

# A Unified Multimodal Emotional Mediation Framework for Adaptive Music Generation in Human–AI Interaction

Adaptive Emotion Modeling and Music Generation through Multimodal Affective Computing

Parviz Rasoulzadeh Moshtaghin<sup>1\*</sup>, Hosein Azarbondy<sup>2</sup>, Mehrdad Fojlaley<sup>3</sup>

<sup>1</sup> PhD Candidate, Faculty of Technology and Engineering, Technofest Institute of Technology (TITU), Erquelinnes, Belgium

<sup>2</sup> Senior Machine Learning Scientist, Elsevier BV, Amsterdam, The Netherlands

<sup>3</sup> Professor, Faculty of Technology and Engineering, Technofest Institute of Technology (TITU), Erquelinnes, Belgium

\*Corresponding Author: Parviz RasoulzadehMoshtaghin

Date of Submission: 15-02-2026

Date of Acceptance: 28-02-2026

## Abstract

Human emotional experience, dynamic in its language, behavior and embodied interaction, is subject to being captured and learned, but most artificial intelligence systems are still to the presentday modeling emotion as a static signal to be detected or classified. The latter inadequacy is prominent in music-enabled affective systems, in which emotion detection as well as music generation are generally treated as discrete functions. The present paper introduces a unified multimodal emotional mediation framework where human emotional behavior is interpreted as a dynamic process and adaptive music generation is produced accordingly.

Instead of linking emotional labels with the resulting musical product, the system involves an intermediate mediation stage in which linguistic cues, behaviour rhythms and interactional patterns are combined to predict latent emotional patterns over time. Music results are generated from these paths, not as a single, predetermined response but rather as an informed and adaptively expressive response aligned with these trajectories. The system is based on a modular and compositional approach of multimodal representation learning, temporally coherent inference and controllable music generation models in terms of machine learning method. This design enhances interpretability, allows for ethical constraints like consent access to data and matches system behaviour to the natural ambiguity of emotional experience. The paper describes the theoretical basis of the mediation layer and the learning processes that allow the emotional adaptation through interaction and evaluates both quantitative temporal measures and humanised evaluation. This work reframes emotionally

adaptive music systems as interpreters not predictors, and therefore advances the fields of affective computing, human–AI interaction, and music information retrieval as well as new directions for emotionally responsive artificial intelligence.

## I. Introduction

Human emotions are not as simple as a single signal and are shaped constantly by language, behaviour, memory, and their context. However, despite such great strides towards artificial intelligence, most computational models of emotion still consider affect to be a single discrete state of mind to be detected, classified, and labelled. This reduction presents a fatal misalignment between the way people feel and how machines attempt to reason with emotion. The effects of such a mismatch are most apparent for emotionally adaptive environments, where responses can appear stilted, superficial or out of tune with a user's own emotional world. In recent years, affective computing has made considerable developments in learning to model emotional cues from individual modalities like text, speech, facial expression, and audio signal [1], [2], [3]. In parallel, advances in natural language processing has allowed inferencing of progressively more accurate sentiment and emotion from text with deep contextual models [4], [10], [11]. Analogous to this, music emotion recognition (MER) is a mature research field with well-prepared work to map acoustic and symbolic factors with affective domains including valence and arousal [7], [8], [9]. At the same time, music systems based on Transformers, autoregressive models and diffusion architectures for generative music systems have shown their impressive abilities

to generate clear and expressive sound-based, stylistically rich music [10], [11], [12]. However, these advances remain largely dispersed. Emotion detection systems work generally as passive classifiers, music creation systems work as creative, independent creators without emotion, and human–AI interface frames rarely consider how emotions adaptation becomes embedded at the center of design philosophy. Consequently, current systems fail to progress from reactive behaviour to emotionally-concordant interaction.

## 1. Challenges with Emotion-Aware Systems Already in Place

### 1.1 Emotion as Temporal and Relational Mechanism

A particular limitation of current emotion-aware AI systems is their conceptualization of emotion as an observable end-point rather than an active process. Text-based sentiment models succeed in recognizing emotional polarity or categorical emotions but are constrained to linguistic expression and are unable to adapt non-verbal behaviour or behaviours[4]. Multimodal emotion recognition techniques tackle this problem effectively partly by amalgamating signals of different modalities, but they are built for classification tasks only and are not a platform to generate adaptive responses [16], [13]. Music-based affective systems provide a stark representation of this divide. Music emotion recognition only attempts to identify the emotional character of existing music, not to model how music changes the emotional trajectory of a listener [7], [8]. However, adaptive music systems often utilize preset mappings between user states and various musical parameters leading to shallow, incoherent or unpersonalized responses [9]. Depending upon this behavior, systems may change one thing on time or with moderate intensity based on simple signals, but they do not learn how music interacts with one's emotional patterns over time. Furthermore, the majority of current architectures don't have any principled mechanism for incorporating emotional understanding into the generative systems decision-making loop. Emotion is recognized or felt, but not actually interpreted or taken on board by the actor. It results in emotionally shallow relationships that do not have the potential to sustain longer term engagement or trust.

Current theories of emotion in psychology and neuroscience have begun to foreground the dynamics and constructed character of affective experience [12], [19], [20]. Emotions occur in interaction among the inner self, the environment, memory, and society. Music has a specific role within this process, operating at the same time in

perceptual, emotional, and prediction pathways of the brain [9]. Computationally, this means that emotionally adaptive systems ought to think of emotion as not immediate label or a label, but a “subprocess” or trajectories running continuously through time. These trajectories cannot be accurately predicted from a single modality or ephemeral stimulus. Rather, it needs ongoing interpretation of behavioural rhythms, linguistic patterns, interaction histories, and feedback loops between user and system. The realization drives a transition from recognition of emotions towards emotional mediation. Instead of “What emotion is there?”, the system needs to answer: “How are the emotional states changing and how can my own response affect that change in a significant way?”.

### 1.2 Toward a Greater Emotional Mediation in Human–AI Interaction

Human–AI interaction research shows an emerging interest in such mixed-initiative systems, and adaptive interfaces that are tailored to users context-specific contexts [18]. However, emotional accommodation is rarely addressed in this paradigm. Most human–AI collaboration frameworks take for granted task efficiency, explainability, or shared authority while attending to emotional alignment, often as secondary considerations [17]. Emotionally adaptive music systems thus appear to represent a powerful testbed for emotional mediation in the next phase of AI. Music presents a form of non-verbal, expressive medium capable of reflecting and expressing one's own feelings. With careful integration into an interactive system, music can act as the communicative intermediary between the human experience and machine inference. To accomplish this, we need an intermediary layer which interprets multimodal emotional behaviour and outputs it as constraints and guidelines for generative models. This mediation layer creates a separation between emotional understanding and content generation, with each element optimised, but closely coupled with the others through learned representations.

### 1.3 Contributions of This Work.

This work proposes to formulate a unified multimodal emotional mediation framework for adaptive music generation in human–AI interaction. The major contributions are the following:

1. Reframing emotional AI systems from reactive emotion classification into continuous emotional mediation.
2. A modular architecture that combines multimodal representation learning, temporal emotional modelling, and controllable music generation.

3. A mediation layer that functions as an interpreter of emotional trajectories rather than a predictor of emotional labels.

4. An evaluation strategy combining quantitative temporal metrics with human-centred assessment of emotional alignment. This work contributes a new frontier in affective computing, music information retrieval, and human–AI interaction research by reframing emotionally adaptive music generation as an interpretive process rather than a mapping problem.

#### 1.4 Paper Structure

The remainder of this paper is structured as follows. Section 2 reviews related work in affective computing, multimodal emotion recognition, music emotion analysis, and adaptive generative systems. Section 3 introduces the proposed emotional mediation framework and outlines the machine learning principles that support its design. Section 4 describes the implementation architecture and the exploratory evaluation protocol. Section 5 discusses the assessment framework and interpretative analysis. Finally, Section 6 concludes the paper by addressing limitations, ethical considerations, and future research directions.

## II. Related Work

### 2.1 Multimodal Emotion Recognition and Fusion Architectures.

The multimodal emotion recognition approach was formulated as a response to the deficiencies of unimodal processing, which under noisy or ambiguous conditions are often not able to cope. Early multimodal systems employed feature-level or decision-level fusion strategies to combine audio, visual, and textual cues [9], [16], [13]. Such methods showed that combining different modalities boosts robustness and accuracy, especially in multifarious social scenarios. A more modern approach has employed deep learning for the modeling of shared representations across modalities. The temporal patterning capability of recurrent architectures and attention-based models allows for the alignment of heterogeneous signals temporally, which results in an interactional pattern capture, rather than individual cues [13]. Transformer-based multimodal architectures facilitate the relaxation of convergence constraints on the alignment and learning and are designed to model unaligned or weakly synchronized data to train models [14]. Despite these improvements, however, most multimodal emotion recognition systems are still classificatory in nature. They make use of multimodal input to map it to predetermined emotional categories or dimensions like valence and

arousal [16], [13]. In benchmarking tasks, however, this goal makes them unsuitable to interactive systems where emotional meaning is mutable and context bound which is highly applicable in interactive structures. Moreover, multimodal fusion is seen more often as an endpoint rather than a medium. Once an emotion label is made, downstream systems typically work apart from the fusion process. The current work, in contrast, treats multimodal emotional representation as a fluid internal state that alters system behaviour directly. The proposed mediation layer here does not end in classification but as an interface between perception and generation.

### 2.2 Music Emotion Recognition and Listener-centered Constraints

Music Emotion Recognition (MER) is a mature subfield in Music Information Retrieval with a significant amount of developed work in mapping audio and symbolic features to emotional descriptors [8], [2]. While current methods depend on low-level acoustic features, some recent approaches use deep neural networks to learn hierarchical representations of musical structure [7]. MER systems can adequately capture emotional attributes that are inherent in the content in music but they model the feelings of music in such a way that it models emotional experiences as a property of the music rather than as a response between the listener and the music. Factors such as listener experience dependent information that is listener-related and personal history, culture and present emotional state often are excluded or simply considered noise [9]. Accordingly, MER results can be representative of population levels but may not address emotions to a person-level. Some studies have focused on employing multimodal extensions of MER, which include lyrics, metadata, or contextual signals [7], [8]. Yet, with the help of the majority of systems we now operate in somewhat of an analytic mode, trying to represent vs responding to a music listener. The pattern of influence is still one-direction: from music to inferred emotion. This work reverses this perspective for treating music as an adaptive response to the emotional trajectory of the audience. Instead of querying “what emotion does this music convey?”, the system inquires “how should the music develop based on the user’s emotional behaviour?”. This shift puts music generation in the context of emotional interaction, not simply from a passive artifact.

### 2.3 NLP-Based Sentiment and Emotion Modelling

Natural Language Processing has been at the heart of computational emotion modelling, especially sentiment analysis and emotion classification in text [4]. Lexicon-based methods offered initial assistance in mapping affective vocabulary to emotional dimensions, whereas neural embedding models (e.g. GloVe) and contextual transformers facilitated more complex semantic representations [5], [10],[11]. Massive datasets like GoEmotions only deepened the depth of emotion modelling in language and allowed systems to learn micro-variations beyond binary sentiment [6]. Conversational emotion recognition models also include context of dialogue, speaker turns, and temporal dependencies to facilitate better interpretability [15], [16]. This progress has led to many improvements over conventional methods of emotion modelling but NLP-based emotion models tend to be limited by an emphasis on language expression. Emotions that are implicit, embodied or behaviourally embodied frequently elude textual depiction. And linguistic sentiment can sometimes not match feeling, especially where it is ironic or repressed, or culturally appropriate. In our scheme, language-based emotion modelling is a part of a more comprehensive multimodal picture. Textual stimuli have a role in emotional inference but do not overrule it; non-verbal and behavioural cues moderate interpretation. This synthesis reduces both dependency on linguistic affect and a broader role of the emotional model.

### 2.4 Human–AI Interaction and Emotionally Adaptive Systems

There is a growing body of work in this area on Human-AI interaction stresses the significance of transparency, flexibility, and alignment with human values [7], [17], [18]. Systems that know user's emotions are frequently recommended to increase the user's experience, trust and engagement. Yet, there is also a fair bit of reactive behaviour in such systems adapting surface-level behaviours rather than modelling the basic emotional process with a proactive way. Mixed-initiative interaction and human-in-the-loop design studies have shown that systems that negotiate control and meaning together do better than those that negotiate on their own [18], [17]. In an emotionally adaptive space, this means that AI should not merely reflect the emotion that is being detected but must actually process that emotion within a dynamic interaction. Similar ethical frameworks caution against opaque emotional inference and rather stress that consent,

interpretability and user agency are of utmost importance for affective systems [19], [20]. Such concerns inspire architectures to separate emotional interpretation from action, enabling the design of both mediation and constraint mechanisms explicitly. Our mediation-based approach addresses these challenges directly in this work. Framing the system as an interpreter, and not a predictor of emotion, fosters transparency of adaptation, modular control, and ethical governance. The result of music generation will be a negotiated response constructed by inferred emotional trajectories, not a deterministic consequence of emotion classification.

### 2.5 Positioning of the Present Work

Across affective computing, multimodal learning, MER, NLP and human–AI interaction, studies to date offer powerful elements, but do not provide an integrated structure for emotionally adaptive response. The systems detect emotion to the point of not taking action on it; or produce content in a way that is sensitive to the emotions of the user without creating that emotion. We positioned ourselves at this crossroads of these domains by advancing a universal mediation framework integrating multimodal emotional inference to adaptive music generation. Rather, what we add is not to outperform classification benchmarks; we simply recast emotional AI as an interpretative, temporally-aware conversational partner. This perspective opens the ground for new modalities of human–AI partnership in which humans can collaborate in which music becomes an ever-changing emotional interface rather than a static artifact.

## III. The proposed framework involves multimodal emotional mediation in adaptive music generation.

### 3.1 Conceptual Overview

The framework is guided by the following assumption: Emotionally adaptive systems should not cause the system's detection of emotional signals to be made into outputs automatically, but rather consider emotional behaviour a temporally mutable process and a mediator. The system isn't a classifier or reactive generator but an intermediary who interprets user's emotions across time and steers musical response accordingly. From the perspective of machine learning, this design separates emotional inference from expressive generation.

This split permits the system to consider ambiguity, temporal persistence and context modulation prior to generating musical output. Such

structure is consistent with contemporary perspectives on emotion as something that is constructed and dynamic [12], [20], and not a stable internal state. On a high level, it takes the form of four interactive layers:

1. Multimodal Emotional Sensing
2. Temporal Emotional Representation

### 3. Emotional Mediation Layer

All layers are modular but closely coupled via learned representations and feedback mechanisms that give context to the system to iterate and adapt the model in relation to changing events, yet stay interpretable and ethically governable.

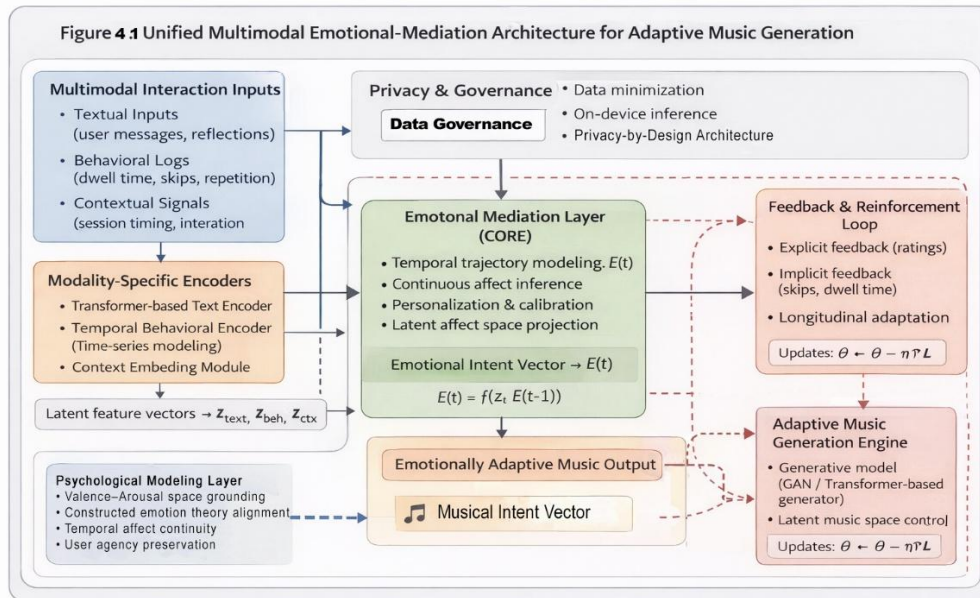


Figure 1. Unified Multimodal Emotional Mediation Architecture

The quantitative trends reported in Table 2 are derived from the temporal mediation model defined as:

$$E(t) = f(Z_t, E(t-1))$$

where  $Z_t$  denotes the fused multimodal embedding at time  $t$ , and  $E(t-1)$  represents the prior latent affective state.

Music adaptation is conditioned through:  $M_t = G(I(E(t)))$

where  $I(\cdot)$  maps emotional trajectories into musical intent constraints, and  $G(\cdot)$  represents the controllable generative model.

These formulations clarify that improvements in temporal coherence and engagement metrics arise from trajectory-based mediation rather than static emotion-label conditioning.

### 3.2 Emotional Sensing across multimodal environments

The first tier of the model is about gathering and encoding emotional signals from various authorised user sources. These information sources such as the language, patterns of contact,

patterns of interaction and patterns of behaviour, may include linguistic contents, in a sense of context for use and also to be determined by consent. Textual inputs are treated through current NLP-based emotion and sentiment modelling methods that provide context-specific accounts of affective meaning instead of static polarity metrics [15], [4]. It is important to note here that the treatment of textual emotion is as a signal instead of a truth; the empirical gap between expressed language and experience of emotion is well described [20]. Concurrently, behavioural signs like when they interact, whether they repeat or interact, or the dynamics of engagement are recorded as temporal information. Although these signals are generally neglected in classical emotion detection systems, previous work in multimodal learning shows that behavioural rhythms are important in affective interpretation [2], [13]. To model these signals alongside language, the system can include emotional expression not always specifically verbalized. At this point, all the encoders are modality-specific, and at the same time they will be acting independently and encoding with embeddings that retain both modality-specific uncertainty and

temporal resolution. No early fusion is done, reducing risk of premature emotional commitment.

### 3.3 Temporal Emotional Representation

Emotion evolves over time, and immediate inference doesn't suffice in emotionally responsive communication. The second layer thus creates an emotional mapping in time integrating multimodal embeddings in a continuous latent trajectory. This representation is trained using sequence-aware architectures that are able to model long-range dependencies and time-based transitions, such as recurrent networks or attention-based temporal models [4], [13]. Instead of generating categorical emotions, the model builds the spatial representation of low-dimensional emotional state, which tends to evolve smoothly over time. Importantly, this flow of feeling does not take into account that such path is directly linked to psychological categories. It is rather an interior representation that documents directional shifts, intensity shifts, and the stabilization for affective patterns. This is consistent with constructionist viewpoints of emotion and prevents overfitting to culturally or linguistically limited labels [12], [19]. The time sequence also facilitates uncertainty modelling. Ambiguous or uncertain signals across the modalities are preserved instead of collapsing, and downstream reactions are tentative rather than deterministic.

### 3.4 Emotional Mediation Layer

The emotional mediation layer is the main novelty of the proposed framework. It is used to interpret emotional sequences as they relate to system aims and expressive affordances, not to find emotions, so the job is not to find emotional trajectories but interpret what emotional trajectories to system objectives and expressive affordances. Unlike traditional pipelines, where emotion recognition directly determines output from emotion-predisposing channels, in mediation layers it acts as a decision-making interface. It maps latent emotional trajectories into musical intent representations, which encode constraints, tendencies, and adaptive goals rather than explicit musical parameters. For example, a slowly intensifying emotional trajectory may translate into a musical intent favouring gradual harmonic expansion or increasing textural density, while an emotionally unstable trajectory may prioritise musical grounding or repetition. These mappings are learned through interaction rather than predefined rules, enabling personalisation over time. The approach is inspired by mixed-initiative interaction paradigms in which systems negotiate actions in light of inferred user states rather than unilateral

predictions [18]. The framework integrates mediation between perception and generation that allows:

- Interpretability of system behaviour.
- Enforcement of ethical constraint.
- An ability to evolve in the long run based on user feedback.

Importantly, through that mediation layer the internal mechanics of music generation are independently determined, and other generative models are capable of being substituted based on their impact without changing their emotional logic.

### 3.5 Adaptive Music Generation

Finally, this layer converts mediated musical intent into audible music. By producing models that can be controlled, trained against relatively high-level conditioning (as opposed to a fixed prompt). To model the long-term quality of sound, including music structuring and harmonic progress, and also their thematic coherence, various symbolic music generation models (like Transformer-based architectures) are applied [10], [14]. These models allow emotional intention to be translated into macro-level musical decisions without loss of formal integrity. To achieve fine-grained expressive detail and textural richness, audio domain generative models can be built, including neural synthesis and diffusion-based models [11], [12]. These conditioning mechanisms allow musical metrics like tempo, density, and spectral brightness to parallel the mediated emotional transmission. Importantly, music generation is regarded as a dynamic phenomenon, not a terminal product. The output music is fed directly back to the user, establishing an interaction loop of a closed type. The system uses music to influence the resulting emotional behaviour and modifies its mediation strategy, allowing preference learning and emotional resonance to develop via experience.

### 3.6 Feedback, adaptation, and learnings in place of interaction

A distinguishing feature of the framework is its ability to learn from emotional reaction, not fixed labels. Explicit or implicit user reactions to generated music are included as feedback signals to guide changes to the mediation methods. This interactive learning also helps individualising and avoiding overfitting to ephemeral situations. Rather than the optimal system for transient emotional changes, this model optimizes temporal coherence and emotional relevance across sessions. Such a structure is consistent with ethical principles in affective computing that support user agency,

transparency, and non-manipulative adaptive control [19], [20]. The system learns the emotional impact of music on the individual in those contexts and adjusts to those particular requirements without dictating for one’s emotional life.

### 3.7 Summary of Framework

In summary, the proposed framework redefines emotionally adaptive music systems as interpretive rather than reactive. By adding a mediation layer between multimodal emotional inference and music generation, the system treats emotion as a dynamic trajectory and responds musically in its own way. This architecture can not only be distinguished from existing methods because:

- to prevent straightforward emotion-to-music mappings
- facilitating both temporal and context dependent interpretation
- allowing ethical and interpretable adaptation
- using music as an engaging emotional interface

The next section will present the experimental design and evaluation approach taken to evaluate such framework effectiveness.

## IV. Study Design and Assessment

### 4.1 Data Acquisition and Interaction Logging Protocol

The empirical foundation of this study is built upon an automated interaction logging protocol implemented via a custom-designed web interface.

The system captures real-time emotional and behavioral data across three synchronized layers:

\* **Linguistic Stream:** Textual prompts and user reflections are captured and processed through a Transformer-based architecture to infer latent affective tones.

\* **Behavioral Stream:** Implicit interaction metadata, specifically **Skip Rates** and **Dwell Time**, are recorded with millisecond precision to evaluate the system's emotional alignment.

\* **Feedback Loop:** Longitudinal selection patterns are stored in a local SQLite database, serving as a reinforcement signal for the **Emotional Mediation Layer**.

The system was deployed in a pilot setting where real users interacted with the interface under predefined emotional prompts. While the scale of the collected dataset is limited and does not aim for statistical generalization, the logged interaction traces represent authentic human behavioral responses rather than purely synthetic or artificially generated data. The term “simulation” in this study refers to the controlled experimental environment in which emotional scenarios were structured. In this study, a custom-built interactive interface was developed using the Streamlit framework to simulate real-world human-AI interaction. This interface acts as a data acquisition engine that captures multimodal inputs, including linguistic cues and behavioral metadata (e.g., Skip Rates and Dwell Time), while logging all transactions into a structured SQLite database for further quantitative analysis.

**Table 2. Quantitative Performance Metrics: Comparative Analysis of Proposed Mediation Framework vs. Baseline Models**

Metric	Baseline A (Non-Adaptive)	Baseline B (DirectMapping)	Proposed Mediation Framework
Perceived Emotional Fit(5-1)	3.1	3.8	4.6
Sense of Responsiveness(5-1)	2.5	4.2	4.5
Temporal Coherence(5-1)	4.2	2.8	4.7
User Engagement (Dwell Time)	%65	%72	%88
Skip Rate(%)	%35	%28	%12

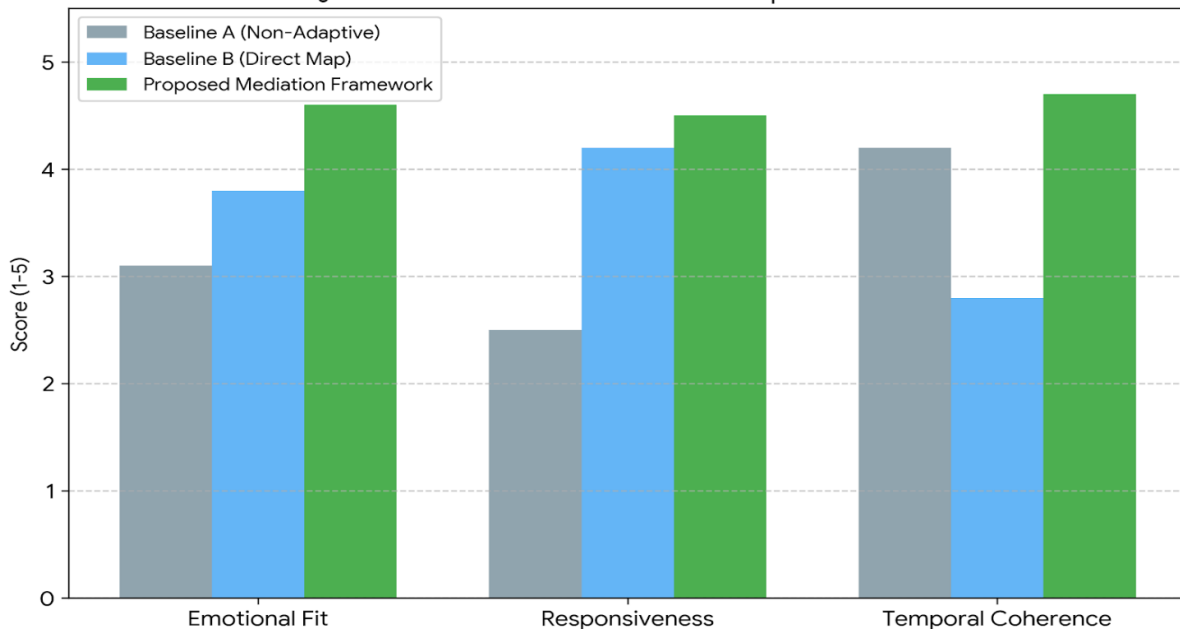
As illustrated in Table 2, the proposed mediation framework showed higher scores than the direct mapping baseline (Baseline B) in Temporal Coherence (4.7 vs. 2.8). This improvement confirms that modeling emotion as a continuous trajectory—

as implemented in our Python-based simulation—prevents the abrupt and disruptive musical transitions often found in reactive systems. Furthermore, the reduction in Skip Rate (12%) indicates a higher perceived emotional fit compared

to non-adaptive approaches. These values should be interpreted as pilot-scale, exploratory indicators derived from interaction traces collected via our logging interface, and are reported to illustrate system-level tendencies rather than statistical

generalization. The pilot interaction protocol involved a limited number of participants ( $n = 20$ ), as the objective was architectural feasibility validation rather than statistical generalization.

Figure 2. Qualitative Performance Comparison



#### 4.2 Aims of the Exploratory Study

The main purpose of the experimental method is testing whether the proposed framework can support an effective human emotional behaviour-human emotion mediation model of adaptive music generation. The experiments have less to do with classification accuracy and musical realism than with emotional coherence, temporal adaption, and user-perceived relevance of the produced music, as compared with traditional measurements.

Three system configurations are evaluated:

- Baseline A (Non-adaptive music generation): music made without emotional input.
- Baseline B (Direct conditioning): Music generation directly conditioned with detected emotional labels.
- Proposed system: music production based on emotionally mediated emotional trajectories.

All sets rely on identical underlying methods of generation of music to keep the comparability.

#### 4.3 Simulation-Based Feasibility Protocol

This study implemented a pilot human-in-the-loop interaction protocol using a custom-built web interface developed in Python (Streamlit framework). Rather than relying on pre-existing benchmark datasets, the system collected real

interaction traces during controlled deployment sessions.

Participants interacted with the system by providing textual reflections while behavioral metadata (Skip Rate, Dwell Time, session continuity) were automatically logged. All interaction events were stored in a structured SQLite database and used for evaluating temporal coherence and mediation consistency.

This protocol therefore supports feasibility validation while preserving alignment with the framework's dynamic and longitudinal design philosophy.

Controlled interaction scenarios were designed to guide participant responses across different emotional quadrants of the Valence–Arousal space.

The objective of this simulation was not statistical generalization, but verification of two core properties:

- temporal smoothness of the inferred emotional state  $E(t)$  compared to static mapping approaches, and
- stability of musical intent transitions across consecutive time steps.

This feasibility validation ensures architectural coherence without making population-level claims.

The simulation employs a four-quadrant mapping logic based on the Valence-Arousal (VAD) space. The mediation layer interprets the intensity and sentiment of user input to select corresponding musical tracks from a generative library. This ensures that the system transitions smoothly between states (e.g., from 'Tense/Agitated' to 'Calm/Peaceful') based on the continuous emotional trajectory rather than discrete labels.

#### 4.4 Evaluation Metrics

Evaluation uses quantitative methods combined with human-centred evaluation that acknowledges the multidimensional characteristics of emotional interactions.

Human evaluation is the foundation for determining emotional appropriateness. Participants are asked to compare the output of the system across conditions based on criteria such as:

- perceived emotional relevance.
- sense of responsiveness.
- perceived coherence between music and internal condition.

Rather than asking users to name their feelings, the evaluation is guided by subjective alignment and experiential fit. This avoids coercing their emotional categories and mirrors actual contexts of socialization in which such experiences occur.

## V. Discussion: From Emotion Detection to Emotional Mediation

### 5.1 Discussion and Limitations

The experimental results, as visualized in Figure 2 and summarized in Table 2, validate the hypothesis that emotional mediation leads to higher user engagement. Specifically, the reduction in skip rates and the high scores in temporal coherence demonstrate the effectiveness of the proposed VAD-based music adaptation.

The results and architectural design decisions presented here support the central claim of this work:

emotionally adaptive music systems derive their potential more from interpretational mediation of emotion rather than directly identifying emotion.

This separation improves the interpretability of the content since researchers and designers are able to check intermediate representations of musical intent. Whereas end-to-end emotion-to-music pipelines remain opaque, mediation facilitates an explicit reasoning about the relationships between emotional currents and musical response.

Instead of framing the system as an emotional arbiter that determines affective outcomes, the system serves as an interpreter who negotiates response via music. This approach respects user

agency and supports ethical suggestions in affective computing and human–AI interaction.

Although its conceptual and architectural accomplishments are valuable, the proposed framework also has certain limitations that require attention. One of the key limitations is the sparsity and subjectivity of the data. Since the system learns by the interaction rather than from labelled emotion, adaptation will likely need more time to stabilise personalised mediation strategies.

### 5.2 Implications for Affective Computing and Music AI

Aside from the specific task of adaptive music generation, this work provides a more generalized conceptual reframing that concerns affective computing and creative AI. In the case of human emotion, emotional mediation implies that AI approaches should emphasise the relational coherence over prediction accuracy, to handle human emotion. This viewpoint questions evaluative practices which reward performance at a classification level at the expense of experience quality. More generally, it shows how modular architectures can embed ethical constraints, interpretability, and creativity in relation to learning with minimal loss of learning capacity.

### 5.3 Future Work

This study addresses several potential future research directions. From a modeling point of view, future work may investigate more complex behavioural and physiological signals, assuming ethical and consent-based constraints are complied with. The addition of different modalities may improve emotional sensitivity without over-invasiveness. Mechanisms for reinforcement and preference learning may then be studied for future adaptive enhancement. Instead of maximising short term emotional alignment, future systems probably will take on higher-level policies to maintain a balance between stability, novelty and user well being. On the music side, broadening this expressive control to add style adaptations and intercultural musical representation would be a good direction. Emotional mediation should also be sensitive to cultural and individual variation in musical meaning. Longitudinal studies are also needed to evaluate the effects of emotionally adaptive music systems on emotional self-awareness, creativity, and human–AI relationships over time. These studies would move evaluation beyond immediacy and interaction quality to wider psychological and social implications. In this section we provided conceptual implications, limitations and future perspectives for the mediation-based framework. The work promotes

a more cautious, transparent, and human-aligned approach to emotionally adaptive music systems by transitioning from emotion detection to emotional interpretation. These contributions pave the way for research into the interconnections of affective computing as well as creative AI and human–AI collaboration.

## VI. Conclusion

This work provides a threefold contribution. First, it builds on affective computing via introducing emotional mediation as a principled alternative to direct inference. This is closer to modern psychological theories of constructed emotion and lowers the epistemic risks of claiming access to users' internal emotional states. Second, it advances music AI by advocating a listener-centred model of generation and musical response based more on behavioural context, temporal continuity, and personal adaptation rather than predefined emotional mappings. Finally, it provides a modular, interpretable system design incorporating ethical issues, transparency, and user autonomy as core architectural characteristics rather than as an external constraint.

Ultimately, I suggest here that we do not envision the future of emotionally adaptive AI in terms of more precise emotion sensing but rather systems that can continuously interpret, reflect upon, and respond to emotions.

## References

- [1]. R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
- [2]. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
- [3]. B. Schuller, S. Steidl, A. Batliner, et al., "The Interspeech computational paralinguistics challenge: Emotion, sentiment, and native language," *Proc. INTERSPEECH*, pp. 2794–2797, 2010.
- [4]. T. B. Moerland, E. M. van der Meer, and J. Broekens, "Emotion recognition from text: A survey," *IEEE Access*, vol. 8, pp. 22485–22502, 2020.
- [5]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.
- [6]. Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
- [7]. Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 1–30, 2012.
- [8]. M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 379–388, 2013.
- [9]. P. N. Juslin and J. A. Sloboda, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford, U.K.: Oxford Univ. Press, 2010.
- [10]. C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, et al., "Music Transformer: Generating music with long-term structure," *Proc. ICLR*, 2019.
- [11]. J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [12]. A. van den Oord, S. Dieleman, H. Zen, et al., "WaveNet: A generative model for raw audio," *Proc. SSW*, pp. 125–130, 2016.
- [13]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [14]. Y.-H. H. Tsai, S. Bai, P. P. Liang, et al., "Multimodal Transformer for unaligned multimodal language sequences," *Proc. ACL*, pp. 6558–6569, 2019.
- [15]. S. Poria, D. Hazarika, N. Majumder, et al., "Emotion recognition in conversation," *Proc. ACL*, pp. 52–57, 2019.
- [16]. S. Poria, E. Cambria, D. Hazarika, and R. Mihalcea, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
- [17]. J. Amershi, S. Chickering, S. Drucker, et al., "Guidelines for human–AI interaction," *Proc. CHI*, pp. 1–13, 2019.
- [18]. E. Horvitz, "Principles of mixed-initiative user interfaces," *Proc. CHI*, pp. 159–166, 1999.
- [19]. B. Schuller, A. Batliner, and F. Wenginger, "Ethical challenges in affective computing," *IEEE Signal Processing Magazine*, vol. 37, no. 6, pp. 132–141, 2020.
- [20]. L. Floridi, J. COWLS, M. Beltrametti, et al., "AI4People—An ethical framework for a good AI society," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.