Process Modeling and Emotional Response Simulation of Musical Tension Evolution

Kexia Huang
School of Art, Guangxi University of Foreign Languages, Nanning 530222, Guangxi, China
Corresponding author: Kexia Huang (e-mail: hkx198642hjy@sina.com)

ABSTRACT Modeling the dynamic relationship between musical tension and emotional response remains challenging because existing approaches typically treat tension prediction and emotion recognition as independent tasks. This study proposes a process-oriented computational framework that integrates cognitive expectation with continuous emotional response simulation through hierarchical signal modeling and dynamic coupling mechanisms. A twelve-dimensional tension descriptor is extracted from symbolic scores and audio signals, while a Transformer-XL-enhanced IDyOM model estimates note-level surprise to characterize expectation violations. The resulting features are recursively encoded by a gated recurrent unit with expectation-aware modulation to generate continuous tension trajectories. Furthermore, a conditional variational autoencoder models listener-specific cognitive sensitivity, and a differential-equation-based coupling mechanism links tension evolution with arousal and valence dynamics to establish a closed-loop perception–expectation–response simulator. Experimental results demonstrate a peak localization error of 0.42 s, an arousal tracking RMSE of 0.152, a valence tracking RMSE of 0.176, and statistically significant causal effects of expectation modulation on simulation accuracy. The proposed framework provides an effective strategy for temporal signal fusion, dynamic state estimation, and adaptive response modeling, offering valuable references for intelligent audio signal processing, multimodal information perception, and computational sensing systems in electromagnetic signal analysis and next-generation interactive media applications.

INDEX TERMS musical tension modeling, audio signal processing, cognitive expectation, emotional response simulation, electromagnetic signal analysis

I. INTRODUCTION

MUSICAL tension, as a core dynamic element driving the listener's emotional experience, has its evolutionary process and the real-time coupling mechanism with emotional response as cutting-edge issues in the field of music cognitive computing. Existing research mainly follows two independent paths: one is tension prediction modeling based on information theory and statistical learning, focusing on the estimation and accumulation of surprise of note sequences; the other is emotional dimension labeling fitting based on acoustic features and regression models, focusing on the static mapping of valence and arousal [1-2]. However, the above paradigms regard tension prediction and emotional response as separate sequential tasks, lacking a unified process model framework to characterize the dynamic interaction and temporal causal pathway between the two. In particular, how tension evolution affects the emotional response through the real-time adjustment of the listener's internal cognitive expectations has not yet been clearly modeled and empirically verified [3-4].

This paper focuses on three specific research objects: first, the continuous quantitative estimation of the trajectory of musical tension evolution, requiring the model to capture the multi-scale accumulation process from note-level surprise to phrase-level structural tension [5-6]; second, the dynamic coupling relationship between tension and emotional response, i.e., how the rate of change of tension drives the fluctuation of arousal state on the time axis [7-8]; and third, the moderating function of cognitive expectation in the above coupling process, i.e. how the listener's prior expectations of musical events amplify or inhibit the actual driving effect of tension on emotion [9-10]. The above three objects correspond to three progressive levels of tension modeling, emotional simulation, and cognitive pathways, respectively. Their joint characterization is a key prerequisite for understanding the "perception-expectation-response" closed loop in music emotion computing.

Existing research has made attempts in the above three aspects. Barchet A V developed an open-source automated prediction model called TenseMusic, which used music

audio as the only input and automatically extracted six features such as loudness, pitch, tonality tension, roughness, tempo and onset frequency using music information retrieval methods to quantify tension [11]. IDyOM estimates the conditional probability of note events through a long short-term memory network and is the most widely used music statistical model in the field of music neuroscience [12-13]. Orjesek R performed dynamic emotion regression prediction of continuous arousal-valence on the DEAM (Database for Emotion Analysis in Music) dataset. The network architecture adopts one-dimensional convolutional layer stacking, autoencoder iterative reconstruction layer and bidirectional gated recurrent unit to verify the effectiveness of GRU (Gated Recurrent Unit) network in dynamic music emotion modeling [14]. However, existing methods still have limitations. The IDyOM framework relies on the fixed-order Markov assumption. Existing emotion models do not explicitly model the driving role of tension as an intermediate latent variable. The expected regulation pathway has not yet been quantified into a computational module that can be embedded in the simulation loop [15-16].

This paper constructs a model of the evolution of musical tension that integrates cognitive expectation to drive dynamic emotional simulation. Methodologically, a twelve-dimensional tension descriptor is extracted from the score and audio. Surprise is calculated using IDyOM enhanced by Transformer-XL, and the tension trajectory is estimated using a GRU with an expectation modulation term. Subsequently, a conditional variational autoencoder modulation pathway is constructed, coupling the rate of change of tension to both arousal and valence dimensions through differential equations, forming a closed-loop simulator. Experimentally, the model has been found to have a significantly higher performance than the baseline with respect to the localization of tension peaks as well as the tracking of arousal levels. The temporal causatory relationship between the expectation modulator and end-to-end procedural modeling of musical emotion computation is validated through Granger causality tests.

II. METHODS

A. MUSICAL TENSION FEATURE EXTRACTION AND HIERARCHICAL CODING

Tension-related elements from the musical score and the audio file are identified simultaneously from both pieces of information. MusicXML (Music Extensible Markup Language) produces a MusicXML file that contains the information needed to analyze the score, and note sequences are parsed through the music21 library to analyze melodic contours and rhythmic structures [17-18]. The pitch profile vectors for adjacent pairs of notes indicate the vertical distances (in semitones) that a melody climbs or falls. In addition, the amount of tension that a chord evokes has also been calculated using Lerdahl's tonality pitch space model [19-20]: the method of mapping the pitches of all degrees in a chord to the basic spatial hierarchy is provided by using the local tonality centre as the starting point and measuring tension with the shortest path step from the centre of the chord to the tonality centre. A rhythm density curve can also be formed for each measure by analysing statistically the occurrence of each event in terms of its frequency of occurrence (coefficient of variation) or amount of segmentation complexity within an event. The audio end is framed with a frame length of 100ms and an overlap rate of 50%, and the acoustic parameters are extracted frame by frame using the Librosa library [21-22]. Loudness fluctuation is defined as the absolute value of the first difference $|\Delta E(t)|$ of the frame-level root mean square energy $E(t)$, capturing the instantaneous contribution of sudden volume changes to tension. Spectral roughness is calculated using the Vassilakis model [23]: for any two frequency components f_m and f_n in the critical band, the roughness contribution is $R_{mn} = (A_m A_n)^{0.1} \cdot g(f_{mn})$, where A_m and A_n are the amplitudes, f_{mn} is the frequency difference, and $g(\cdot)$ is the frequency difference weighting function. The entire roughness spectrum is calculated by combining all of the roughness spectra of all of the roughness pairs that are obtained from the analysis. The dissonance shown in dissonance function is from the Plomp-Levelt pure tone dissonance curve, and a weighted integral of each of the component pairs obtained through sinusoidal harmonic decomposition from the current frame can be conducted. The weighting of the integral is based solely on the critical bandwidth and the perceived dissonance threshold for pure tones as described in references [24-25]. Both musical notation and audio features are downsampled from their respective sources to 10Hz to ensure temporal alignment. To mitigate potential information loss—particularly the erosion of audio transients and their logical coupling with symbolic semantics—a cross-domain synchronization check was performed. The alignment preserves the primary rhythmic structure while the 10Hz resolution maintains sufficient granularity for modeling emotional response trajectories without significant loss of critical tension markers. This aligned 12-dimensional fused feature vector serves as the foundation for subsequent tension evolution modeling.

B. TENSOR QUANTIZATION MODEL BASED ON COGNITIVE EXPECTATION

Using MIDI (Musical Instrument Digital Interface) pitch sequences as input, the instantaneous surprise level is calculated per event using the IDyOM framework in combination with Transformer-XL [26-27]. Transformer-XL is pre-trained on the MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) dataset with a note prediction task. Its segment-level recursive mechanism allows the current segment to reuse the hidden state of the previous segment, enabling long-term dependency modeling for large-scale formal structures such as sonata form [28-29]. For the t-th note event, the model outputs a conditional probability distribution $P(e_t \mid e_{<t})$, and then calculates the Shannon surprise value $S_t = -\log_2 P(e_t \mid e_{<t})$.

The larger the value of S_t, the lower the probability of the note appearing in the given historical context, and the higher the degree of violation of cognitive expectation. The surprise value sequence is concatenated with a 12-dimensional feature vector and input into a gated recurrent unit network to recursively estimate the tension evolution state [30]. The GRU hidden state h_t encodes the accumulated tension level up to the current time. Its update rule introduces an expectation satisfaction modulation term to reflect the modulating effect of cognitive expectation on the tension accumulation rate. Specifically, the update gate z_t is calculated as follows:

$$z_t = \sigma\left(W_z[h_{t-1}, S_t] + b_z + \lambda \cdot I[P(e_t) > \theta]\right) \quad (1)$$

In Equation (1), W_z is the trainable weight matrix; b_z is the bias vector; $\sigma(\cdot)$ is the sigmoid activation function. λ is the learnable scalar modulation intensity parameter; $I[\cdot]$ is the indicator function; θ is the preset probability threshold (the probability value corresponding to the median of surprise on the validation set). When the predicted probability $P(e_t)$ is higher than θ, the expectation being satisfied weakens the activation of the update gate at the current moment, thereby reducing the tension accumulation rate. The reset gate r_t and the candidate state $\tilde{h}_t$ follow the standard GRU form, and the final hidden state update is $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$. $\odot$ represents element-wise multiplication.

The continuous tension trajectory is compressed from the hidden state to the interval [-1, 1] through linear projection and hyperbolic tangent. Long-range dependency capability is guaranteed by the inter-segment state transfer of Transformer-XL, while GRU focuses on the recursive accumulation of short-to-medium-range tension, jointly modeling multi-level expectation violations.

C. TENSION-EMOTION DYNAMIC RESPONSE COUPLED SIMULATOR

This paper formalizes the process modeling of the constructed coupled simulator using a continuous-time dynamic system. The system state vector is defined as $s(t) = [T(t), A(t)]^T$, where $T(t)$ is the continuous tension trajectory and $A(t)$ is the arousal level; the external input vector is $u(t) = [S(t), x_t^T, \beta_t]^T$, which includes the Shannon surprise $S(t)$, the 12-dimensional fusion feature x_t, and the adjustment factor β_t.

$$\dot{s}(t) = f(s(t), u(t); \theta),$$

where θ is the dynamic parameter. The system evolution is described by the state equation; the observation equation $y(t) = g(s(t))$ outputs the emotional intensity. This form satisfies the three core properties of process modeling: temporal continuity, state memory, and causal closed loop. The continuous tension trajectory $T(t)$ is used as the driving source to construct the differential equation for the emotional state evolution. The tension trajectory $T(t)$ itself also follows a first-order dynamic process with an inertial term to avoid the instantaneous jumps in the output of the gated loop unit (GRU) from causing false high-frequency fluctuations in the emotional drive. The tension evolution equation is defined as:

$$\frac{dT(t)}{dt} = -\kappa \cdot (T(t) - T_{\mathrm{GRU}}(t)) + \eta \cdot S(t) \quad (2)$$

In Equation (2), $T_{\mathrm{GRU}}(t)$ is the instantaneous tension estimate output by the GRU network at time t (compressed to [-1, 1] by linear projection and hyperbolic tangent); κ is the tension regression rate, controlling the strength of the convergence of $T(t)$ to $T_{\mathrm{GRU}}(t)$; $S(t)$ is the Shannon surprise of the current frame, and η is the direct driving gain of the surprise on the tension change rate. This equation treats tension as a first-order state variable with memory and inertia, rather than an instantaneous static mapping of the GRU. The dynamics of arousal $A(t)$ follow a damped first-order drive response process:

$$\frac{dA(t)}{dt} = \alpha \cdot \frac{dT(t)}{dt} - \gamma \cdot A(t) \quad (3)$$

In Equation (3), α is the positive coupling gain from the rate of change of tension to arousal, which characterizes the sensitivity of the sympathetic branch of the listener's autonomic nervous system to tension fluctuations; γ is the baseline damping coefficient of emotional regression, which simulates the rate at which arousal decays exponentially to a neutral state ($A = 0$) when there is no new tension drive, and its physiological correspondence is the recovery effect of the parasympathetic nervous system. The valence dimension is also constructed as a first-order dynamic system with inertia to avoid emotional jumps caused by frame-level regression [31]. The valence evolution equation is defined as:

$$\begin{aligned}\frac{dV(t)}{dt} = \delta \cdot (V_{\mathrm{base}}(t) - V(t)) + \xi \cdot \frac{dT(t)}{dt} \\ + \zeta \cdot A(t) \cdot (1 - |V(t)|)\end{aligned} \quad (4)$$

In Equation (4), $V_{\mathrm{base}}(t)$ is the tonality-modality baseline valence. While initially grounded in the Gomez model (+0.3 for major, -0.2 for minor), this baseline is dynamically modulated by real-time rhythmic density and tempo-arousal interactions. This dynamic adjustment accounts for musical contexts where mode-valence congruency varies, such as high-tempo minor passages evoking positive valence, thereby resolving the logical limitations of a purely static baseline in complex emotional transitions. The resulting value is further adjusted by pitch contour weighting within the range of [-1, 1]. δ is the valence regression rate. ξ is the driving coefficient of the tension change rate on valence. ζ is the arousal-valence cross-coupling coefficient. This equation ensures that the valence response has temporal inertia and together with the arousal equation, and constitutes a two-dimensional emotional dynamics system.

To characterize the modulation of the emotional response amplitude by expectation, a conditional variational autoencoder (CVAE) is introduced. This CVAE uses 10 seconds of music context features as conditional input and outputs a latent variable z through the encoder, and the decoder is used to predict the surprise sequence frame by frame within the corresponding time period [32-33]. The latent variable captures the prior distribution of the listener's sensitivity to subsequent tension events. During simulation, the normalized difference between the surprise predicted by the CVAE decoder and the actual calculated surprise is used as a modulating factor β_t, which amplifies or suppresses the tension driving efficiency. The revised arousal equation is:

$$\frac{dA(t)}{dt} = \alpha \cdot \frac{dT(t)}{dt} \cdot (1 + \beta_t) - \gamma \cdot A(t) \tag{5}$$

Valence and arousal are updated by Euler numerical integration at each step. Closed-loop feedback is reflected in the emotional state vector output in the current frame as part of the CVAE conditional input for the next time step, enabling the simulator to continuously correct subsequent emotional trajectories in the perception-expectation-response loop [34-35]. This coupled framework unifies bottom-up acoustic tension driving and top-down cognitive expectation adjustment into the same differential dynamics system. The simulator architecture diagram is shown in Figure 1.

Figure 1 illustrates the closed-loop simulation process from the continuous tension trajectory to the evolution of emotional state. The tension trajectory drives the differential response of arousal and the baseline calculation of valence, respectively. The conditional variational autoencoder extracts moderating factors from the musical context to dynamically correct the driving efficiency of tension on arousal. The corrected arousal and valence form the emotional state and influence subsequent moderating processes through feedback loops. Similar to arousal, the valence dimension also evolves dynamically through the aforementioned differential equations, rather than through static regression, thus ensuring the consistency of the emotional response process.

D. NUMERICAL SOLUTION AND PARAMETER OPTIMIZATION OF COUPLED SIMULATOR

A fourth-order Runge-Kutta method is used to numerically integrate the arousal differential equation with an integration step size of 10 Hz (consistent with the audio frame rate). Model training consists of two phases: the first phase involves independent training of a conditional variational autoencoder (CVAE), inputting 10 seconds of contextual features (including a 12-dimensional fused tension descriptor and a preceding surprise sequence), with the output being the surprise sequence for the corresponding time period. The loss function is a weighted sum of the reconstruction mean square error and the KL (Kullback-Leibler) divergence. The second phase involves joint optimization of the GRU tension network and coupling coefficients, freezing the CVAE encoder weights, calculating the adjustment factor online, and using a loss function that combined the simulated arousal with the labeled mean square error plus a tension trajectory smoothing regularization term. The optimizer used is AdamW (Adam with Weight Decay) (weight decay of 0.01), with a cosine annealing learning rate, 150 training epochs, and a batch size of 16. All dynamic parameters are automatically differentiated and backpropagated using the Runge-Kutta solver, and updated synchronously with the GRU network weights.

III. EXPERIMENTS AND VERIFICATION

A. EXPERIMENTAL DESIGN

744 complete music clips labeled in 2013 from the DEAM dataset are selected, covering eight Western music genres: blues, classical, country, electronic, folk, jazz, pop, and rock. Each audio clip is 45 seconds long and includes time series of valence and arousal continuously labeled by multiple listeners. The original sampling rate is 2 Hz. To align the 10 Hz simulation outputs with the 2 Hz ground truth labels for metric calculation, a decimation filter (low-pass filtering followed by downsampling) was applied to the simulated trajectories to prevent temporal aliasing errors, ensuring a rigorous frame-by-frame comparison for RMSE and Pearson r calculations.

The dataset is randomly stratified in an 8:1:1 ratio, based on eight genre labels, to ensure that the genre distribution of each subset is consistent with the whole dataset. The training set is used for model parameter optimization; the validation set is used for hyperparameter selection and early stopping detection; the test set is only used for the final performance report.

Three baseline models are established for comparative evaluation. Baseline 1: Baseline 1 is a pure audio feature linear regression, with input consisting only of six acoustic features including loudness fluctuation, spectral roughness, and dissonance. Ridge regression is used to fit the arousal sequence. Baseline 2: Baseline 2 is a standard GRU tension model, removing the expected satisfaction modulation term from the update gate. The remaining structure is consistent with the method in this paper, used to examine the contribution of the modulation term to the tension fitting accuracy. Baseline 3: Baseline 3 is an LSTM (Long Short-Term Memory) fully connected direct mapping model, inputting 12-dimensional fused features into a two-layer LSTM. The hidden state directly outputs arousal through a linear layer, without explicit intermediate tension representation, used to verify the necessity of intermediate tension modeling.

B. EVALUATION OF THE FITTING ACCURACY OF THE TENSION EVOLUTION TRAJECTORY

The true value of tension is obtained using audio clips from the same source as DEAM. Five music experts are invited to annotate the tension intensity frame by frame at a sampling rate of 10Hz, and the average value of the five experts is

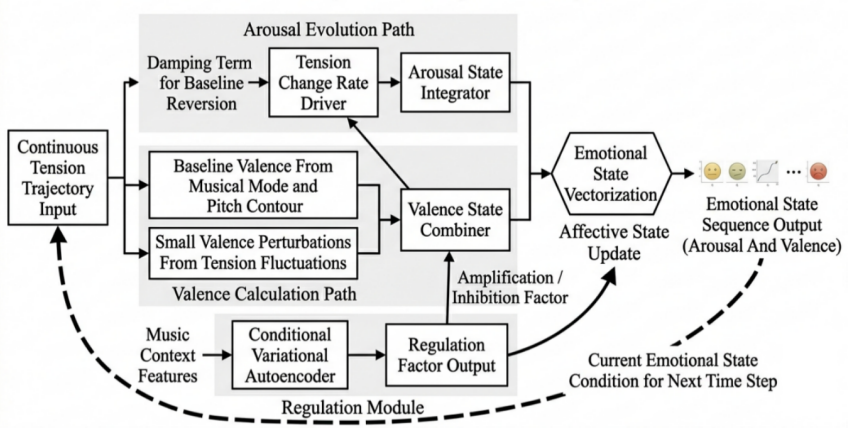

FIGURE 1. Architecture diagram of the tension-emotion dynamic response coupled simulator

taken as a reference. The model output tension is linearly mapped to [0,1] and then aligned. Three metrics are used for evaluation: Pearson correlation coefficient (calculated frame by frame as the ratio of the product of covariance and standard deviation, measuring the linear consistency of the trend; the correlation coefficients of all segments are reported after Fisher z-transformation, mean, and inverse transform), dynamic time warping distance (using FastDTW (Fast Dynamic Time Warping) for nonlinear alignment, calculating the cumulative Euclidean distance of the warped path; the smaller the distance, the tighter the sequential matching; the mean and standard deviation of all segments are reported), and peak position absolute error (detecting local maxima; peak height threshold ≥ 0.3 times the maximum value, peak spacing ≥ 2 seconds; the mean error of the global highest peak or the top three peaks is taken; the mean and median error of the test set are reported). The three metrics can be used together to assess how well the tension trajectory fits its amplitude trend, temporal structure, and climax positioning as illustrated in Figure 2.

Figure 2 illustrates the three dimensions of the fitting accuracy evaluation of the tension evolution trajectory. To compare the trend consistency, Figure 2(a) uses the Pearson correlation coefficient. With a mean value of 0.82 for the Proposed Method and a mean of 0.71 for Baseline 2, the Proposed Method has a higher numerical value than the Baseline 2 method. Figure 2(b) utilizes the dynamic time warping distance to evaluate the temporal structure matching degree. The mean for the Proposed Method is 12.3, and the mean for Baseline 2 is 18.7, indicating that the distance value of the Proposed Method is lower than that of the Baseline 2 method. Figure 2(c) employs the absolute error of the peak position (climax positioning accuracy) to reflect how accurate the peak position is. The mean and median for both methods are 0.42 seconds for the Proposed Method and 0.78 seconds for Baseline 2. The combined data from all three subfigures presents the model's fitting performance regarding amplitude trend, temporal alignment, and climax positioning.

C. DYNAMIC TRACKING PERFORMANCE EVALUATION OF EMOTIONAL AROUSAL RESPONSE

To measure the arousal generated by the simulator against the DEAM-annotated arousal metric, two metrics are devised to evaluate the performance of the simulator. The first metric used is the root mean square error (RMSE), which is determined by taking the square root of the sum of squares of discrepancies - that is, the difference between predicted and true values - over all frames, divided by the number of frames. A smaller RMSE value represents that simulated trajectories deviate less from truth; therefore both the mean and standard deviation of all segment RMSE values can be presented. The second metric is the Kendall τ-b coefficient, which assesses the direction of first order differences - rising, falling, or zero - for both simulated and annotated sequences and uses the number of frames with consistent, conflicting and unidirectional changes to calculate the coefficient. If the coefficient is positive, this implies that there is a positive correlation in terms of direction of change; therefore, the closer to 1 the value is, the more successfully the emotion fluctuation timing is captured. In combination, these two metrics evaluate tracking from the amplitude fidelity and direction consistency perspectives. When used together, they prevent issues such as having a low RMSE with inherent random swings in direction and a high Kendall τ-b coefficient with systematic amplitude differences. All metrics are calculated and summarized on a segment-by-segment basis in the test set. The dynamic tracking performance of emotional arousal is shown in Figure 3.

Figure 3 shows the evaluation results of the dynamic tracking performance of emotional arousal, comparing the four methods using two metrics: root mean square error (RMSE) and Kendall's τ-b. For RMSE, the proposed method has a mean of 0.152 and a standard deviation of 0.023; Baseline 1 has a mean of 0.284 and a standard deviation of 0.041; Baseline 2 has a mean of 0.198 and a standard deviation of 0.035; and Baseline 3 has a mean of 0.176 and a standard deviation of 0.029. For Kendall's τ-b, the proposed method has a mean of 0.74 and a standard

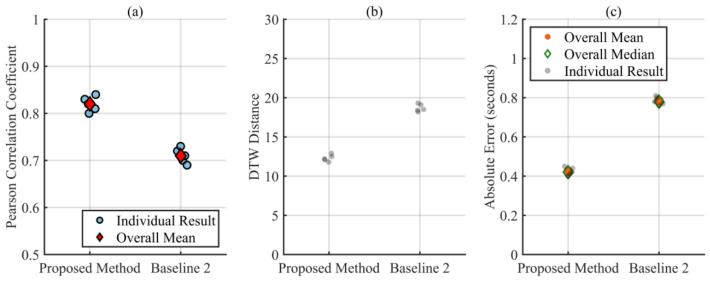

FIGURE 2. Accuracy assessment of tension evolution trajectory fitting: (a) Pearson correlation coefficient; (b) Dynamic time warping distance; (c) Absolute error of peak position

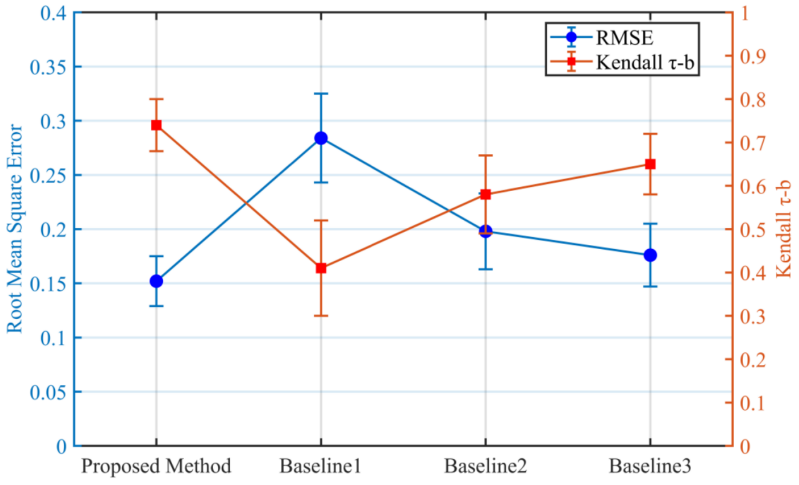

FIGURE 3. Dynamic tracking performance of emotional arousal

deviation of 0.06; Baseline 1 has a mean of 0.41 and a standard deviation of 0.11; Baseline 2 has a mean of 0.58 and a standard deviation of 0.09; and Baseline 3 has a mean of 0.65 and a standard deviation of 0.07. When evaluated on the numerical distribution metric, the Proposed Method exhibits numerically superior performance on both RMSE and Kendall's τ-b. The mean RMSE values for the Proposed Method are lower than all baseline methods, while it also exhibits the highest average Kendall's τ-b. Furthermore, the standard deviations for both metrics are relatively small, indicating more stable performance across the various test segments of the proposed method than shown by any of the baseline methods. Baseline 1 has gaps in both metrics when compared against the other methods, which illustrates how various modeling strategies affect the accuracy of sentiment being tracked.

D. DYNAMIC TRACKING AND EVALUATION OF THE VALENCE DIMENSION

The valence sequence output by the simulator is compared point-by-point with the DEAM-annotated true valence values at a frame rate of 10 Hz. Valence calculation employs a Gomez modal valence regression model to determine the baseline value: for each audio frame, the local modality (major/minor) and pitch contour are extracted, weighted, and superimposed to obtain the baseline valence, followed by a linear perturbation term for tension fluctuations (the perturbation coefficient is calibrated to 0.12 on the validation set). The evaluation metrics used are root mean square error (RMSE) and Pearson correlation coefficient (Pearson r). RMSE measures the absolute deviation between the simulated valence and the annotated true value; Pearson r measures the linear consistency of their trends. The RMSE and Pearson correlation coefficient are shown in Table 1.

TABLE 1. Root mean square error of valence tracking for each school of thought and Pearson correlation coefficient

Methods	RMSE mean	RMSE std	Pearson r mean	Pearson r std
Proposed Method	0.176	0.031	0.68	0.09
Baseline 1	0.312	0.048	0.39	0.14
Baseline 2	0.234	0.042	0.51	0.12
Baseline 3	0.208	0.036	0.59	0.10

Table 1 shows that the proposed method outperforms the three baselines in valence tracking with a mean RMSE of 0.176 and a mean Pearson r of 0.68, exhibiting the smallest standard deviation and indicating the best accuracy and stability. This performance stems from the influence of abstract structures such as mode and harmony on valence.

The proposed method effectively simulates valence inertia by dynamically correcting tension driven by the expected modulation factor of CVAE and combining it with the differential equation of the tonality baseline. In contrast, the baseline method lacks explicit cognitive regulation and dynamic coupling.

E. TEMPORAL CAUSAL VERIFICATION OF EXPECTED REGULATORY PATHWAYS

To verify whether the moderating factors output by CVAE constitute a temporal cause of the improvement in emotion simulation, the moderating factor sequence and simulation error sequence are extracted from each audio segment in the test set and aligned at 10 Hz to form a binary time series. The Granger causality test is used to calculate the F-statistic. The test is performed on 74 audio segments, and the number and proportion of segments with $p < 0.05$ are counted. A repeated test is conducted by randomly shuffling the temporal order of the moderating factor sequences to rule out spurious causal associations. The distribution of the F-statistic and the proportion of significant segments are shown in Figure 4.

Figure 4 shows the results of the Granger causality test. As shown in Figure 4(a), the number of segments exhibits a relatively concentrated distribution in the medium F-statistic intervals. The interval [3.0, 4.5] contains 20 segments, the most of any interval; the interval [4.5, 6.0] has 10 segments; and the intervals [6.0, 7.5] and [7.5, 9.0] each have 8 segments, totaling 26 segments, constituting the main part of the distribution. In the low-value intervals, the interval [0.0, 1.5] has only 2 segments, and the interval [1.5, 3.0] has 7 segments, indicating that the proportion of segments with weaker temporal explanatory power of the moderating factor for simulation errors is relatively small. The high-value intervals are relatively dispersed, with 3 segments in the [9.0, 10.5] interval and 7 segments in the [12.0, 13.5] interval. Higher intervals such as [16.5, 18.0], [18.0, 19.5], [19.5, 21.0], [21.0, 22.5], and [24.0, 25.5] have 2, 1, 1, 2, and 3 segments respectively, exhibiting a sparse but persistent tail characteristic. Overall, the number of segments initially increases with increasing F-statistics, then fluctuates and decreases, with segments exhibiting moderate causal relationships making up the majority. Figure 4(b) shows that in 74 test audio segments, 67.6% of the segments have a Granger test $p < 0.05$; after randomly shuffling the moderating factor sequence, the proportion of significant segments decreases to 5.4%, indicating a significant statistical precedence and predictive robustness of the moderating factors regarding simulation error fluctuations.

F. CROSS-VALIDATION OF MODEL COUPLING EFFECTIVENESS

To quantify the overall predictive stability of the coupled framework, the Mean Absolute Percentage Error (MAPE) is used as the cross-validation metric. The absolute difference between the simulated wakefulness and the DEAM labeled value is calculated frame by frame, divided by the absolute value of the label. Frames with labeled values below 0.05 are discarded (to avoid division-by-zero singularities). The percentage errors of the valid frames are summed, divided by the number of valid frames, and then multiplied by 100% to obtain the MAPE for each audio segment (45 seconds/450 frames). The experiment is repeated five times. The model coupling performance is shown in Figure 5.

Figure 5 shows the comparison of mean absolute percentage errors of the coupling effectiveness of the four models. The proposed method has a mean error of 8.2%, a standard deviation of 0.2%, a minimum of 7.9%, a maximum of 8.5%, and a median of 8.2%. Baseline 3 has a mean error of 11.9%, a standard deviation of 0.3%, a range of 11.5% to 12.3%, and a median of 11.9%. Baseline 2 has a mean error of 15.7%, a standard deviation of 0.7%, a minimum of 14.8%, a maximum of 16.5%, and a median of 15.9%. Baseline 1 has a mean error of 23.1%, a standard deviation of 1.0%, a range of 21.9% to 24.2%, and a median of 23.1%. From the perspective of mean indicators, the error values of the proposed method are lower than those of the three baseline models, and the medians of each method are basically consistent with the mean, reflecting a relatively symmetrical error distribution. The standard deviation of the proposed method (0.2%) is lower than the standard deviation of Baseline 1 (1.0%) and Baseline 2 (0.7%), and its level of error (0.6%) is much smaller than that of Baseline 1 (2.3%), indicating that predictions with the proposed method fluctuate less than predictions when using the other two methods.

G. SIMULATION ROBUSTNESS AND GENERALIZATION ABILITY TEST

The DEAM genre label divides the test set into eight different sets. Additionally, the root mean square error (RMSE) of simulated arousal and labeled arousal for each genre has been calculated using frame-by-frame analysis of the model and its capabilities to track each particular style of music (i.e., the model's is not as accurate tracking all genres equally). The eight genres mean RMSE values and their standard deviations are recorded to assist in determining systematic errors within certain styles of music so that this can also be used to measure the relationship between specific music genre acoustic features or tension patterns as well as by evaluating the transfer ability of the combined simulation framework on new styles of music through the mean of the RMSE results represents the overall result for this study. The mean RMSE values for each genre are displayed in Table 2.

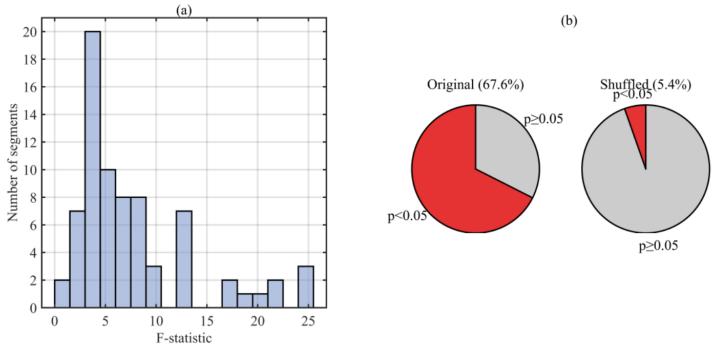

FIGURE 4. Granger causality test results for the expected regulation pathway: (a) Distribution of F-statistic; (b) Significant fragment proportions

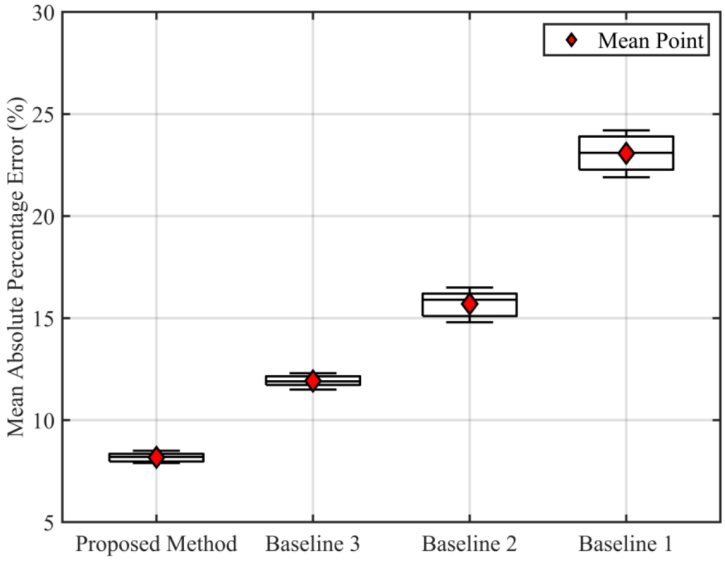

FIGURE 5. Comparison of model coupling effectiveness

TABLE 2. Mean RMSE values of each method across different music genres

Genre	Proposed Method	Baseline 1	Baseline 2	Baseline 3
Blues	0.168	0.312	0.224	0.195
Classical	0.145	0.285	0.201	0.172
Country	0.159	0.298	0.215	0.188
Electronic	0.172	0.335	0.241	0.206
Folk	0.152	0.290	0.208	0.179
Jazz	0.185	0.358	0.258	0.221
Pop	0.148	0.278	0.198	0.170
Rock	0.175	0.342	0.249	0.213

According to the Mean Root Mean Square Error (RMSE) in Table 2, a lower value is better. This table shows the trajectory accuracy of the four methods in predicting the emotional arousal of eight music genres. The Proposed Method exhibits the lowest errors across all genres: Classical (0.145), Pop (0.148), Folk (0.152), and Jazz (0.185). Overall, Baseline 1 displays the largest error values of the four methods, and has errors that range from 0.278 for Pop to 0.358 for Jazz. There is a range of error values for Baseline 2 and Baseline 3; Baseline 3 (for example, Classical 0.172, Pop 0.170) performs slightly better than Baseline 2 (for example, Classical 0.201, Pop 0.198), but still does not match the accuracy of the Proposed Method. Generally, Classical and Pop all have lower than average calculations across most of the methods, while Jazz and Electronic music perform with relatively larger than average calculations, supporting the fact that there are differences in the level of challenge involved in modeling the emotions for various styles of music. The standard deviation of the RMSE is provided in Table 3.

TABLE 3. Standard deviation of RMSE for each method across different music genres

Genre	Proposed Method	Baseline 1	Baseline 2	Baseline 3
Blues	0.021	0.045	0.038	0.029
Classical	0.018	0.041	0.032	0.024
Country	0.020	0.043	0.035	0.027
Electronic	0.023	0.049	0.041	0.031
Folk	0.019	0.042	0.034	0.025
Jazz	0.026	0.053	0.044	0.034
Pop	0.018	0.040	0.031	0.023
Rock	0.024	0.050	0.042	0.032

Table 3 shows the root mean square error standard deviation of arousal prediction for the four methods across eight music genres, used to measure the differences in prediction stability of each model across different music segments. From the data distribution, the proposed method generally performs worse than the three baselines across all genres. For example, in Blues, the proposed method has a standard deviation of 0.021, compared to 0.045 for Baseline 1, 0.038 for Baseline 2, and 0.029 for Baseline 3. In Classical and Pop, the proposed method has a standard deviation of 0.018, which is relatively low compared to all genres. In Jazz, the standard deviations of the methods are relatively high, with the proposed method at 0.026 and Baseline 1 reaching 0.053. Overall, the standard deviation of RMSE for the proposed method ranges from 0.018 to 0.026, while Baseline 1 ranges from 0.040 to 0.053. Baseline 2 and Baseline 3 range from 0.031 to 0.044 and 0.023 to 0.034, respectively, reflecting the differences in error fluctuation control among different modeling strategies.

IV. CONCLUSION

A coupled model of musical tension evolution and emotional response is created from a combination of cognitive expectation and musical properties. A 12-dimensional tension descriptor is obtained from musical score and audio channels. Note-by-note surprise is estimated using the IDyOM framework enhanced with Transformer-XL, so long-range dependencies across measures can be captured. Finally, a gated recurrent unit with a term for expectation satisfaction modulation, subject to recursive estimation, is used to estimate the continuous tension trajectory, thereby quantifying the multi-scale accumulation of tension. Finally, a conditional variational autoencoder generates a modulation factor, and a damped differential equation couples the rate of change of tension to the arousal dimension of the Russell emotional space, forming a perception-expectation-response closed-loop simulator. Experimental results show that the absolute error of the model's peak position is 0.42 seconds; the root mean square error of arousal tracking is 0.152; the Granger causality test confirms that the expectation modulation pathway has a temporal causal effect on improving simulation accuracy. This paper provides a new perspective for end-to-end procedural modeling of musical emotion computing, from symbolic expectation to physiological response.

AUTHOR CONTRIBUTIONS

Kexia Huang designed, collected and analyzed the data, and drafted the manuscript. Kexia Huang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Kexia Huang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Kern, M. Heilbron, F. P. de Lange, and E. Spaak, "Cortical activity during naturalistic music listening reflects short-range predictions based on long-term experience," elife, vol. 11, no. 1, Art. no. e80935, 2022, doi: 10.7554/eLife.80935.

[2] X. Wang, L. Wang, and L. Xie, "Comparison and analysis of acoustic features of Western and Chinese classical music emotion recognition based on VA model," Applied Sciences, vol. 12, no. 12, pp. 5787-5799, 2022, doi: 10.3390/app12125787.

[3] V. K. M. Cheung, P. M. C. Harrison, S. Koelsch, M. T. Pearce, A. D. Friederici, and L. Meyer, "Cognitive and sensory expectations independently shape musical expectancy and pleasure," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 379, no. 1895, pp. 1-16, 2024, doi: 10.31234/osf.io/z76hg_v1.

[4] N. Singer, N. Jacoby, T. Hendler, and R. Granot, "Feeling the beat: Temporal predictability is associated with ongoing changes in music-induced pleasantness," Journal of Cognition, vol. 6, no. 1, pp. 34-51, 2023, doi: 10.5334/joc.286.

[5] N. Zhang, L. Sun, Q. Wu, and Y. Yang, "Tension experience induced by tonal and melodic shift at music phrase boundaries," Scientific Reports, vol. 12, no. 1, pp. 8304-8315, 2022, doi: 10.1038/s41598-022-11949-4.

[6] K. Basiński, D. R. Quiroga-Martinez, and P. Vuust, "Temporal hierarchies in the predictive processing of melody-From pure tones to songs," Neuroscience & Biobehavioral Reviews, vol. 145, no. 1, pp. 105007-105023, 2023, doi: 10.31234/osf.io/nua2j.

[7] A. S. Turrell, A. R. Halpern, K. Bannister, D. Chai-Wi-Ting, and A. H. Javadi, "Building the anticipation: How variation in tension mediates emotions in music," Music Perception: An Interdisciplinary Journal, vol. 42, no. 3, pp. 256-268, 2025, doi: 10.1525/ mp.2024.aa004.

[8] Y. Shang, Q. Peng, Z. Wu, and Y. Liu, "Music-induced emotion flow modeling by ENMI Network," PLoS One, vol. 19, no. 10, Art. no. e0297712, 2024, doi: 10.1371/journal.pone.0297712.

[9] S. You, L. Sun, and Y. Yang, "The effects of contextual certainty on tension induction and resolution," Cognitive Neurodynamics, vol. 17, no. 1, pp. 191-201, 2023, doi: 10.1007/s11571-022-09810-5.

[10] T. Daikoku, M. Tanaka, and S. Yamawaki, "Bodily maps of uncertainty and surprise in musical chord progression and the underlying emotional response," Iscience, vol. 27, no. 4, pp. 109498-109510, 2024, doi: 10.1016/j.isci.2024.109498.

[11] A. V. Barchet, J. M. Rimmele, and C. Pelofi, "TenseMusic: An automatic prediction model for musical tension," Plos one, vol. 19, no. 1, Art. no. e0296385, 2024, doi: 10.31234/osf.io/xck3w.

[12] G. Marion, F. Gao, B. P. Gold, G. M. Di Liberto, and S. Shamma, "IDyOMpy: A new Python-based model for the statistical analysis of musical expectations," Journal of Neuroscience Methods, vol. 415, no. 1, Art. no. 110347, 2025, doi: 10.1016/j.jneumeth.2024.110347 .
[13] X. Guan, Z. Ren, and C. Pelofi, "py2lispIDyOM: A Python package for the information dynamics of music (IDyOM) model," Journal of Open Source Software, vol. 7, no. 79, pp. 4738-4753, 2022, doi: 10.21105/joss.04738.
[14] R. Orjesek, R. Jarina, and M. Chmulik, "End-to-end music emotion variation detection using iteratively reconstructed deep features," Multimedia Tools and Applications, vol. 81, no. 4, pp. 5017-5031, 2022, doi: 10.1007/s11042-021-11584-7.
[15] P. Bodily and D. Ventura, "Steerable music generation which satisfies long-range dependency constraints," Transactions of the International Society for Music Information Retrieval, vol. 5, no. 1, pp. 71-86, 2022, doi: 10.5334/tismir.97.
[16] M. R. Bjare, S. Lattner, and G. Widmer, "Differentiable short-term models for efficient online learning and prediction in monophonic music," Transactions of the International Society for Music Information Retrieval, vol. 5, no. 1, pp. 190-206, 2022, doi: 10.5334/tismir.123.
[17] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Late multimodal fusion for image and audio music transcription," Expert Systems with Applications, vol. 216, no. 1, Art. no. 119491, 2023, doi: 10.1016/j.eswa.2022.119491.
[18] W. Yu, "Music source feature extraction based on improved attention mechanism and phase feature," Systems and Soft Computing, vol. 6, no. 1, Art. no. 200149, 2024, doi: 10.1016/j.sasc.2024.200149.
[19] M. Navarro-Cáceres, M. Caetano, G. Bernardes, M. Sánchez-Barba, and J. Merchán Sánchez-Jara, "A computational model of tonal tension profile of chord progressions in the tonal interval space," Entropy, vol. 22, no. 11, pp. 1291-1302, 2020, doi: 10.3390/e22111291.
[20] F. Zhu, C. Wu, Q. Huang, N. Zhu, and T. Leng, "Rhythm-Based Attention Analysis: A Comprehensive Model for Music Hierarchy," Applied Sciences, vol. 15, no. 11, pp. 6139-6152, 2025, doi: 10.3390/app15116139.
[21] M. S. M. Mendjel, S. Ghazi, A. Dib, and H. Seridi, "A new audio approach based on user preferences analysis to enhance music recommendations," Revue d'Intelligence Artificielle, vol. 37, no. 5, pp. 1341-1349, 2023, doi: 10.18280/ria.370527.
[22] N. He and S. Ferguson, "Music emotion recognition based on segment-level two-stage learning," International Journal of Multimedia Information Retrieval, vol. 11, no. 3, pp. 383-394, 2022, doi: 10.1007/s13735-022-00230-z.
[23] A. J. Milne, E. A. Smit, H. S. Sarvasy, and R. T. Dean, "Evidence for a universal association of auditory roughness with musical stability," PLoS One, vol. 18, no. 9, Art. no. e0291642, 2023, doi: 10.1371/journal.pone.0291642.
[24] R. Marjieh, P. M. C. Harrison, H. Lee, F. Deligiannaki, and N. Jacoby, "Timbral effects on consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales," Nature Communications, vol. 15, no. 1, pp. 1482-1501, 2024, doi: 10.1038/s41467-024-45812-z.
[25] V. Etxebarria, "Dissonance, Sound Spectrum and Musical Scale for Ancient Idiophones and Aerophones," Journal of Mathematics and Music, vol. 1, no. 1, pp. 1-15, 2025, doi: 10.1080/17459737.2025.2560922 .
[26] J. Min, Z. Gao, and L. Wang, "Application and research of music generation system based on cvae and Transformer-XL in video background music," IEEE Transactions on Industrial Informatics, vol. 21, no. 2, pp. 1409-1418, 2024, doi: 10.1109/TII.2024.3477561.
[27] J. Liang, "Harmonizing minds and machines: survey on transformative power of machine learning in music," Frontiers in Neurorobotics, vol. 17, no. 1, Art. no. 1267561, 2023, doi: 10.3389/fnbot.2023.1267561.
[28] Y. Liang, H. Abudukelimu, J. Chen, A. Abulizi, and W. Guo, "MAML-XL: a symbolic music generation method based on meta-learning and Transformer-XL: Y," Liang et al. Multimedia Systems, vol. 31, no. 3, pp. 206-225, 2025, doi: 10.1007/s00530-025-01803-8.
[29] Z. Li, Q. Huang, X. Yang, Q. Chen, and L. Zhang, "Automatic composition system based on transformer-xl," Applied Sciences, vol. 14, no. 13, pp. 5765-5776, 2024, doi: 10.3390/app14135765.
[30] Y. Zhang, M. Li, and S. Pan, "Deep learning-based emotion recognition algorithms in music performance," Scalable Computing: Practice and Experience, vol. 25, no. 6, pp. 4712-4719, 2024, doi: 10.12694/scpe.v25i6.3286.
[31] N. B. Maimon, D. Lamy, and Z. Eitan, "Do Picardy thirds smile? Tonal hierarchy and tonal valence: Explicit and implicit measures," Music Perception: An Interdisciplinary Journal, vol. 39, no. 5, pp. 443-467, 2022, doi: 10.1525/mp.2022.39.5.443.
[32] S. Ji and X. Yang, "EmoMusicTV: Emotion-conditioned symbolic music generation with hierarchical transformer VAE," IEEE Transactions on Multimedia, vol. 26, no. 1, pp. 1076-1088, 2023, doi: 10.1109/tmm.2023.3276177.
[33] L. Comanducci, D. Gioiosa, M. Zanoni, F. Antonacci, and A. Sarti, "Variational Autoencoders for chord sequence generation conditioned on Western harmonic music complexity," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2023, no. 1, pp. 24-35, 2023, doi: 10.1186/s13636-023-00288-5.
[34] Y. Xin, "MusicEmo: transformer-based intelligent approach towards music emotion generation and recognition," Journal of Ambient Intelligence and Humanized Computing, vol. 15, no. 8, pp. 3107-3117, 2024, doi: 10.1007/s12652-024-04811-0.
[35] S. L. Wu and Y. H. Yang, "MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, no. 1, pp. 1953-1967, 2023, doi: 10.1109/TASLP.2023.3270726 .