: 10.1145/3672554.
- [21] Daly *et al.*, “Music-induced emotions can be predicted from a combination of brain activity and acoustic features,” *Brain and Cognition*, vol. 101, pp. 1–11, Dec. 2015, doi: 10.1016/j.bandc.2015.08.003.
- [22] H. Guo *et al.*, “EMO-Music: Emotion Recognition Based Music Therapy with Deep Learning on Physiological Signals,” in *2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC)*, IEEE, Feb. 2024, pp. 10–13. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/aimhc59811.2024.00008>
- [23] J.-H. Su, T.-P. Hong, Y.-H. Hsieh, and S.-M. Li, “Effective Music Emotion Recognition by Segment-based Progressive Learning,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct. 2020, pp. 3072–3076. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/smc42975.2020.9283500>
- [24] C. Kang *et al.*, “EmoGen: Eliminating Subjective Bias in Emotional Music Generation,” arXiv.org. Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/abs/2307.01229>
- [25] P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, and R. Panda, “A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition,” *Sensors*, vol. 24, no. 7, p. 2201, Mar. 2024, doi: 10.3390/s24072201.
- [26] Z. Huang, S. Ji, Z. Hu, C. Cai, J. Luo, and X. Yang, “ADFF: Attention Based Deep Feature Fusion Approach for Music Emotion Recognition,” in *Interspeech 2022*, ISCA: ISCA, Sep. 2022. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.21437/interspeech.2022-726>
- [27] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, “EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation,” arXiv.org. [Online]. Available: <https://arxiv.org/abs/2108.01374>
- [28] R. Khan and A. Venugopal, “Exploring Deep Learning Methods for Text Augmentation to Handle Imbalanced Datasets in Natural Language Processing,” in *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, IEEE, Nov. 2024, pp. 1–8. Accessed: Apr. 10, 2025. [Online]. Available: <https://doi.org/10.1109/delcon64804.2024.10866632>
- [29] X. Kong, “Deep Learning in Music Generation: A Comprehensive Investigation of Models, Challenges and Future Directions,” *ITM Web of Conferences*, vol. 70, p. 04027, 2025, doi: 10.1051/itmconf/20257004027.
- [30] X. Hua, “Towards controllable neural generation of arguments,” Northeastern University Library. Accessed: Apr. 25, 2025. [Online]. Available: <https://doi.org/10.17760/d20412319>
- [31] S. Ji and X. Yang, “EmoMusicTV: Emotion-Conditioned Symbolic Music Generation With Hierarchical Transformer VAE,” *IEEE Transactions on Multimedia*, vol. 26, pp. 1076–1088, 2024, doi: 10.1109/tmm.2023.3276177.
- [32] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, “MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11312.
- [33] Grekow and T. Dimitrova-Grekow, “Monophonic Music Generation With a Given Emotion Using Conditional Variational Autoencoder,” *IEEE Access*, vol. 9, pp. 129088–129101, 2021, doi: 10.1109/access.2021.3113829.
- [34] Z. Ning, X. Han, and J. Pan, “Semi-supervised emotion-driven music generation model based on category-dispersed Gaussian Mixture Variational Autoencoders,” *PLOS ONE*, vol. 19, no. 12, p. e0311541, Dec. 2024, doi: 10.1371/journal.pone.0311541.
- [35] S. Sulun, M. E. P. Davies, and P. Viana, “Symbolic Music Generation Conditioned on Continuous-Valued Emotions,” *IEEE Access*, vol. 10, pp. 44617–44626, 2022, doi: 10.1109/access.2022.3169744.
- [36] S. Kundu, S. Singh, and Y. Iwahori, “Emotion-Guided Image to Music Generation,” arXiv.org. Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/abs/2410.22299>
- [37] “Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning.” Accessed: May 20, 2025. [Online]. Available: <https://research.google/pubs/generating-music-by-fine-tuning-recurrent-neural-networks-with-reinforcement-learning/>
- [38] G. Cideron *et al.*, “MusicRL: Aligning Music Generation to Human Preferences,” arXiv.org. Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/abs/2402.04229>
- [39] B. P. Gold, M. J. Frank, B. Bogert, and E. Brattico, “Pleasurable music affects reinforcement learning according to the listener,” *Frontiers in Psychology*, vol. 4, 2013, doi: 10.3389/fpsyg.2013.00541.
- [40] Mycka and J. Mańdziuk, “Artificial intelligence in music: recent trends and challenges,” *Neural Computing and Applications*,

vol. 37, no. 2, pp. 801–839, Nov. 2024, doi: 10.1007/s00521-024-10555-x.

- [41] “Emotion-driven Piano Music Generation via Two-stage Disentanglement and Functional Representation.” Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/html/2407.20955v1>
- [42] “A New Dataset, Notation Software, and Representation for Computational Schenkerian Analysis.” Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/html/2408.07184v1>
- [43] Y. Zhang, “Utilizing Computational Music Analysis and AI for Enhanced Music Composition: Exploring Pre- and Post-Analysis,” *Journal of Advanced Zoology*, vol. 44, no. S6, pp. 1377–1390, Dec. 2023, doi: 10.17762/jaz.v44is6.2470.
- [44] M.-C. Dinu, C. Leoveanu-Condrei, M. Holzleitner, W. Zellinger, and S. Hochreiter, “SymbolicAI: A framework for logic-based approaches combining generative models and solvers,” arXiv.org. Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/abs/2402.00854>
- [45] “SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation.” Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/html/2502.13128v1>
- [46] “Video2Music: Suitable Music Generation from Videos using an Affective Multimodal Transformer model.” Accessed: May 20, 2025. [Online]. Available: <https://arxiv.org/html/2311.00968v2>
- [47] “The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation,” IEEE Xplore. Accessed: May 20, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8665362>