Cross-Modal Emotion-Aware Music Generation for Therapeutic Healing Using LoRA-Tuned Transformers

Abstract—**The World Health Organization (WHO) reported that in 2019, nearly 970 million people worldwide suffered from mental health conditions, with anxiety and depression being the most frequently occurring. These disorders cause considerable distress, a decline in productivity, and elevated medical expenses. Although music therapy has been proven to have results in decreasing stress and stabilizing emotions, its application is limited by expense, inaccessibility, and reliance on professional practitioners. To overcome these shortcomings, this paper proposes a Generative AI approach to personalized and adaptive music therapy that incorporates cross-modal learning and large language models (LLMs). The system takes speech and text inputs, predicts Valence–Arousal–Dominance (VAD) scores with a custom Transformer, projects them onto emotion categories, and synthesizes structured prompts for music with MusicGen. By removing the necessity for human intervention, the model enables scalable and real-time therapeutic music delivery. Experimental assessment examines the Transformer's ability in VAD-based emotion prediction, and the outcomes demonstrate the potential for automated emotion support with music. The best-performing LoRA configuration achieves a Pearson Correlation Coefficient of 0.7622 for valence, 0.7811 for arousal, and 0.6724 for dominance prediction.**

Index Terms—**emotion recognition, music therapy, cross-modal learning, LoRA adaptation, transformers**

I. Introduction

Mental well-being is the foundation of health and determines how individuals perceive, respond to, and interact with their surroundings. The increasing number of cases of mental health issues is a worldwide concern, and millions of people suffer from depression and anxiety, among other conditions, contributing to emotional suffering, decreased productivity, escalated healthcare expenses, and social impact. Addressing this problem requires cost-effective, accessible, and scalable methods of emotional care that current approaches fail to achieve.

Music, the universal language, possesses therapeutic value, reducing stress, alleviating symptoms of mental illness, and stabilizing emotion. However, conventional music therapy, typically conducted by qualified practitioners, is hampered by expense and limited accessibility, denying many people access. Traditional music therapy methods rely on static playlists or therapist-guided improvisation, which, while effective, often lack real-time personalization and adaptability to users' changing mental states.

To bridge these limitations, this paper introduces a Generative AI framework that leverages cross-modal learning and large language models (LLMs) to provide customized therapeutic music. The system processes user input in the form of speech and text, estimates continuous Valence–Arousal–Dominance (VAD) scores, and maps them to emotional categories. Based on the emotional category, musical descriptions for therapy are retrieved and reformulated into structured prompts via LLM, which are then used to generate music. Human intervention is avoided through this method, bringing the advantages of music therapy to users via any intelligent device. This paper presents the end-to-end system design and implementation, including emotion modeling via a custom Transformer, a MusicGen-based retrieval-augmented generative music pipeline, and performance evaluation of the Transformer in predicting VAD-based emotions.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed methodology. Section IV presents experimental results and discussion. Section V concludes the paper.

II. Related Work

A. Cross-Modal Emotion Recognition

Cross-modal emotion recognition has become a prominent research topic that uses multiple modalities to enhance emotion detection precision. State-of-the-art research has demonstrated the superiority of attention-based fusion mechanisms over conventional concatenation strategies. Multimodal transformer models have exhibited strong potential in this field. Yu et al. introduced the Modular Co-Attention Network (MCAN), which utilizes self-attention units for intra-modal interactions and guided attention units for cross-modal interactions [8]. Similarly, the Cross-Modal RoBERTa (CM-RoBERTa) model presented by Luo et al. employs parallel self- and cross-attention mechanisms to extract intermodal and intramodal interactions among audio and text modalities [9]. The MemoCMT system proposed by Nam et al. addresses multimodal emotion recognition with cross-modal transformers that properly examine local and global speech patterns [10]. Hybrid fusion strategies combining both feature-level and model-level approaches have also been explored recently. The Multimodal Transformer Augmented Fusion (MTAF) technique proves that these kinds of hybrid methods can provide better performance by retaining fine-grained intramodal and intermodal information interactions [11].

B. Parameter-Efficient Fine-Tuning With LoRA

Low-Rank Adaptation (LoRA) has attracted significant attention as a parameter-efficient fine-tuning method aimed at overcoming the computational challenge of adapting large pretrained models to specific tasks. Recent LoRA applications span different domains, including natural language processing and computer vision. LoRA has emerged as well-suited for emotion recognition tasks where computational efficiency is paramount. Experiments have indicated that LoRA configurations with ranks of 4–32 can achieve performance comparable to full fine-tuning while having significantly lower memory requirements and training times [12]. The selection of appropriate LoRA hyperparameters, especially rank (r) and target modules, has been found to be crucial for optimization quality. Experimental work has shown that targeting all linear layers instead of only attention blocks may lead to higher adaptation quality [13].

C. Emotion Modeling and VAD Framework

The Valence-Arousal-Dominance (VAD) model is currently the leading continuous emotion representation framework, providing a superior alternative to discrete category-based models. The three-dimensional VAD model captures all emotional states, where valence represents pleasantness, arousal represents activation level, and dominance represents the sense of control. Recent work has utilized VAD representations for various emotion recognition tasks. Li et al.'s VADLE method uses continuous space VAD emotion knowledge to represent emotion correlations and create rich emotion distributions [14]. Similarly, brain signal analysis studies have revealed the effectiveness of the 3D VAD space in evaluating emotions with EEG signals [15].

D. Music Therapy and AI Applications

The convergence of artificial intelligence and music therapy represents an evolving field with vast therapeutic potential. Conventional music therapy, though effective, is limited by accessibility and customization issues that AI systems can address. Recent advances in AI-based music generation have demonstrated potential for therapeutic applications. Ferreira et al. proposed MusicGen-based systems for generating emotional music to help patients transition between emotional states [16]. Similarly, the EmotionBox system illustrated the possibility of producing music with specific emotional properties through deep neural networks [17]. These developments suggest significant prospects for democratizing access to music therapy via intelligent systems. The convergence of these research areas—cross-modal learning, parameter-efficient fine-tuning, emotion modeling, and AI-based music therapy—provides the rich foundation for developing comprehensive emotion-aware therapeutic music systems. Our research builds upon these existing techniques and introduces novel architectural innovations for real-time music therapy applications.

III. Methodology

This section outlines the comprehensive end-to-end pipeline for our emotion-aware therapeutic music generation system, illustrating the complete process from data preprocessing to music synthesis. The approach includes multimodal input extraction and processing, emotional state prediction via state-of-the-art neural architectures, emotional mapping to therapeutic music descriptions, and the generation of personalized audio outputs. Figure 1 shows the complete system pipeline, outlining the process in which a user narrates their day, triggering audio feature extraction and speech-to-text conversion, input to a custom transformer model to estimate valence-arousal-dominance (VAD) scores, which are translated to an emotion word, used to query a knowledge base for a healing music description, transformed by a language model to a music prompt, and finally synthesized into therapeutic music by MusicGen. The following subsections detail each component, including technical discussions of the design and implementation.

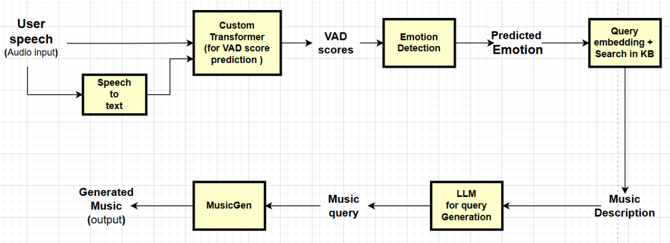

Fig. 1. Overview of the proposed emotion-aware therapeutic music generation pipeline.

A. Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [1], obtained from Hugging Face, is used for training and testing our multimodal emotion recognition system. IEMOCAP comprises approximately 12 hours of audio recordings collected from ten actors (five males and five females) engaging in dyadic conversations, including scripted and spontaneous interactions. The dataset includes 16 kHz audio recordings and manual transcriptions of the audio, making it suitable for cross-modal emotion recognition experiments using audio and text modalities. Emotion annotations are provided as continuous Valence-Arousal-Dominance (VAD) ratings between 1 and 5, where valence represents the degree to which the stimulus is perceived as pleasant or unpleasant, arousal represents the intensity of the emotion, and dominance represents the sense of control. The database comprises approximately 10,039 utterances, each lasting 4.5 seconds on average.

To enhance the emotion mapping module, IEMOCAP is extended with the NRC VAD Lexicon v2.1 [2], which provides human-rated VAD scores for more than 55,000 English words and phrases, including approximately 10,000 frequent multiword phrases. As described in the lexicon documentation, these VAD dimensions—valence, arousal, and dominance—encode the primary dimensions of word meaning, supporting detailed representation of emotional states. In contrast to most emotion recognition datasets that are confined to a limited number of discrete emotion classes (e.g., happiness, sadness, anger), the NRC VAD Lexicon's extensive coverage enables our system to capture a broad spectrum of emotions in a continuous three-dimensional space. This capability addresses the common limitation of other datasets, which typically lack sufficient granularity to capture the full range of human emotions, as demonstrated by the lexicon's comprehensive coverage and its usage across various domains including psychology, natural language processing, and social sciences. By integrating IEMOCAP's rich audio-text data with the NRC VAD Lexicon's fine-grained emotional annotations, our system undergoes robust training and achieves precise emotion-to-music mapping to handle a diverse set of emotional states for personalized therapy music generation.

B. Data Preprocessing

During training, text transcripts are already available, so no speech-to-text conversion is required in the preprocessing pipeline.

Audio Preprocessing: Raw audio waveforms are processed through a frozen HuBERT model [3] to obtain high-level speech representations. To enhance model robustness and generalizability, waveform-level augmentations such as additive noise and speed perturbation are applied during training.

Text Preprocessing: Transcripts are tokenized using a lowercased BERT tokenizer [4]. Synonym replacement through lexical substitution is employed to add variability to text inputs. These text features are combined with audio features in our multimodal architecture.

Label Smoothing: To address the inherent subjectivity in valence annotations, label smoothing [23] is employed during training, which encourages the model to generalize well on uncertain samples.

Data Splitting: Stratified splitting methods are utilized to ensure data balance across train, validation, and test sets. For the model that predicts valence, valence scores are divided into bins of equal lengths, and stratified sampling is performed according to the bins. For the Arousal-Dominance (AD) model, both arousal and dominance scores are divided into 3 bins, combined to create 9 different AD groups, and these composite labels are used for stratified splitting.

C. Model Architecture

Our multimodal emotion recognition model utilizes an advanced neural architecture that combines audio and text modalities using cross-modal attention mechanisms and LoRA (Low-Rank Adaptation) [5] for fine-tuning. Since valence prediction is a challenging task with high loss values, independent architectures are created for valence regression and arousal-dominance prediction to achieve optimal performance for each task.

1) Overall Architecture Overview: Our multimodal emotion recognition model is based on a dual-stream architecture that processes audio and text modalities simultaneously before their fusion for final prediction. Figure 2 depicts the complete architecture for the valence regression model, showing the detailed step-by-step flow from raw inputs to the final valence prediction.

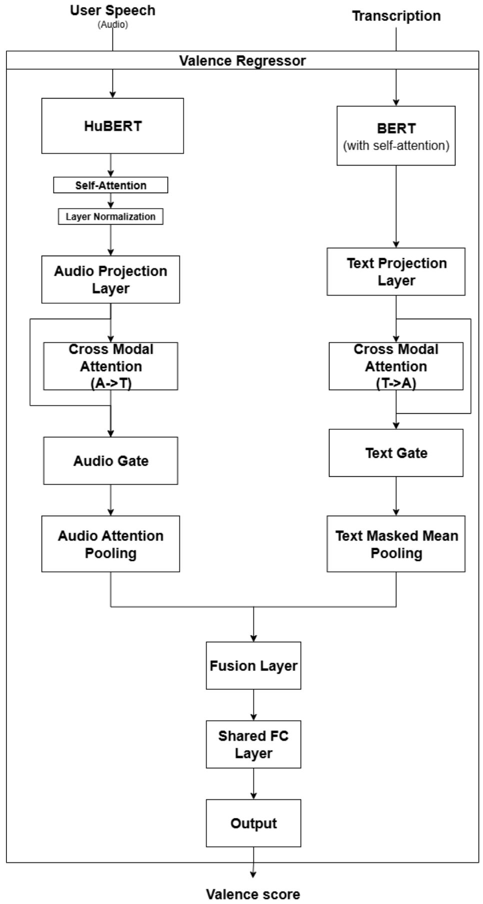

Fig. 2. Detailed architecture of the valence regression model showing the dual-stream processing of audio and text modalities with cross-modal attention and fusion mechanisms.

The architecture processes raw waveforms and transcriptions through specialized pathways as shown in Figure 2.

Audio Stream Processing: The audio pathway processes raw waveforms using: (1) HuBERT Encoder for acoustic feature extraction, (2) Custom Self-Attention Layers for emotion-relevant representation, (3) Normalization for feature stability, (4) Audio Projection to shared hidden dimension, (5) Cross-Modal Attention (A→T) attending to text features, (6) Audio Gate for information flow control, and (7) Audio Attention Pooling for temporal aggregation.

Text Stream Processing: The text processing pathway handles transcriptions through: (1) BERT Encoder utilizing built-in self-attention for semantic understanding, (2) Text Projection to shared hidden dimension, (3) Cross-Modal Attention (T→A) attending to audio features, (4) Text Gate controlling information flow, and (5) Text Masked Mean Pooling for token-level aggregation.

Fusion and Output: The two streams merge through: (1) Fusion Layer combining multimodal representations, (2) Shared FC Layer for common feature transformation, and (3) Output Layer for task-specific prediction, as illustrated in the architectural flow presented in Figure 2.

2) Encoder Components: **Audio Encoder:** The audio processing pipeline employs HuBERT (Hidden-Unit BERT) [3] as the base model to extract rich acoustic features. HuBERT provides strong feature extraction capability for emotion-relevant acoustic patterns through its 768-dimensional features from the `facebook/hubert-base-ls960` model. Audio processing handles fixed-length sequences (128,000 samples at 16 kHz $\approx$ 8 seconds) with LoRA integration on query and value projection layers (`q_proj`, `v_proj`).

Text Encoder: For processing textual content, BERT (Bidirectional Encoder Representations from Transformers) [4] was utilized to extract semantic and contextual information from transcriptions. The study employs BERT-base-uncased with 768-dimensional feature space, BERT tokenizer with a maximum sequence length of 512 tokens, and LoRA integration applied to query and value layers.

3) LoRA Mathematical Framework: Low-Rank Adaptation modifies the weight update mechanism by decomposing the adaptation into low-rank matrices:

LoRA Weight Update Formula:

$$W' = W_0 + \Delta W = W_0 + BA \tag{1}$$

where $W_0 \in \mathbb{R}^{d\times k}$ is the frozen pre-trained weight matrix, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable low-rank matrices, $r \ll \min(d, k)$ is the rank constraint, and $\Delta W = BA$ represents the learnable adaptation.

Forward Pass Computation:

$$h = W_0 x + BAx = W_0 x + B(Ax) \tag{2}$$

This decomposition reduces trainable parameters from $d\times k$ to $(d + k) \times r$, achieving significant parameter efficiency.

4) Modality-Specific Processing Strategies: **Text Processing: BERT Self-Attention:** For encoding text, BERT's native self-attention was used without additional layers. BERT has been extensively pre-trained and fine-tuned on diverse NLP tasks and is well-specialized for various text understanding scenarios including emotion recognition from transcripts.

Audio Processing: Custom Transformer Stack: In contrast to BERT, HuBERT was primarily developed for learning general speech representations and is not task-specific for emotion recognition. Therefore, specialized transformer encoder layers were addded to enhance HuBERT output for emotion-specific acoustic pattern extraction with 2 enhancement layers, 768 model dimension, multi-attention heads, and regularization to prevent overfitting.

The motivation for custom audio layers includes: (1) Task Adaptation—HuBERT requires additional processing for emotion recognition, (2) Temporal Modeling—Better capture of emotional temporal patterns in speech, (3) Feature Refinement—Enhanced alignment with emotion-related acoustic properties, and (4) Domain Gap—Bridges the gap between general speech representation and emotion classification.

5) Cross-Modal Attention Mechanism: The architecture implements bidirectional cross-modal attention to enable context enrichment between audio and text modalities:

Audio-to-Text Attention:

$$A2T = \mathrm{MultiHeadAttention}(Q_{\text{audio}}, K_{\text{text}}, V_{\text{text}}) \tag{3}$$

Text-to-Audio Attention:

$$T2A = \mathrm{MultiHeadAttention}(Q_{\text{text}}, K_{\text{audio}}, V_{\text{audio}}) \tag{4}$$

The MultiHeadAttention mechanism can be expressed as:

$$\mathrm{MultiHeadAttention}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O \tag{5}$$

where each head is computed as:

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{QW_i^Q (KW_i^K)^\top}{\sqrt{d_k}}\right)(VW_i^V) \tag{6}$$

Here, Q, K, and V are the query, key, and value matrices; W_i^Q, W_i^K, and W_i^V are learned projection matrices for the i-th head; W^O is the output projection matrix; h is the number of attention heads; and d_k is the dimensionality of the keys.

6) Gating and Pooling Mechanisms: **Adaptive Gating:** Sigmoid-based gating mechanisms regulate information flow across modalities and modulate the influence of self- and cross-attention outputs, allowing for selective amplification or suppression of attended representations.

Audio Gating:

$$g_{\text{audio}} = \sigma(W_a^g[\text{AudioProj}; A2T] + b_a^g) \tag{7}$$

$$\text{audio\_gated} = \text{AudioProj} \odot g_{\text{audio}} \tag{8}$$

Text Gating:

$$g_{\text{text}} = \sigma(W_t^g[\text{TextProj}; T2A] + b_t^g) \tag{9}$$

$$\text{text\_gated} = \text{TextProj} \odot g_{\text{text}} \tag{10}$$

where $[\cdot;\cdot]$ denotes concatenation and $\odot$ represents element-wise multiplication.

Pooling: Following gating, modality-specific pooling operations are used to aggregate the temporally distributed features into fixed-length representations for downstream fusion.

Audio Attention Pooling:

$$\text{pooled\_audio} = \sum_i \alpha_i \cdot \text{audio\_gated}_i, \tag{11}$$

where α_i is the attention weight.

Text Mean Pooling:

$$\text{pooled\_text} = \frac{\sum_i \text{text\_gated}_i \cdot \text{mask}_i}{\sum_i \text{mask}_i} \tag{12}$$

7) Task-Specific Architecture Variants: The system employs a modular structure for Valence, Arousal, and Dominance (VAD) regression, consisting of three primary stages: *Fusion*, *Shared FC*, and *Output*. Let $d =$ `hidden_dim`.

a) Fusion Layer:: Pooled audio and text representations are concatenated along the feature dimension:

$$\mathbb{R}^d_{\text{audio}} \oplus \mathbb{R}^d_{\text{text}} \to \mathbb{R}^{2d} \to \mathbb{R}^d$$

This operation allows the model to jointly leverage complementary cues from each modality in a shared feature space. Projecting the concatenated vector back to $\mathbb{R}^d$ maintains dimensional consistency between tasks and minimizes parameter overhead. The fusion stage is crucial for learning cross-modal interactions not achievable when handling modalities separately.

b) Shared FC Layer:: The joint representation undergoes transformation:

$$\mathbb{R}^d \to \mathbb{R}^d$$

This layer adds learnable transformations to prepare the merged representation for regression, and also serves as a regularization step to prevent direct overfitting of the fusion result to the target variable.

c) Output Layer:: The intermediate representation is mapped to the final scalar output via:

$$\mathbb{R}^d \to \mathbb{R}^1$$

The raw prediction is then scaled to the target range using:

$$\hat{y} = 1.0 + 4.0 \cdot \sigma(\text{raw}) \tag{13}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

d) Training Strategy:: Valence is optimized separately, while Arousal and Dominance are jointly trained, each having independent parameters for all three phases but sharing identical dimensional transformations. This design captures the empirical observation that valence prediction benefits from specialized optimization, while arousal and dominance exhibit greater mutual correlation and can successfully share intermediate representations.

8) Loss Functions: **Valence Loss (MSE with Label Smoothing):** To mitigate overfitting and account for uncertainty in valence labels, label smoothing is applied prior to loss computation:

$$\text{SmoothMSE}(\hat{y}, y) = \text{MSE}(\hat{y}, y_{\text{smooth}}) \tag{14}$$

$$y_{\text{smooth}} = y \cdot (1 - \varepsilon) + v_{\text{neutral}} \cdot \varepsilon \tag{15}$$

where ε is the smoothing factor and v_{neutral} denotes the neutral valence value.

Arousal–Dominance Loss (Huber Loss with Dynamic Weighting): A robust regression objective is adopted for arousal and dominance, with task-specific weights adjusted dynamically based on their relative errors.

Huber Loss [22]:

$$L_{\text{Huber}}(\hat{y}, y) = \begin{cases} 0.5(\hat{y} - y)^2, & \text{if } |\hat{y} - y| \leq \delta \\ \delta|\hat{y} - y| - 0.5\delta^2, & \text{otherwise} \end{cases} \tag{16}$$

Dynamic Weight Computation:

$$w_{\text{arousal}} = \alpha \cdot \frac{L_{\text{arousal}}}{L_{\text{arousal}} + L_{\text{dominance}}} \tag{17}$$

$$w_{\text{dominance}} = \alpha \cdot \frac{L_{\text{dominance}}}{L_{\text{arousal}} + L_{\text{dominance}}} \tag{18}$$

If $L_{\text{dominance}} > \beta \cdot L_{\text{arousal}}$:

$$w_{\text{dominance}} \leftarrow w_{\text{dominance}} \times \gamma,$$
$$w_{\text{arousal}} \leftarrow w_{\text{arousal}} \times \delta$$

The weights are subsequently normalized and clipped to $[r_{\min}, r_{\max}]$ to prevent any single loss term from dominating the optimization process.

Combined Loss:

$$\begin{aligned} L_{\text{total}} = {} & w_{\text{arousal}} \cdot L_{\text{Huber}}(\text{arousal}_{\text{pred}}, \text{arousal}_{\text{true}}) \\ & + w_{\text{dominance}} \cdot L_{\text{Huber}}(\text{dominance}_{\text{pred}}, \text{dominance}_{\text{true}}) \end{aligned} \tag{19}$$

9) Architectural Design Principles: **Separate Training Strategy:** The choice to train independent models is driven by: (1) Valence Complexity—High loss values during joint training suggest valence requires differential optimization, (2) Task-Specific Adaptation—Independent architectures enable superior hyperparameters for each task, and (3) Convergence Stability—Separate training prevents any single task from dominating learning.

Parameter Efficiency Through LoRA: LoRA integration provides ∼95% reduction in trainable parameters without compromising model expressiveness through frozen base models, low-rank updates, and storage efficiency, reducing model checkpoints from ∼2 GB to ∼50 MB.

Regularization Strategy: The regularization approach prevents overfitting through dropout on multiple layers, layer normalization for training stability, gradient clipping, and label smoothing to reduce overfitting to noisy valence labels.

D. Emotion Mapping

The emotion mapping module converts the continuous Valence-Arousal-Dominance (VAD) values predicted by the cross-modal transformer and rescaled to [1, 5] into discrete emotion categories used for therapeutic music generation. The NRC VAD Lexicon v2.1 dataset [2], with valence, arousal, and dominance scores for approximately 55,000 English words and multi-word expressions, defines a three-dimensional affective space. Each lexicon word is annotated with VAD values in [0, 1], rescaled to [1, 5] for compatibility with the transformer's output space.

The resulting VAD values are normalized to the range [-1, 1] using a MinMaxScaler, defined as:

$$v_{\text{norm}} = 2 \cdot \frac{v - v_{\min}}{v_{\max} - v_{\min}} - 1, \tag{20}$$

where $\mathbf{v} = (v, a, d) \in \mathbb{R}^3$ is the input VAD vector, and $v_{\min}, v_{\max}$ are the dataset's extremal VAD values to provide a normalized coordinate system. A K-means clustering algorithm [24] divides the normalized VAD space into 40 clusters that correspond to specific emotion categories, which minimizes the within-cluster sum of squares:

$$J = \sum_{i=1}^{40} \sum_{\mathbf{v} \in C_i} \|\mathbf{v} - \mu_i\|^2, \tag{21}$$

where C_i denotes the i-th cluster, $\mu_i \in \mathbb{R}^3$ is its centroid, and $\| \cdot \|^2$ refers to the Euclidean norm.

A convolutional neural network (CNN) is trained on the dataset to predict the emotion category for a normalized VAD vector $\mathbf{v}_{\text{norm}} \in \mathbb{R}^{3\times1}$. The proposed CNN model consists of two convolutional blocks followed by a fully connected layer and a softmax output layer for 40 clusters. Each convolutional block comprises convolution, normalization, and pooling operations. The model is optimized with categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{40} y_{ij}\log(\hat{y}_{ij}), \tag{22}$$

where N is the batch size, $y_{ij} \in \{0, 1\}$ is the indicator of the true label, and $\hat{y}_{ij} \in [0, 1]$ is the probability prediction for the j-th cluster. Training uses the Adam optimizer [20] and early stopping with patience = 5. Testing achieved an accuracy of 0.9220, precision of 0.9253, recall of 0.9220, and F1-score of 0.9220, demonstrating strong classification performance. The predicted cluster index corresponds to one of 40 pre-defined emotion labels, which is passed to the knowledge base retrieval module.

E. Knowledge Base Retrieval

The knowledge base, implemented as a Retrieval-Augmented Generation (RAG) pipeline using a ChromaDB vector database [19], stores and retrieves therapeutic music descriptions by emotion. The knowledge base is constructed from the NRC VAD Lexicon v2.1 dataset [2], which informs the discretization of the VAD space into 8 octants, defined by binary thresholding on valence, arousal, and dominance (e.g., high/low valence × high/low arousal × high/low dominance). Each octant contains 5 representative emotions, selected based on their proximity to the octant centroid in VAD space and therapeutic relevance, yielding 40 emotion categories in total (8 octants × 5 emotions). The categories are matched with professionally-written music descriptions for tempo, tonality, instrumentation, and dynamics that are tailored to address therapeutic objectives, such as relaxation or emotional release.

Descriptions are represented as dense vector embeddings $\mathbf{e}_d \in \mathbb{R}^{384}$ using the all-MiniLM-L6-v2 sentence transformer model [18] and stored in ChromaDB, indexed by emotion labels, for efficient similarity-based retrieval. Similarity-based retrieval uses cosine similarity to find the most similar description to a query embedding $\mathbf{e}_q$, obtained from the predicted emotion category, as follows:

$$\text{sim}(\mathbf{e}_q, \mathbf{e}_d) = \frac{\mathbf{e}_q \cdot \mathbf{e}_d}{\|\mathbf{e}_q\|_2 \|\mathbf{e}_d\|_2}, \tag{23}$$

where $\| \cdot \|_2$ is the Euclidean norm. The RAG pipeline searches the ChromaDB database and returns the best match ($n_{\text{results}} = 1$) along with its description, emotion label, and similarity score. The octant-based model, guided by the NRC VAD Lexicon v2.1 data, constrains the search space, improving precision by linking queries to emotionally consistent regions. ChromaDB's Hierarchical Navigable Small World (HNSW) indexing supports sub-millisecond query latency with a time complexity of $O(\log N)$, where N is the number of embeddings, and is designed to scale to large knowledge bases. The retrieved description, embedding musically informed therapeutic strategies, is passed to the prompt generation process to resonate with the user's emotional state.

F. LLM Prompting and Music Generation

The music synthesis and prompt generation phase combines a BART-large-CNN large language model (LLM) [6] with MusicGen [7] to generate therapeutic audio outputs. The LLM interprets the retrieved music description and VAD (Valence-Arousal-Dominance) scores to create a structured prompt. Description summarization uses a text summarization pipeline with constraints: 150 tokens maximum, 20 tokens minimum, and 4-beam search strategy. The resulting prompt includes the detected emotion, VAD scores, and interpretive annotations.

The prompt is fed into MusicGen, a transformer-based sequence-to-sequence model with a pre-trained audio tokenizer. Audio generation is guided by a conditional probability distribution:

$$P(\mathbf{a}|\mathbf{p}) = \prod_{t=1}^{T} P(a_t | a_{1:t-1}, \mathbf{p}), \tag{24}$$

where $\mathbf{a}$ is the audio sequence, $\mathbf{p}$ is the prompt, and T is the audio frame count (e.g., 2048 frames at 32 kHz). Utilizing CUDA-accelerated GPU processing, the system produces real-time music synthesis, facilitating the delivery of personalized emotionally intelligent music through smart devices at scale.

IV. Experiments and Results

A. Experimental Setup

Comprehensive experiments were conducted to evaluate the performance of our multimodal emotion recognition model in different architectural configurations. The experiments were designed to demonstrate the incremental improvement achieved by architectural enhancements, from baseline unimodal approaches to our final cross-modal transformer with LoRA fine-tuning.

All models were trained and evaluated on the same dataset splits to ensure fair comparison. Test loss (via appropriate loss functions for each model category) and Pearson Correlation Coefficient (PCC) for correlation evaluation of predicted vs. ground truth VAD scores serve as the evaluation criteria. In multimodal experiments, separate metrics were reported for valence (V), arousal (A), and dominance (D) prediction.

B. Baseline Unimodal Results

The study initially established baseline performance with unimodal architectures to understand the contribution of individual modalities prior to introducing cross-modal interactions.

GRU-based Unimodal Models: Our baseline employs GRU networks to model sequential audio features. The shared FCN approach, where a single fully connected network predicts all three VAD dimensions simultaneously, achieved an average test loss of 0.6045. This relatively high loss indicates the difficulty of joint prediction of VAD dimensions with a simple shared model.

Using separate fully connected networks for each VAD dimension resulted in significant improvement, with the average test loss reducing to 0.4313. This 28.7% reduction demonstrates that task-specific prediction heads are crucial for effective emotion recognition since different emotional dimensions may require distinct feature representations.

Transformer-based Unimodal Models: Replacing GRU with transformer architecture yielded considerable performance enhancement. The shared FCN unimodal transformer achieved an average test loss of 0.3607, a 40.4% improvement over the GRU baseline. This progress highlights the superior ability of transformer attention mechanisms [21] to capture complex temporal dependencies in emotional speech.

The unimodal transformer with separate FCN for each dimension further improved performance to an average test loss of 0.3473, though the improvement was less dramatic (3.7%) than the substantial performance shift when switching from GRU to transformer architecture.

C. Cross-Modal Integration Results

Initial Cross-Modal Architecture: Introducing cross-modal attention between audio and text modalities provided additional improvements. The cross-modal transformer achieved an average test loss of 0.3391, representing a 2.4% improvement over the best unimodal configuration. While the gain may appear modest, it demonstrates the value of multimodal fusion for emotion recognition.

Separate Model Training without LoRA: Due to the challenging nature of valence prediction as observed in our initial experiments, separate training strategies were implemented for valence and arousal-dominance models. Without LoRA fine-tuning, this approach showed promising performance with distinct behavior across dimensions as shown in Table I.

The significantly lower loss for arousal compared to valence and dominance supports the understanding that arousal is the most predictable dimension in our dataset, consistent with previous emotion recognition studies.

TABLE I
MODEL PERFORMANCE COMPARISON (WITHOUT LORA FINE-TUNING)

Model	**Avg Loss**	**V Loss**	**A Loss**	**D Loss**
GRU (Shared)	0.6045	-	-	-
GRU (Separate)	0.4313	-	-	-
Transformer (Shared)	0.3607	-	-	-
Transformer (Separate)	0.3473	-	-	-
Cross-modal (Joint)	0.3391	-	-	-
Cross-modal (Sep V&AD)	0.20034	0.3289	0.1046	0.1675

D. LoRA Fine-tuning Results

Different LoRA configurations were systematically explored to identify optimal parameter-efficient fine-tuning settings. The experiments validated the effectiveness of LoRA in improving model performance while maintaining computational efficiency.

LoRA Configuration: $r = 4$**,** $\alpha = 8$ This low-rank configuration performed exceptionally well across all dimensions, with valence achieving test loss 0.2947 and PCC 0.7478, arousal achieving test loss 0.0955 and PCC 0.7780, and dominance achieving test loss 0.1614 and PCC 0.6765. The high PCC values, particularly for arousal (0.7780), indicate strong correlation between predicted and actual values.

LoRA Configuration: $r = 8$**,** $\alpha = 16$ Scaling alpha proportionally with rank yielded mixed results. This configuration showed slight valence performance degradation with comparable arousal performance. This suggests that the lower-rank configuration ($r = 4$) may be better suited for valence prediction, possibly due to its stronger regularization effect.

LoRA Configuration: $r = 16$**,** $\alpha = 32$ The high-rank configuration performed best overall as indicated in Table II. This configuration maintained the lowest valence loss (0.2815) and highest arousal PCC (0.7811), representing our best-performing model. The superior performance of higher-rank adaptation for valence (PCC 0.7622) indicates that higher-rank LoRA adaptation can learn more complex emotion-specific patterns without compromising parameter efficiency.

TABLE II
LORA FINE-TUNED CROSS-MODAL TRANSFORMER RESULTS

2***LoRA Config**	**Valence**		**Arousal**		**Dominance**	
(r, α)	**Loss**	**PCC**	**Loss**	**PCC**	**Loss**	**PCC**
(4, 8)	0.2947	0.7478	0.0955	0.7780	0.1614	0.6765
(8, 16)	0.3115	0.7320	0.0941	0.7750	0.1648	0.6702
(16, 32)	**0.2815**	**0.7622**	**0.0919**	**0.7811**	0.1649	0.6724

E. Performance Analysis

Dimensional Performance Comparison: Across all configurations, arousal consistently showed the best performance with the lowest test losses and highest PCC values. This aligns with existing research suggesting that arousal is more directly reflected in acoustic features such as energy, pitch, and speaking rate. Valence prediction proved most challenging, requiring specialized architectures and higher-rank LoRA configurations to achieve satisfactory performance. Dominance showed intermediate difficulty, with stable but moderate correlation scores across configurations.

Architectural Progression Impact: The progression from unimodal GRU (average loss 0.6045) to our best cross-modal LoRA model (average loss ~0.1461 across VAD) represents a 75.8% improvement. This substantial enhancement demonstrates the cumulative benefit of architectural innovations: transformer attention, cross-modal fusion, separate task training, and LoRA fine-tuning.

LoRA Configuration Analysis: The $r = 16$, $\alpha = 32$ configuration was determined to be optimal, suggesting that emotion recognition benefits from higher capacity adaptation without sacrificing LoRA's parameter efficiency. The relatively small variation in performance across different LoRA settings (within 10% for most metrics) demonstrates that LoRA is robust to hyperparameter changes and is a reliable choice for parameter-efficient emotion recognition.

Cross-Modal Effectiveness: The substantial gains from cross-modal integration validate our hypothesis that audio-text modality integration provides complementary information for emotion recognition. The attention mechanism effectively learns to emphasize informative cross-modal features, as evidenced by consistent gains over unimodal baselines.

V. Conclusion

This work proposes an integrated AI-based personalized therapeutic music generation system for scalable emotional care. By leveraging a cross-modal Transformer for robust emotion recognition, an established knowledge base for emotion-to-music conversion, and MusicGen for real-time music synthesis, the system offers an end-to-end solution for automated music therapy. Experimental results demonstrate the system's effectiveness in accurately identifying emotional states, with remarkable improvement on the challenging valence dimension. The convergence of affective computing and generative AI enables the creation of a scalable and accessible approach to emotional well-being. This research represents a significant contribution toward democratizing emotional care through technology. By personalizing and providing access to music therapy, the proposed system holds potential for addressing the global mental health crisis through intelligent, compassionate intervention.

References

[1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[2] S. M. Mohammad, "NRC VAD Lexicon v2: Norms for valence, arousal, and dominance for over 55,000 English terms," National Research Council Canada, Tech. Rep., 2018.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learning Representations*, 2022.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[7] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Proc. 40th Int. Conf. Machine Learning*, vol. 202, 2023, pp. 6328–6343.

[8] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition*, 2019, pp. 6281–6290.

[9] J. Luo, H. Phan, and J. D. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," *IEEE Trans. Multimedia*, vol. 25, pp. 7394–7405, 2023.

[10] T. P. Nam, H. Kim, S. Lee, and J. Park, "MemoCMT: Multimodal emotion recognition using cross-modal transformer," *Scientific Reports*, vol. 13, art. no. 15234, 2023.

[11] Y. Tian, J. Ma, C. Wang, and L. Zhang, "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers in Neurorobotics*, vol. 17, art. no. 1181598, 2023.

[12] W. Cai, R. Zhang, J. Liu, and S. Wang, "LoRA-MER: Low-rank adaptation of pre-trained speech models for efficient multimodal emotion recognition," in *Proc. Interspeech*, Dublin, Ireland, 2023, pp. 2344–2348.

[13] B. Qin, D. Yang, J. Tao, and Z. Wen, "EELE: Exploring efficient and extensible LoRA integration in emotional text-to-speech," *IEEE Signal Process. Lett.*, vol. 31, pp. 2156–2160, 2024.

[14] X. Li, S. Zhao, J. Zhang, and Y. Liu, "VADLE: Practical continuous emotion representation learning with VAD knowledge," *Pattern Recognition*, vol. 120, art. no. 108157, 2021.

[15] C. Liu, W. Zhao, T. Zhang, and H. Chen, "Emotion recognition based on 3D VAD space using EEG signals," *Brain Sciences*, vol. 11, no. 4, art. no. 470, 2021.

[16] V. Ferreira, D. Santos, A. Silva, and M. Costa, "Emotional music generation using MusicGen for therapeutic applications," *Computers in Biology and Medicine*, vol. 165, art. no. 107426, 2023.

[17] F. Liu, K. Wang, T. Li, and H. Zhang, "EmotionBox: A music-element-driven emotional music generator based on music psychology," *Frontiers in Psychology*, vol. 13, art. no. 841926, 2022.

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, 2019, pp. 3982–3992.

[19] ChromaDB Team, "ChromaDB: The AI-native open-source embedding database," ChromaDB Inc., San Francisco, CA, USA, Tech. Rep., 2023. [Online]. Available: https://docs.trychroma.com/

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations*, San Diego, CA, USA, 2015.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Information Process. Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[22] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[24] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.