Movie Soundtrack Emotion Matching Algorithm Based on CLIP Model

Junyan Yang[1], Xiang Gao[2], Wenxi Zhang[1]
[1]Department of Integrated Arts, Silla University, Busan 46958, South Korea
[2]School of Art and Design, Huaiyin Institute of Technology, Huai'an 223001, Jiangsu, China
Corresponding author: Xiang Gao (e-mail: 15951262609@163.com)

ABSTRACT To address the semantic distortion and temporal misalignment introduced by existing movie soundtrack matching methods that rely on textual intermediaries, this paper proposes an end-to-end audio-visual cross-modal temporal alignment algorithm based on the CLIP model. Considering that robust multimodal signal fusion and temporal synchronization are increasingly important for intelligent information processing in advanced electromagnetic sensing and communication environments, the proposed framework constructs a unified embedding space without textual bridging by extracting visual semantics through CLIP-ViT-B/32 and integrating AudioMAE with BiLSTM to capture deep temporal emotional characteristics of music. Musical representations are directly aligned with visual emotion prototypes to reduce the semantic gap, while a sliding-window temporal attention mechanism enables fine-grained dynamic soft alignment between scene transitions and musical emotional evolution. Furthermore, a multi-dimensional weighted scoring function incorporating semantic consistency, rhythm synchronization, and emotional intensity is designed, and global recommendation sequences are optimized under scene continuity constraints. Experimental results demonstrate that the proposed method achieves an emotion category accuracy of 76.3 $\pm$ 0.6%, an average cross-modal similarity of 0.732 $\pm$ 0.088, and an average temporal alignment deviation of only 1.24 $\pm$ 0.22 s. The framework significantly improves temporal synergy and emotional consistency between visual and audio modalities, providing an effective solution for cross-modal signal alignment and offering valuable methodological insights for multimodal information fusion and intelligent signal processing in electromagnetic wave perception and communication-oriented applications.

INDEX TERMS cross-modal alignment, CLIP model, temporal attention mechanism, multimodal signal processing, electromagnetic information processing

I. INTRODUCTION

WITH the development of intelligent film and television production technology, automated recommendation of film soundtracks has become a key step in improving creative efficiency, a drive towards automation that parallels the increasing reliance on intelligent, real-time decision-making systems to enhance production efficiency and quality control in modern manufacturing [1-3]. Traditional soundtracks rely on manual experience, are costly, and difficult to adapt to the needs of large-scale content production [4,5]. In recent years, emotion matching methods based on multimodal learning have gradually emerged, aiming to automatically match appropriate background music by analyzing the emotions of the picture [6,7]. However, existing methods generally rely on text descriptions as an intermediary bridge between images and audio, resulting in semantic distortion and dynamic alignment deviations in the temporal dimension, making it difficult to accurately capture the fine-grained co-evolution of audio and video emotions [8,9]. Therefore, how to overcome the modal gap and achieve end-to-end dynamic matching of audio and video emotions has become a core issue that needs to be addressed urgently.

The research object of this paper is the key emotional fragments in film narratives and their corresponding background music sequences, covering typical emotional categories such as sadness, tension, joy, tranquility, and excitement. Su Z [10] proposed a hybrid matching model based on feature canonical correlation analysis and fine-grained emotional similarity, which refines its fusion features by analyzing matched and unmatched music-video pairs; In response to the problem of emotional matching between sound and picture, researchers have proposed a variety of solutions. Shen Q [11] used the multimodal model ImageBind to perform preliminary cross-modal retrieval between images and music, and selected images related to music for further processing; Li J [12] proposed a knowledge distillation hash based on CLIP for cross-modal retrieval, which can continuously provide the best results; Yang J [13] proposed an emotion assessment method based on

multimodal information, which achieved an unweighted accuracy of 68.02% and an F1 score of 67.11%. The current research paradigm is trapped in the text-dependent method that sacrifices semantic fidelity, while the method that focuses on time sequence breaks the direct connection between modalities. These two fundamental flaws together lead to the unstable performance of existing systems in real movie scenes, making it difficult to meet the needs of high-quality film and television creation.

This paper makes systematic innovations in three aspects: signal processing, model structure, and training objectives. At the signal processing level, a keyframe selection strategy based on optical flow amplitude difference and brightness threshold is introduced, as well as a pipeline for audio pre-emphasis → STFT → Mel spectrum → logarithmic compression. Continuous wavelet transform (CWT) and optical flow rate of change are used to extract music BPM curves and visual rhythm signals to support rhythm synchronization measurement. In terms of model structure:
(1) Low-rank adaptive (LoRA style) fine-tuning is proposed for the 12th layer MHSA query projection of CLIP-ViT-B/32 to enhance attention to emotionally sensitive areas (facial expression, color temperature, composition) without destroying the backbone weights;
(2) The pre-trained AudioMAE output is time-sequentially aggregated by a single layer BiLSTM and mapped to CLIP through a learnable bilinear projection. The emotional anchor points of text prototype alignment share the embedding space, and orthogonal regularization is introduced to maintain the geometric stability of projection;
(3) A sliding window bidirectional temporal attention mechanism (forward/backward attention fusion, weighted dot product similarity, temperature coefficient dynamically generated by a lightweight network) is designed to achieve end-to-end local soft alignment, replacing the dependence on global DTW/CTC.

In terms of training and evaluation objectives, a differentiable multidimensional emotional matching scoring function (semantic similarity + rhythm synchronization rate + emotional intensity adaptation) is proposed and the sequence-level recommendation is modeled as an MDP with emotional topology and tone penalty. The global optimal path is solved by forward dynamic programming and backtracking, and BPM fluency reward is introduced to encourage smooth transition. The above modules are not redundantly stacked: ablation experiments show that removing the "joint projection space" causes the accuracy to drop from 76.3% to 69.3% (a decrease of 7.0 percentage points) and increases the average alignment deviation from 1.24s to 1.87s (+0.63s); removing the sliding window temporal attention or scene continuity constraints also causes significant decreases in accuracy and alignment performance, respectively, quantitatively demonstrating the substantial contribution of each design in this paper to the final performance.

II. ALGORITHM DESIGN

Visual Semantic Extraction Based on CLIP

Given an input movie video stream $V = \{I_t\}_{t=1}^{T_v}$, where $I_t \in \mathbb{R}^{H\times W\times 3}$ represents the original image of the t-th frame, key frame sampling and preprocessing are first performed. A dynamic frame selection strategy based on the difference in optical flow amplitude is adopted to calculate the normalized optical flow field $\mathrm{F}_{t\to t+1}$ between adjacent frames, and its amplitude mean [14,15] is:

$$\mu_t = \frac{1}{H'W'}\sum_{i,j}\|F_{t\to t+1}(i,j)\|_2 \tag{1}$$

In formula (1), H' and W' are the height and width of the optical flow map, respectively. If $\mu_t > \tau_{\text{flow}}$ (threshold) or it is the starting frame of the scene, the frame is retained; at the same time, black screen frames with a brightness mean lower than τ_{dark} (threshold) are eliminated, and finally a sparse valid frame sequence $\{I_{t_k}\}_{k=1}^{N_{\text{frame}}}$ is obtained. For each frame I_{t_k}, geometric and color normalization is performed: scale to 224×224 resolution and normalize according to ImageNet statistics [16]:

$$\hat{I}_{t_k} = \frac{I_{t_k} - \mu}{\sigma} \tag{2}$$

In formula (2), μ is the mean vector and σ is the standard deviation vector. The normalized image is input into the visual encoder of the pre-trained CLIP-ViT-B/32 model. ViT divides the image input into 32×32 image blocks, linearly embeds them to generate the sequence $\mathbf{x}_0 \in \mathbb{R}^{(N_p+1)\times d}$, $N_p = 49$, $d = 768$, and adds a learnable [CLS] tag vector. The Vision Transformer structure is used for feature extraction. Each layer contains a multi-head self-attention module (Multi-Head Self-Attention, MHSA) and a feed-forward network (Feed-Forward Network, FFN) [17]. The calculation process of the l-th layer is as follows:

$$\mathbf{z}'_l = \mathrm{MHSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1,\ldots,L \tag{3}$$

$$\mathbf{z}_l = \mathrm{FFN}(\mathrm{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \tag{4}$$

In formulas (3) and (4), LN is layer normalization, where MHSA is defined as:

$$\begin{aligned}\mathrm{MHSA}(\mathbf{Z}) &= [\mathrm{head}_1;\ldots;\mathrm{head}_h]W^O,\\ \mathrm{head}_i &= \mathrm{Softmax}\left(\frac{(\mathbf{Z}W_i^Q)(\mathbf{Z}W_i^K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}W_i^V\end{aligned} \tag{5}$$

In formula (5), W_i^Q, W_i^K, W_i^V, and W^O are learnable parameters. The final output feature $\mathbf{z}_L$ is the vector v_{cls} corresponding to [CLS] as the global image semantic representation. To further enhance the model's sensitivity to emotion-related visual features, while freezing the backbone network parameters, low-rank adaptation is introduced only

for the query projection matrix W_{12}^{Q} in the 12th layer of MHSA:

$$\Delta W_{12}^{Q} = AB \tag{6}$$

During the fine-tuning process, only the matrix is updated, so that the attention mechanism can focus more on emotional discriminative areas such as facial expressions, warm and cold tones, and spatial composition. Finally, the adapted visual feature sequence is output for the subsequent cross-modal alignment task, as shown in Figure 1.

Figure 1 shows the visual semantic extraction process based on CLIP-ViT-B/32: first, key frames are sampled from the input video stream, and frames with significant dynamic changes are retained by judging whether they are the scene start frames or the average optical flow amplitude, and black screen frames with an average brightness below 20 are removed to ensure the validity of the input; then the retained frames are preprocessed, scaled to 224×224 and normalized according to the ImageNet standard; after linear embedding, the [CLS] tag is added, and features are extracted layer by layer through the 12-layer Vision Transformer, of which the first 11 layers are standard MHSA and FFN modules, and the 12th layer introduces low-rank adaptation to enhance attention to emotion-related areas; finally, a deep feature sequence containing the [CLS] vector is output.

AUDIO EMOTION REPRESENTATION LEARNING BASED ON AUDIOMAE

Given a candidate soundtrack audio signal a(t), it is first digitally sampled at the sampling rate to obtain a discrete sequence:

$$a[n] = a(t)\big|_{t=n/f_s} \tag{7}$$

This is followed by pre-emphasis to enhance the high frequency content:

$$a_{\text{pre}}[n] = a[n] - \alpha \cdot a[n-1] \tag{8}$$

In formula (8), α is the pre-emphasis coefficient, which controls the degree of high-frequency enhancement. It is then converted to a time-frequency representation using a short-time Fourier transform (STFT) [18,19]. Assuming the window function is a Hamming window $w[n]$, the STFT output of the t-th frame is:

$$X(t,f) = \sum_{n=0}^{N-1} a_{\text{pre}}(tM+n)w[n]e^{-j2\pi fn/N} \tag{9}$$

In formula (9), M is the frame shift, N is the window length, and the amplitude spectrum $|X(t,f)|$ is mapped to the Mel scale to construct the Mel spectrum graph [20]:

$$S_{\text{mel}}(t,m) = \sum_{f=0}^{N/2} |X(t,f)|F_m(f) \tag{10}$$

And logarithmically compressing the result:

$$\hat{S}(t,m) = \log(1 + S_{\text{mel}}(t,m)) \tag{11}$$

This forms the input feature $F_{\text{in}} \in \mathbb{R}^{T' \times 512}$, where T' is the number of time frames. F_{in} is fed into the pre-trained AudioMAE model. This model uses an asymmetric encoder-decoder structure, where the encoder receives only unmasked spectral blocks. The encoding process passes through 12 Transformer layers, resulting in a deep representation. The decoder receives position cues and mask information for the full sequence length and combines this with the encoder output to reconstruct the original spectrum. This paper extracts the [CLS] vector sequence from the last encoder layer to form the original audio semantic sequence c_t. To further aggregate rhythmic and dynamic features, c_t is fed into a single-layer bidirectional LSTM:

$$\begin{aligned} \overrightarrow{h}_t &= \text{LSTM}_{\text{forward}}(c_t, h_{t-1}), \\ \overleftarrow{h}_t &= \text{LSTM}_{\text{backward}}(c_t, h_{t+1}) \end{aligned} \tag{12}$$

After splicing:

$$h_t^{\text{bi}} = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in \mathbb{R}^{512} \tag{13}$$

Finally, it is mapped to the CLIP semantic space through linear transformation:

$$m_t = W_m h_t^{\text{bi}} + b_m, \; W_m \in \mathbb{R}^{768\times512}, \; b_m \in \mathbb{R}^{768} \tag{14}$$

Outputting the music emotion embedding sequence $\{m_t\}_{t=1}^{T_a}$ for subsequent cross-modal alignment.

This paper employs a hybrid emotion space design of "main class prototype + continuous valence - arousal supplement". Subjective data and candidate music libraries are labeled as six emotion sets {sadness, tension, joy, tranquility, excitement, serenity}. The natural language descriptions of these six categories are input into the CLIP text encoder to generate structured emotion prototype vectors, which serve as discrete semantic anchors, facilitating inter-class discrimination learning using InfoNCE loss. Simultaneously, to characterize the intensity and subtle differences of emotions (e.g., fragments all expressing "sadness" but with different arousal levels), each sample is also equipped with continuous V-A annotations (valence ∈ [-1, 1] represents extreme negative → extreme positive, arousal ∈ [0, 1] represents extreme calm → intense). A small MLP is used to map (V, A) to an intensity vector of the same dimension as the CLIP/Audio embedding, which is then incorporated into the final matching score. During training, two objectives are optimized simultaneously: minimizing the InfoNCE contrast loss between the music embedding and its corresponding discrete emotion prototype to ensure semantic alignment; and applying L2 to the V-A mapping. An error term (or a cosine-based intensity consistency constraint) is used to preserve the continuity of emotion intensity. The final scoring function employs learnable weights (tuned on a validation

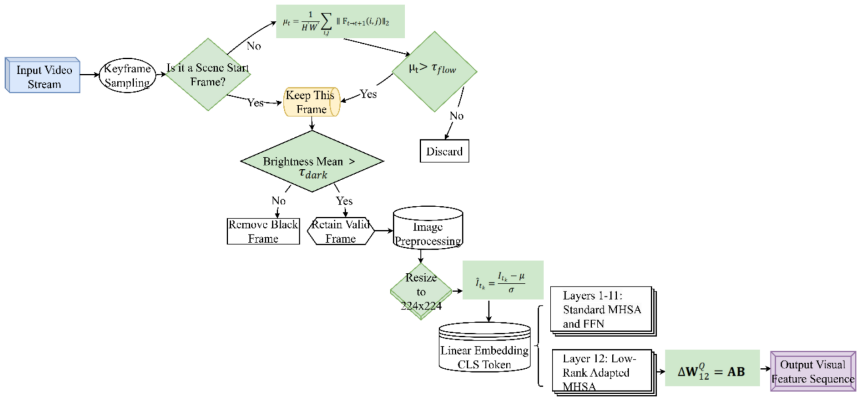

FIGURE 1. Key frame sampling and encoding flow chart

set) to weight and fuse semantic similarity and intensity consistency. This hybrid strategy leverages the discriminative and interpretable nature of discrete prototypes while preserving the support of continuous emotion characterization for fine-grained matching and intensity adaptability.

CONSTRUCTION OF CROSS-MODAL SPACE WITH EMOTIONAL ANCHOR AS THE CORE

To achieve comparability between audio and visual modalities in a unified semantic space, this paper constructs a shared embedding space to structurally align the music emotion embedding m_t with the visual emotion prototype vector generated by the CLIP text encoder. Let the set of emotion categories be $E = \{e_k\}_{k=1}^{K}$, corresponding to predefined emotion semantic templates. Each template text e_k is input into the CLIP Transformer text encoder to obtain its corresponding semantic vector t_k, forming the emotion prototype matrix $T = [t_1, t_2, \ldots, t_K]^\top \in \mathbb{R}^{K\times 768}$. Learnable bilinear projection mapping functions $\phi : \mathbb{R}^{768} \to \mathbb{R}^{d_s}$ and $\psi : \mathbb{R}^{768} \to \mathbb{R}^{d_s}$ are introduced to map the different modal features to a common dimension d_s. Specifically:

$$z_m = \phi(m_t) = W_m^\top m_t + b_m \tag{15}$$

$$z_t^{(k)} = \psi(t_k) = W_t^\top t_k + b_t \tag{16}$$

It further designs a cross-modal similarity metric function and uses normalized temperature-scaled cosine similarity:

$$s(m_t, t_k) = \frac{1}{\tau} \cdot \frac{z_m^\top z_t^{(k)}}{\|z_m\|_2 \|z_t^{(k)}\|_2} \tag{17}$$

In formula (17), τ is a learnable temperature parameter that controls the degree of distribution sharpness. The contrastive learning objective is used in the training phase. Construct positive and negative sample pairs: for the true matching audio-text pair $d(m_t, t_k^+)$, the remaining $K-1$ emotion vectors are used as negative samples. Defining the InfoNCE loss:

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(s(m_t, t_k^+))}{\sum_{j=1}^{K} \exp(s(m_t, t_j))} \tag{18}$$

By minimizing this loss, the model learns to place the music feature m_t close to its corresponding emotional semantic anchor t_k^+ in the shared space and away from other irrelevant categories. To enhance emotional semantic anchor t_k spatial geometric consistency, an orthogonal regularization constraint is introduced to prevent the projection matrix from degenerating:

$$\mathcal{R} = \lambda_1(\|W_m^\top W_m - I\|_F^2) + \lambda_2(\|W_t^\top W_t - I\|_F^2) \tag{19}$$

In formula (19), W_m and W_t are two learnable linear projection matrices, which are used to map the features of different modalities into a shared cross-modal alignment space. The final optimization goal consists of two parts: the contrastive learning loss narrows the distance between the matching music emotion embedding and the corresponding visual emotion prototype, while pushing away the mismatched negative samples; the orthogonal regularization constraint ensures the geometric stability of the two modal projection matrices to prevent model degradation. The joint embedding space constructed in this way enables the music emotion representation to be directly aligned with the visual emotion prototype at the vector level, eliminating the intermediate step of generating the picture description text in the traditional method, and providing a unified and stable metric benchmark for subsequent time series matching.

DYNAMIC TIMING ALIGNMENT MECHANISM BASED ON SLIDING WINDOW

In order to solve the nonlinear synchronization problem of movie images and soundtracks in the process of temporal dynamic evolution, this paper designs a sliding window temporal attention mechanism based on local sensitivity to achieve fine-grained alignment of cross-modal sequences [21]. Let the visual feature sequence be $\{v_k\}_{k=1}^{N_{\text{frame}}}$, where v_k is the semantic vector extracted by CLIP-ViT and low-rank adapted for the k-th frame. On the audio side, the music emotion embedding sequence $\{m_t\}_{t=1}^{T_m}$ is obtained after AudioMAE and LSTM aggregation, where m_t represents the music semantic representation of the t-th time step. In view of the inconsistent sampling frequency and

emotional rhythm differences between the two modalities, a local alignment window centered on the current video frame k is defined:

$$\mathcal{W}_k = [t \in \mathbb{Z}^+ \mid t \in [\max(1, t_c - W), \min(T_m, t_c + W)]] \tag{20}$$

In formula (20), t_c is the coarse alignment position based on the time scale, and W is the window radius. Construct the forward attention distribution within the window:

$$\alpha_{k,t} = \frac{\exp(\text{score}(v_k, m_t)/\tau_k)}{\sum_{t' \in \mathcal{W}_k} \exp(\text{score}(v_k, m_{t'})/\tau_k)}, \quad t \in \mathcal{W}_k \tag{21}$$

The similarity score uses a learnable weighted dot product:

$$\text{score}(v_k, m_t) = v_k^\top W_a m_t \tag{22}$$

The size and overlap stride of the sliding window were empirically determined based on the statistical characteristics of film shot switching frequency and music rhythm changes. A window radius of 4 frames, covering approximately 4 seconds of video time, adequately encompasses short-term visual motion and audio emotional fluctuations while avoiding interference from irrelevant frames caused by excessively large windows. A window overlap stride of 2 frames, meaning adjacent windows overlap by 50%, ensures smooth attention to keyframes and continuity across windows, while improving local alignment accuracy. During model training and validation, different combinations of sliding window sizes and overlap strides were searched on the validation set using a grid search. Alignment accuracy and attention concentration were comprehensively evaluated, and the above parameters were ultimately selected as a compromise, achieving a balance between local sensitivity and computational efficiency, and ensuring fine-grained performance of the sliding window temporal attention mechanism in end-to-end alignment tasks. The temperature coefficient τ_k in formula (22) is dynamically generated by a lightweight network to adapt to the needs of semantic concentration or dispersion in different scenarios. It further introduces a reverse attention mechanism to calculate the attention of the music state m_t to its corresponding visual context:

$$\beta_{t,k'} = \frac{\exp(\text{score}(m_t, v_{k'})/\tau'_t)}{\sum_{k'' \in N_t} \exp(\text{score}(m_t, v_{k''})/\tau'_t)}, \quad k' \in N_t \tag{23}$$

In formula (23), N_t is the reverse window, and k' is the local index variable within the window, which is used to traverse the candidate frames. By fusing the forward and reverse attention, it gets the joint alignment confidence $\gamma_{k,t}$, and finally outputs the alignment path. The optimal music frame index is determined by the maximum confidence:

$$\pi(k) = \arg\max_{t \in \mathcal{W}_k} \gamma_{k,t} \tag{24}$$

This mechanism achieves end-to-end soft alignment without explicit DTW (Dynamic Time Warping) or CTC (Connectionist Temporal Classification), effectively capturing the nonlinear correspondence between scene transitions and musical emotional fluctuations, and providing an accurate time mapping basis for subsequent dynamic matching scoring. The attention weights are shown in Figure 2.

Figure 2 shows the distribution of attention weights within the local alignment window using the sliding window temporal attention mechanism. The horizontal axis represents the music time steps (t-4 to t+4), and the vertical axis represents the video frames (k-4 to k+4). The color depth and the value within the grid represent the normalized attention weight at the corresponding position. As can be seen from the heat map, high-weight regions are concentrated along the main diagonal, forming a continuous bright band slightly to the lower right of the center. The maximum value occurs at (t+2, k+2) (0.38), which is higher than the surrounding area, indicating that the model can dynamically focus on the key areas of audio-visual synchronization. The weights decay in a band-like manner from the center to the periphery, reflecting local sensitivity, and there is no global diffusion phenomenon, indicating that the sliding window temporal attention mechanism effectively suppresses interference from irrelevant frames. The visual emotion prototypes used in this paper are fixed, predefined prototypes based on experience, rather than dynamically learned through clustering during training. Each emotion category (sadness, tension, joy, tranquility, excitement, serenity) corresponds to a natural language description (e.g., "a scene with tense facial expressions and strong contrast in warm and cool colors"). These text descriptions are input into the CLIP text encoder to generate structured prototype vectors, which form an emotion prototype matrix. This matrix remains unchanged throughout training and inference, serving as discrete semantic anchors to ensure stable alignment of the music embedding with the fixed prototypes within a shared cross-modal space. The advantages of this approach are: avoiding the impact of potential clustering instability on alignment quality during training; providing interpretable prototype semantics; allowing the model output to be intuitively mapped to specific emotion categories; and facilitating quantitative evaluation and interpretability analysis of the alignment mechanism. To address the issue that movies typically only have scene or song-level labels, this paper employs a hybrid annotation strategy during the training phase. First, discrete emotion categories (sadness, tension, joy, tranquility, excitement, serenity) are used to provide initial labels for scene levels. Then, fine-grained interpolation is performed within each scene using continuous valence-arousal (V-A) signals to generate virtual temporal labels at the sub-scene level. This strategy combines empirically defined prototype vectors with a continuous mapping of micro-emotional changes, enabling the sliding window temporal attention mechanism to achieve end-to-end dynamic soft alignment of visual and musical sequences without

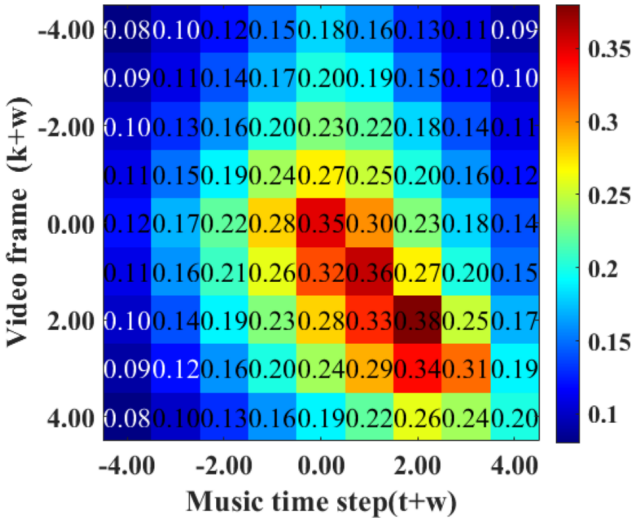

FIGURE 2. Temporal attention weight heat map

explicit manual annotation, thus ensuring the rationality and operability of "fine-grained alignment."

MULTI-DIMENSIONAL EMOTIONAL DYNAMIC MATCHING SCORING FUNCTION

To achieve a comprehensive match evaluation between movie images and soundtracks based on multi-dimensional emotional characteristics, this paper constructs a differentiable and scalable composite scoring function that combines three orthogonal dimensions: semantic consistency, rhythmic synchronization, and emotional intensity adaptability. The overall match score is defined as:

$$S_{\text{match}} = \lambda_1 S_{\text{sem}} + \lambda_2 S_{\text{rhy}} + \lambda_3 S_{\text{int}} \tag{25}$$

Each sub-item is independently modeled as follows. The cross-modal semantic similarity S_{sem} is based on the joint embedding space constructed in Section "Construction of Cross-Modal Space with Emotional Anchor as the Core" and calculates the mean normalized cosine similarity between aligned frame pairs:

$$S_{\text{sem}} = \frac{1}{N} \sum_{(t,\tau)\in A} \frac{\phi(m_\tau)^\top \psi(t_k)}{\|\phi(m_\tau)\|_2 \|\psi(t_k)\|_2} \tag{26}$$

In formula (26), A is the optimal alignment path set output by the sliding window temporal attention mechanism module. Rhythm synchronization rate S_{rhy} measures the synergy between the picture motion rhythm and the music beat. The music BPM (Beat Per Minute) curve is extracted through continuous wavelet transform (CWT) and dynamic programming, the visual rhythm signal is extracted through the optical flow change rate, and dynamic time warping (DTW) is used to align the two [22,23]. Finally, the rhythm synchronization rate is obtained by normalizing the negative warping distance with the Sigmoid function:

$$S_{\text{rhy}} = \sigma(-\gamma \cdot D_{\text{dtw}}) = \frac{1}{1 + \exp(\gamma \cdot D_{\text{dtw}})} \tag{27}$$

In formula (27), D_{dtw} is the cumulative cost, and $\gamma > 0$ controls the response sensitivity. Emotional intensity adaptation S_{int} evaluates the consistency of the emotional intensity between the picture and the music. The emotional intensity $\iota_v(t)$ of the picture comprehensively considers the optical flow amplitude, color contrast, and facial expression activation; the emotional intensity $\iota_m(\tau)$ of the music is calculated by loudness, spectral flux, and harmonic complexity. The smaller the average intensity deviation, the higher the adaptation:

$$S_{\text{int}} = 1 - \Delta_{\text{int}}, \quad \Delta_{\text{int}} = \frac{1}{N} \sum_{(t,\tau)\in A} |\iota_v(t) - \iota_m(\tau)| \tag{28}$$

Ultimately, this scoring function supports end-to-end weight learning or validation set tuning while maintaining the interpretability of each component, forming a comprehensive quantitative evaluation of the emotional synergy between audio and video.

In this paper, the weights for the three dimensions of semantics, rhythm, and sentiment intensity were initially set empirically to 0.4, 0.35, and 0.25, respectively, to reflect the dominant role of semantics in matching while also considering the influence of rhythm and sentiment intensity. To verify the robustness of this empirical weight setting, a sensitivity analysis was conducted to observe the changes in the final comprehensive matching score. The sensitivity results of the score to different weight combinations are shown in Figure 3.

As can be seen from Figure 3, although the weights fluctuated slightly, the overall score remained at a high level, indicating that the empirically set weight combination

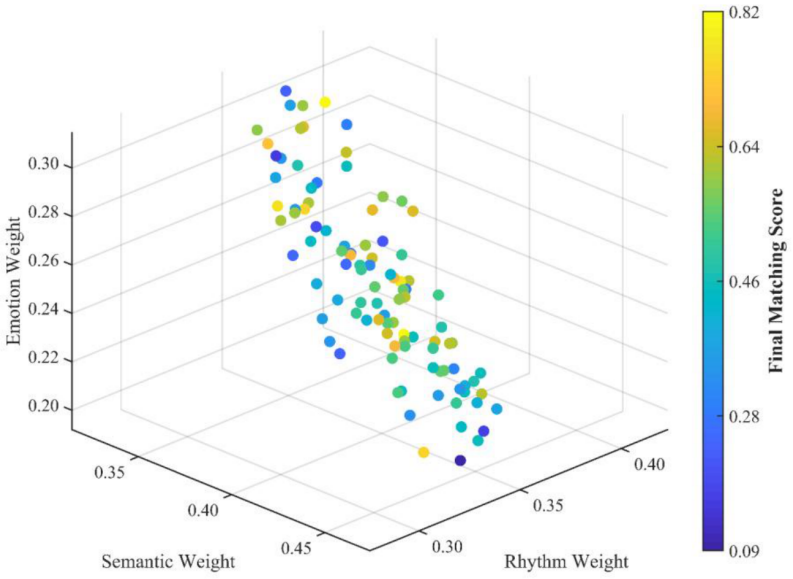

FIGURE 3. Sensitivity Results of Scores to Different Weight Combinations

has good robustness within $\pm 20\%$. Score changes were mainly concentrated in the case of extreme emphasis on a single dimension, while in the region of relatively balanced weights, the final score fluctuation was limited. This verifies the rationality of the selected weights and the robustness of the model output.

GENERATION OF COHERENT MUSIC SEQUENCES BASED ON GLOBAL OPTIMIZATION

In order to generate a coherent and emotionally consistent soundtrack sequence in the temporal dimension, this paper models the final recommendation as a constrained global path optimization problem. Given N consecutive segments $S = \{s_n\}_{n=1}^{N}$ of a movie video divided according to the narrative structure, for each segment s_n, according to the emotional dynamic matching score calculated in Section "Multi-Dimensional Emotional Dynamic Matching Scoring Function", the Top-K' optimal candidates are selected from the music library to form a local candidate set, forming a state space sequence. On this basis, a Markov Decision Process (MDP) model is constructed to optimize cross-segment continuity [24]. The system state is defined as $x_n = (m_n, e_n, p_n)$, where m_n is the selected music, e_n is its emotional category label, and p_n is its pitch tonality. The state transition cost between adjacent segments is determined by the emotional jump penalty and the tonality dissonance:

$$c(x_n, x_{n+1}) = \alpha \cdot d_{\text{emo}}(e_n, e_{n+1}) + \beta \cdot d_{\text{ton}}(p_n, p_{n+1}) \quad (29)$$

In formula (29), the emotional distance $d_{\text{emo}}(\cdot,\cdot)$ is defined as the shortest path length on the preset emotional topology graph; the tonality distance adopts the minimum interval of semitone difference modulo 12 in music theory:

$$d_{\text{ton}}(p_n, p_{n+1}) = \min(|p_n - p_{n+1}|, 12 - |p_n - p_{n+1}|) \quad (30)$$

A smoothness bonus is introduced to encourage rhythm stability. If the BPM difference between two pieces of music is less than a threshold δ_{bpm}, compensation is given:

$$r_{\text{rhy}}(m_n, m_{n+1}) = \gamma \cdot \mathbb{I}\big(|\,\text{BPM}(m_n) - \text{BPM}(m_{n+1})| < \delta_{\text{bpm}}\big) \quad (31)$$

Based on the above, the total cost function is defined as:

$$J = \sum_{n=1}^{N} [-S_{\text{match}}(s_n, m_n)] + \sum_{n=1}^{N-1} [c(x_n, x_{n+1}) - r_{\text{rhy}}(m_n, m_{n+1})] \quad (32)$$

A forward dynamic programming algorithm is used for efficient solution, and the optimal path is finally recovered by backtracking:

$$m_N^* = \arg\min_{m' \in \mathcal{C}_N} V_N(m'), \quad m_n^* = \text{prev}(m_{n+1}^*) \quad (33)$$

Outputting a complete recommendation sequence to ensure smooth emotional transitions and auditory consistency across the entire film soundtrack while maintaining the quality of individual segment matches.

III. EXPERIMENT AND VERIFICATION

EXPERIMENTAL DESIGN

Each video clip is annotated with keyframe images and original soundtracks, sampled at 1 fps. The data is divided into training/validation/test sets at a 7:2:1 ratio to ensure no overlapping videos. The candidate music library covers six emotional categories: sadness, tension, joy, tranquility (a sense of peace induced by the external natural environment or physical space), excitement, and peace (a state of mind characterized by a relaxed facial expression, a restful posture, or a warm moment of reunion after a character has achieved reconciliation, relief, or satisfaction). All methods use the same input resolution and audio preprocessing process to ensure fair comparison, as shown in Table 1. This study was approved by the Ethics Committee of Shilla University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

TABLE 1. Experimental settings

Item	Description
Keyframe Sampling Rate	1 frame per second
Train/Validation/Test Split	7:2:1
Music Emotion Categories	Sadness, Nervous, Joy, Tranquility, Passionate, Peace
Audio Sample Rate	44.1 kHz, 16-bit PCM (Pulse Code Modulation)
Input Image Resolution	224 × 224 pixels (resized using bilinear interpolation)
Batch Size	32
Optimizer	AdamW (learning rate = 3×10^{-5}, weight decay = 0.01)
Training Epochs	50 epochs with early stopping
Hardware Platform	4 × NVIDIA A100 GPU (Graphics Processing Unit), PyTorch 1.13, CUDA (Compute Unified Device Architecture) 11.7

EMOTION CATEGORY ACCURACY AND F1-SCORE EVALUATION

Run each model on the test set, output the top-1 recommended soundtrack for each video, extract its predefined main emotional category, and compare it with the real label. Calculate the emotional category accuracy:

$$Acc = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(c_n = \hat{c}_n) \tag{34}$$

In formula (34), c_n is the true emotion category, $\hat{c}_n$ is the predicted category, and N is the total number of samples.

Figure 4 shows the comparative results of the proposed method, Text-based Retrieval, AudioCLIP-ZeroShot, and Flamingo-Audio on six emotion categories. The horizontal axis represents sadness, tension, joy, tranquility, excitement, and peace, while the vertical axis represents accuracy and F1-Score performance indicators. As the model evolves from text-based retrieval to a multimodal, end-to-end architecture, overall performance continues to improve. The proposed method achieves an average accuracy of 76.3 ± 0.6%, significantly exceeding Flamingo-Audio's 71.2 ± 0.7% and AudioCLIP-ZeroShot's 66.7 ± 0.8%, demonstrating the superiority of the cross-modal alignment strategy without text bridging. Within fine-grained categories, highly expressive emotions such as joy and sadness perform best. While tranquility and peace, due to weaker visual cues, yield relatively low scores across all models, the proposed method still achieves accuracies of 75.7 ± 0.6% and 76.9 ± 0.6%, respectively, demonstrating stronger discriminative capabilities. The proposed method achieves a comprehensive lead across all six emotion categories, with low standard deviations (generally below 0.7%), demonstrating not only high accuracy but also repeatability and robustness.

Traditional methods rely on generating text from images and then matching it to audio, which can lead to semantic distortion and error accumulation. For example, Text-Based Retrieval achieves an overall F1 score of only 59.8 ± 1.3% due to its inability to capture the emotional coupling between music rhythm and image dynamics.

CROSS-MODAL SIMILARITY SCORE ANALYSIS

The image feature v_i generated by this method and the matching music feature m_t are projected into a unified 768-dimensional semantic space and the cosine similarity is calculated:

$$S_{\text{sin}} = \frac{v_i^{\top} m_t}{\|v_i\|_2 \|m_t\|_2} \tag{35}$$

The average score across all correctly matched pairs in the test set is used as a quantitative indicator of cross-modal semantic consistency. A high score indicates that the audio and video are highly aligned in the shared space.

Figure 5 shows the cosine similarity of the image and soundtrack features in a unified semantic space for the proposed method, Text-based Retrieval, AudioCLIP-ZeroShot, and Flamingo-Audio, measuring the degree of cross-modal semantic alignment. Overall, the proposed method achieves an average similarity of 0.732 ± 0.088, significantly outperforming Text-based Retrieval (0.579 ± 0.144), AudioCLIP-ZeroShot (0.629 ± 0.123), and Flamingo-Audio (0.672 ± 0.105), demonstrating a more consistent joint representation of audio and video. Among all categories, the proposed method achieves the highest scores for highly expressive emotions such as joy (0.739 ± 0.086) and sadness (0.742 ± 0.084), while maintaining stable output in low-dynamic scenes such as tranquility (0.725 ± 0.09) and tension (0.728 ± 0.089). The standard deviation of all categories is controlled within ±0.092, which is significantly lower than that of other methods (up to ±0.150). This shows that the proposed method not only has high matching accuracy, but also has stronger generalization ability and stability for different types of emotions.

TIMING ALIGNMENT ACCURACY EVALUATION

Locating the emotional turning points in the real soundtrack (via BPM mutation or spectrum center of mass change

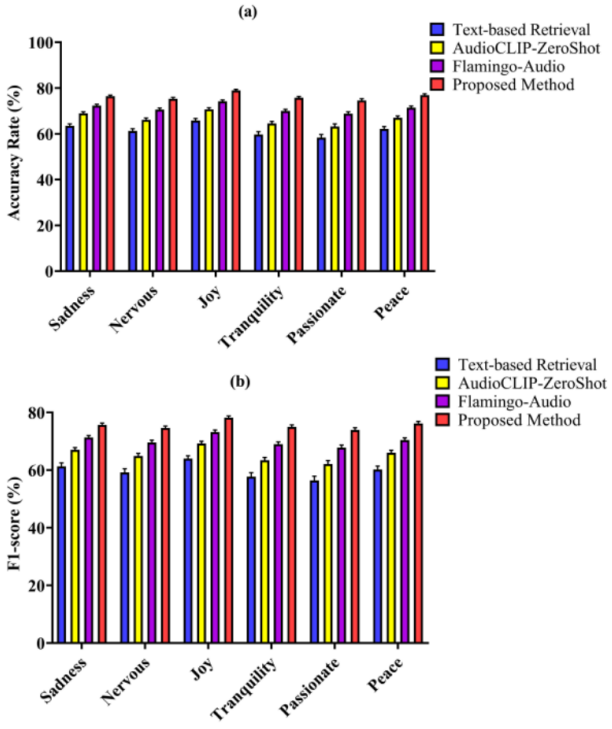

FIGURE 4. Sentiment Category Accuracy and F1-Score. (a) Accuracy; (b) F1-Score

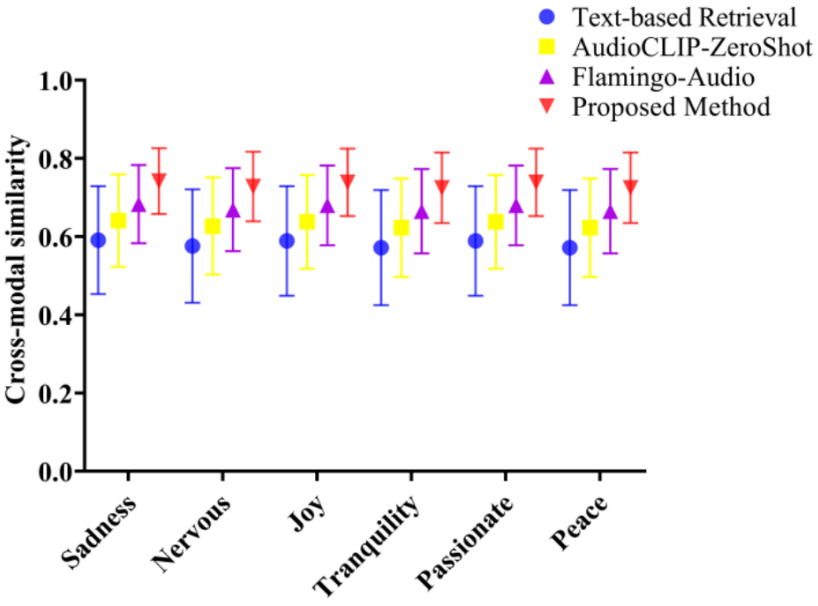

FIGURE 5. Cross-modal similarity score analysis

detection) and extract the corresponding time t_m; simultaneously detecting the frames where the emotion changes in the picture and record the time t_v. Calculating the absolute time deviation:

$$\delta t = |t_m - t_v| \tag{36}$$

Figure 6 compares the temporal alignment accuracy of four methods. The horizontal axis shows Text-based Retrieval, AudioCLIP-ZeroShot, Flamingo-Audio, and the proposed method. The left vertical axis shows the average time deviation, presented as a bar graph. The right vertical axis shows the alignment success rate, including two broken lines

(@2.0s and @1.5s), which represent the percentage of samples where the deviation between the predicted alignment time and the true turning point is within a specified threshold. The average deviation of the proposed method is only 1.24 ± 0.22 seconds, significantly lower than Flamingo-Audio's 1.92 ± 0.31 seconds, AudioCLIP-ZeroShot's 2.35 ± 0.38 seconds, and Text-based Retrieval's 2.87 ± 0.43 seconds. Furthermore, the proposed method achieves an alignment success rate of 83.3% @1.5s and 91.6% @2.0s, significantly exceeding other methods. This demonstrates the proposed method's overwhelming advantage in temporal sensitivity for dynamic emotion matching.

USER PERCEPTION CONSISTENCY VERIFICATION

Ten judges with film and television production experience are invited to watch the original footage and the algorithm-recommended music video under double-blind conditions. They are rated using a 5-point Likert scale (1 = extremely inconsistent, 5 = completely consistent). Each judge independently scored, and the average opinion score is calculated.

Figure 7 shows the scoring results of the proposed method, Text-based Retrieval, AudioCLIP-ZeroShot, and Flamingo-Audio in a user perception consistency test. The proposed method achieves an overall score of 4.35 ± 0.39, significantly higher than Flamingo-Audio (3.69 ± 0.52), AudioCLIP-ZeroShot (3.08 ± 0.62), and Text-based Retrieval (2.22 ± 0.76), indicating that the recommended soundtracks are nearly perfectly aligned with the emotional content of the images. Furthermore, the internal standard deviation within each reviewer is kept between 0.35 and 0.43, significantly lower than that of the other methods, indicating that the judges' judgment of the consistency of the proposed method's output is stronger and the subjective experience is more stable.

RECOMMENDATION DIVERSITY AND COHERENCE EVALUATION

Calculating the Shannon entropy of the recommended sequences to measure diversity:

$$H = -\sum_{k=1}^{K} p_k \log p_k \tag{37}$$

In formula (37), p_k is the frequency of the k-th music in the recommendation results. The coherence is measured using transition smoothness:

$$\text{TS-Score} = \frac{1}{L-1} \sum_{l=1}^{L-1} \exp[-\gamma \cdot d(e_l, e_{l+1})] \tag{38}$$

In formula (38), L is the sequence length, and $d(\cdot,\cdot)$ is the semantic distance between emotion categories.

Figure 8 shows the comprehensive performance of the four methods in terms of recommendation diversity and coherence. The proposed method's Shannon entropy (2.85) and coherence (2.74) significantly outperform Text-based Retrieval (Shannon entropy 2.18, coherence 0.42), AudioCLIP-ZeroShot (2.43, 0.51), and Flamingo-Audio (2.67, 0.58), demonstrating that it not only mines a richer selection of compatible tracks from the music library but also ensures smooth and natural emotional transitions. These results demonstrate that the proposed algorithm balances the richness of artistic expression with the consistency of narrative logic when generating recommendation sequences.

ABLATION EXPERIMENT VERIFICATION MODULE CONTRIBUTION

Four sets of ablation models are constructed:
(1) Removing the joint projection space;
(2) Removing the sliding window temporal attention;
(3) Removing the scene continuity constraint.

Each metric is evaluated on the same test set, comparing the performance degradation of the full model and quantitatively analyzing the contribution of each module.

Table 2 demonstrates the contributions of each of the core modules through systematic ablation experiments. The presence or absence of the three components, "Joint Projection Space," "Sliding Window Temporal Attention," and "Scene Continuity Constraint," are compared across variants in terms of accuracy and mean time deviation. The full model leads with an accuracy of 76.3 ± 0.6%, a mean time deviation of 1.24 ± 0.22 seconds, and a user rating of 4.35. Removing each module leads to a performance decline: removing the joint projection space results in a sharp drop in accuracy to 69.3 ± 0.9%, while mean time deviation increases to 1.87 ± 0.31 seconds. All differences in metrics are significant at the $p < 0.005$ level.

TABLE 2. Ablation experiment table

Model Variant	Joint projection space	Sliding Window Temporal Attention	Scene continuity constraint	Accuracy rate (%)	Average time deviation (s)	
Full Model	✓	✓	✓	76.3 ± 0.6	1.24 ± 0.22	
	–	✓	✓	69.3 ± 0.9	1.87 ± 0.31	$p < 0.001$
	✓	–	✓	71.8 ± 0.8	1.65 ± 0.28	$p < 0.001$
	✓	✓	–	74.5 ± 0.7	1.41 ± 0.24	$p < 0.005$

ABLATION EXPERIMENTS ON THE SYNERGY OF LOCAL AND LONG-RANGE MODELING

To verify the synergistic advantages of BiLSTM (long-range modeling) and sliding window temporal attention (local modeling), several control model variants were constructed under the same data partitioning and training strategy. These included Full (original model, BiLSTM + sliding window temporal attention), BiLSTM-only (removing the sliding window module, retaining BiLSTM for temporal alignment), SWA-only (retaining sliding window attention, replacing BiLSTM with temporal average pooling to weaken long-range context), BiLSTM+GlobalAttention (replacing the sliding window with global multi-head self-attention to test the effect of global rather than local attention), and UniLSTM+SWA (replacing BiLSTM with unidirectional LSTM to verify the value of bidirectional information). Each model was independently trained on the same training set

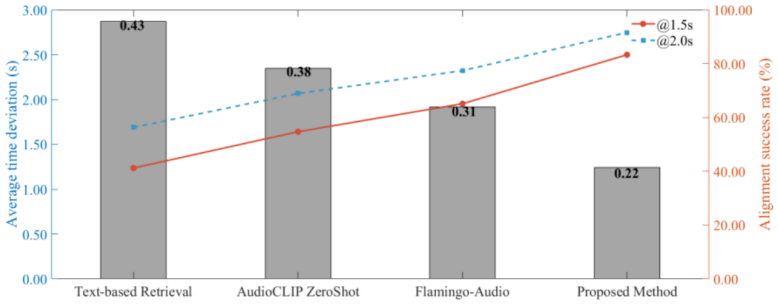

FIGURE 6. Timing alignment accuracy evaluation

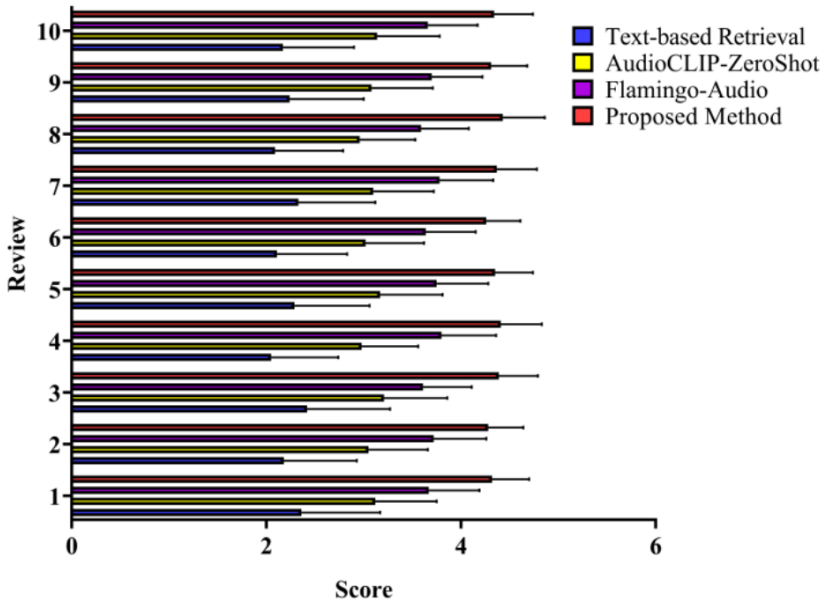

FIGURE 7. User perception consistency test

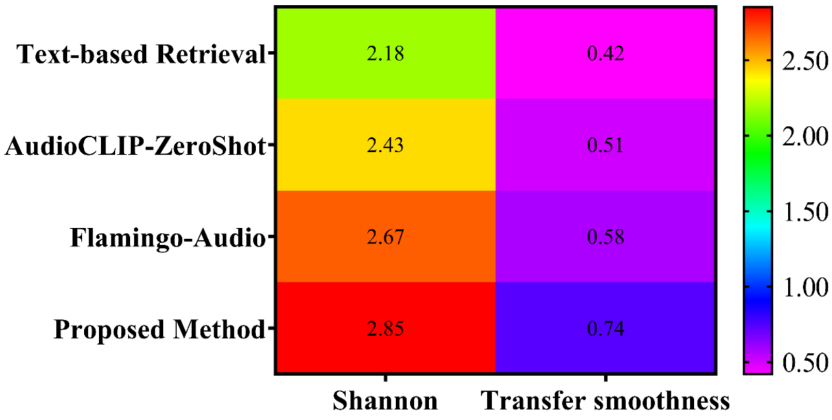

FIGURE 8. Recommendation diversity and coherence evaluation

with 5 different random seeds and evaluated on the test set. Paired t-tests were used for Full and each variant to quantify the significance of performance differences. The results are shown in Table 3.

In the ablation experiments on the synergy of local and long-range modeling (Table 3), the results for each

TABLE 3. Ablation experimental results of synergy between local and long-range modeling

Model Variants	BiLSTM	Sliding Window Temporal Attention (SWA)	Description	Accuracy (%)	Mean Temporal Bias (s)	User Rating (Likert 1–5)	vs Full p-value	Computational overhead (GFLOPs/frame)	Execution efficiency (frames/second)
Full (baseline)	✓	✓	Original Model	76.3 ± 0.6	1.24 ± 0.22	4.35 ± 0.39	–	18.5	54
BiLSTM-only	✓	–	Remove local sliding windows (end-to-end alignment with BiLSTM)	71.8 ± 0.8	1.65 ± 0.28	3.92 ± 0.45	$p < 0.001$	15.2	66
SWA-only	–	✓	Remove BiLSTM (replace long-range context with temporal average pooling)	72.5 ± 0.7	1.50 ± 0.25	4.01 ± 0.42	$p < 0.001$	14.8	68
BiLSTM + GlobalAttention	✓	(Alternative) Global	Replace local sliding windows with global multi-head MHSA self-attention	73.6 ± 0.7	1.38 ± 0.26	4.08 ± 0.41	$p < 0.005$	21	45
UniLSTM + SWA	(One-way LSTM)	✓	Replace BiLSTM with unidirectional LSTM to verify bidirectional information value	74.2 ± 0.6	1.34 ± 0.24	4.15 ± 0.40	$p < 0.01$	16.5	60

model variant clearly demonstrate that the model achieves optimal performance only when both BiLSTM and Sliding Window Temporal Attention (SWA) are present. The Full model has the highest accuracy (76.3% ± 0.6), the lowest mean temporal bias (1.24 ± 0.22 s), and the best user rating (4.35 ± 0.39). Removing either module significantly reduces performance: the accuracy of BiLSTM-only drops to 71.8% and the bias increases to 1.65 s, while the accuracy of SWA-only is 72.5% and the bias is 1.50 s. Both of these are significantly different from Full ($p < 0.001$), indicating that both local and long-range modeling are indispensable. Replacing SWA with global attention (BiLSTM+GlobalAttention) slightly outperformed the version without the module, but was still significantly weaker than Full (accuracy 73.6%, $p < 0.005$), indicating that the "local window" mechanism is more suitable for temporal alignment requirements than global attention. Replacing BiLSTM with unidirectional LSTM (UniLSTM+SWA) also resulted in performance degradation (accuracy 74.2%, $p < 0.01$), emphasizing the importance of bidirectional context in sentiment alignment. Overall, the experiments systematically validated that the collaborative modeling of local (SWA) + long-range (BiLSTM) makes a crucial contribution to improving the accuracy, stability, and perceptual quality of cross-modal temporal alignment. As shown in Table 3, the combination of local and long-range modeling (Full model) achieves optimal performance while maintaining moderate computational overhead (18.5 GFLOPs/frame) and an efficiency of 54 frames/second, meeting real-time processing requirements. Removing local or long-range modules (BiLSTM-only, SWA-only) significantly reduces computation (approximately 14–15 GFLOPs/frame) and increases the frame rate to 66–68 frames/second, but accuracy and user-perceived scores decrease significantly, indicating that the performance loss is not worth the computational overhead savings.

Replacing the sliding window with global attention (BiLSTM+GlobalAttention) introduces global information, but due to the fact that the computational complexity of MHSA increases quadratically with the sequence length, the computational overhead rises to 21 GFLOPs/frame, and the efficiency drops to 45 frames/second, indicating that global attention is costly and has limited practical benefits in time series alignment tasks. Unidirectional LSTM (UniLSTM+SWA) reduces the complexity of long-range modeling, but lacks bidirectional contextual information, resulting in limited efficiency improvement (60 frames/second) and performance inferior to the Full model. Overall, the Full model achieves a better balance between accuracy, user-perceived quality, and computational efficiency, validating the practical feasibility of local + long-range collaborative modeling.

VISUAL EMOTION FEATURE EXTRACTION BENCHMARK

A set of comparative experiments were designed. First, under the same data partitioning (training/validation/testing = 7:2:1) and the same keyframe preprocessing workflow, several visual emotion discriminators (using only visual modalities) were independently trained/evaluated. These included the default CLIP-ViT-B/32 (using only the [CLS] vector output by the visual encoder), two publicly available dedicated visual emotion models (FaceEmotionNet: a face/expression model fine-tuned on AffectNet based on ResNet50; SceneEmotionNet: a scene emotion model fine-tuned on a scene emotion dataset based on SENet), and a simple hybrid fusion (Hybrid = linear projection and orthogonal regularization after concatenating CLIP-CLS and FaceEmotionNet features). The experimental steps are as follows:

(1) Fine-tuning the classification head of each visual model on the training set to predict the six sentiment categories used in the paper;
(2) Evaluating the visual sentiment-only classification performance (accuracy and F1) of each visual model;
(3) Replacing CLIP-visual features with the visual features of each model according to the pipeline in this paper (keeping the other modules unchanged) to measure the downstream cross-modal matching accuracy, which directly reflects the effectiveness of the features in the final system;
(4) Recording the average inference latency per frame and performing a paired t-test on Full vs each variant

to test the significance of the difference. The benchmark comparison results of visual sentiment feature extraction are shown in Table 4.

TABLE 4. Benchmark Comparison Results of Visual Sentiment Feature Extraction

Visual Feature Extractor	Visual Sentiment Classification Accuracy (%)	F1 Scoring	Downstream Matching Accuracy (%)	Average Inference Time/Frame (ms)
CLIP-ViT-B/32 (baseline)	74.0 ± 0.6	0.736	76.3 ± 0.6	12
FaceEmotionNet (AffectNet-ft ResNet50)	70.3 ± 0.6	0.697	73.0 ± 0.7	18
SceneEmotionNet (SENet scene-ft)	69.1 ± 0.8	0.687	73.6 ± 0.6	20
Hybrid (CLIP + Face fusion)	72.4 ± 0.7	0.712	75.5 ± 0.5	28

The results in Table 4 show that different visual feature extractors make significant trade-offs between "visual emotion discrimination ability" and "cross-modal matching performance." In terms of visual classification performance, CLIP-ViT-B/32 achieves the highest accuracy of 74.0% and an F1 score of 0.736, while the two dedicated emotion

models (FaceEmotionNet and SceneEmotionNet) achieve 70.3%/0.697 and 69.1%/0.687, respectively. This indicates that CLIP's semantic pre-training is more generalizable in movie clips with mixed scene and facial information. In terms of downstream cross-modal matching performance, CLIP remains the best at 76.3%, while FaceEmotionNet and SceneEmotionNet drop to 73.0% and 73.6%, respectively. This shows that although both can capture local emotional cues from faces or scenes, their alignment with audio embeddings in the joint space is not as good as CLIP's cross-modal consistency. While the Hybrid model did not surpass CLIP in visual classification (72.4% vs 74.0%), it narrowed the gap to 75.5% in downstream matching, indicating that incorporating facial expression features can compensate for CLIP's shortcomings in capturing emotions in close-up portraits to some extent.

However, its inference overhead is significantly increased (28 ms/frame, compared to CLIP's 12 ms), making it more suitable for offline or non-real-time scenarios. Overall, CLIP-ViT-B/32 achieves the best balance between performance (76.3%) and efficiency (12 ms), while Hybrid provides a slightly better but more expensive alternative.

It's important to note that optimal emotional matching of film scores is influenced not only by visual atmosphere and musical emotion, but also by dialogue semantics, ambient sound effects (such as explosions and rain sounds), and the overall narrative context. This paper's model primarily focuses on visual image semantics and audio-visual features, without explicitly modeling dialogue or ambient sound effect information, nor utilizing plot or narrative structure as auxiliary signals. This means that in scenes containing complex dialogues or strong ambient sound effects, the model may not be able to fully capture multi-source cues triggering emotions, thus limiting its fine-grained emotion matching capabilities. Future work could consider incorporating information such as speech emotion recognition, ambient sound event detection, and narrative context modeling into cross-modal embeddings to further improve the emotional accuracy and narrative consistency of film score recommendations.

IV. CONCLUSION

This paper proposes a movie soundtrack sentiment matching algorithm based on the CLIP model. By constructing a cross-modal joint embedding space for text-free retrieval and combining it with a sliding window temporal attention mechanism, this algorithm achieves fine-grained dynamic alignment of images and music in both semantic and temporal dimensions.

Furthermore, it introduces a multidimensional scoring function and a Markov decision process to optimize the global recommendation sequence. Experimental results demonstrate that this method significantly outperforms existing techniques in terms of sentiment category accuracy, cross-modal similarity, temporal alignment accuracy, and user-perceived consistency. The overall average accuracy reaches 76.3 $\pm$ 0.6%, the average alignment deviation is only 1.24 $\pm$ 0.22 seconds, and the user-perceived consistency test score reaches 4.35 $\pm$ 0.39. The proposed method effectively addresses the semantic distortion and temporal mismatch issues caused by traditional methods' reliance on text, achieving end-to-end, high-precision, and coherent movie soundtrack sentiment matching, demonstrating promising application prospects and artistic expression.

AUTHOR CONTRIBUTIONS

Conceptualization – Junyan Yang, Xiang Gao and Wenxi Zhang; methodology – Junyan Yang, Xiang Gao and Wenxi Zhang; investigation – Junyan Yang and Xiang Gao; writing-original draft preparation – Junyan Yang, Xiang Gao and Wenxi Zhang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Shilla University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] A. H. J. P. J. Permana and A. T. Wibowo, "Movie recommendation system based on synopsis using content-based filtering with TF-IDF and cosine similarity," International Journal on Information and Communication Technology (IJoICT), vol. 9, no. 2, pp. 1-14, 2023, doi: 10.21108/ijoict.v9i2.

[2] C. Jin, M. Lin, F. Wu, X. Wu, Y. Zhou, and J. Wang, "TVMTrailer: A Text-Video-Music AIGC Framework for Film Trailer Generation," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 55, no. 9, pp. 6000-6010, 2025, doi: 10.1109/TSMC.2025.3576988.

[3] B. Millet, J. Chattah, and S. Ahn, "Soundtrack design: The impact of music on visual attention and affective responses," Applied Ergonomics, vol. 93, Art. no. 103301, 2021, doi: 10.1016/j.apergo.2020.103301.

[4] Y. Zhao, M. Yang, Y. Lin, X. Zhang, F. Shi, Wang, Z, et al., "AI-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions," Electronics, vol. 14, no. 6, pp. 1197, 2025, doi: 10.3390/electronics14061197.

[5] J. Gan, "Music feature classification based on recurrent neural networks with channel attention mechanism," Mobile Information Systems, vol. 2021, no. 1, Art. no. 7629994, 2021, doi: 10.1155/2021/7629994.

[6] G. Tong, "Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning," Scientific Programming, vol. 2022, no. 1, Art. no. 2802573, 2022, doi: 10.1155/2022/2802573.

[7] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimedia Tools and Applications, vol. 80, no. 2, pp. 2887-2905, 2021, doi: 10.1007/s11042-020-08836-3.

[8] J. Salas-Cáceres, J. Lorenzo-Navarro, D. Freire-Obregón, and M. Castrillón-Santana, "Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics," Multimedia Tools and Applications, vol. 84, pp. 27327-27343, 2024, doi: 10.1007/s11042-024-20227-6.

[9] S. Moorthy and Y. K. Moon, "Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion," Mathematics, vol. 13, no. 7, pp. 1100, 2025, doi: 10.3390/math13071100.
[10] Z. Su, Y. Feng, J. Liu, J. Peng, W. Jiang, and J. Liu, "An Audiovisual Correlation Matching Method Based on Fine-Grained Emotion and Feature Fusion," Sensors, vol. 24, no. 17, pp. 5681, 2024, doi: 10.3390/s24175681.
[11] Q. Shen, J. Xu, J. Mei, X. Wu, and D. Dong, "EmoStyle: Emotion-aware semantic image manipulation with audio guidance," Applied Sciences, vol. 14, no. 8, pp. 3193, 2024, doi: 10.3390/app14083193.
[12] J. Li, W. K. Wong, L. Jiang, X. Fang, S. Xie, and Y. Xu, "CKDH: CLIP-based knowledge distillation hashing for cross-modal retrieval," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 7, pp. 6530-6541, 2024, doi: 10.1109/TCSVT.2024.3350695.
[13] J. Yang, L. Wang, Y. Qi, H. Chen, and J. Li, "Multimodal Information Fusion and Data Generation for Evaluation of Second Language Emotional Expression," Applied Sciences, vol. 14, no. 19, pp. 9121, 2024, doi: 10.3390/app14199121.
[14] G. Yang, K. Xu, X. Fang, and J. Zhang, "Video face forgery detection via facial motion-assisted capturing dense optical flow truncation," The Visual Computer, vol. 39, no. 11, pp. 5589-5608, 2023, doi: 10.1007/s00371-022-02683-z.
[15] T. K. T. Zizi, S. Ramli, M. Wook, and M. A. M. Shukran, "Optical flow-based algorithm analysis to detect human emotion from eye movement-image data," Journal of Image and Graphics, vol. 11, no. 1, pp. 53-60, 2023, doi: 10.18178/joig.11.1.53-60.
[16] Y. Shao and N. Guo, "Recognizing online video genres using ensemble deep convolutional learning for digital media service management," Journal of Cloud Computing, vol. 13, no. 1, pp. 102, 2024, doi: 10.1186/s13677-024-00664-2.
[17] R. Tan, M. Sun, and Y. Liang, "Transformer-based multi-level attention integration network for video saliency prediction," Multimedia Tools and Applications, vol. 84, no. 13, pp. 11833-11854, 2025, doi: 10.1007/s11042-024-19404-4.
[18] J. Min, Z. Gao, L. Wang, and A. Zhang, "Application research of short-time Fourier transform in music generation based on the parallel WaveGan system," IEEE Transactions on Industrial Informatics, vol. 20, no. 9, pp. 10770-10778, 2024, doi: 10.1109/TII.2024.3397344.
[19] C. Constantinescu and R. Brad, "An overview of sound features in time and frequency domain," International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, vol. 13, no. 1, pp. 45-58, 2023, doi: 10.2478/ijasitels-2023-0006.
[20] S. Seo, C. Kim, and J. H. Kim, "Convolutional neural networks using log mel-spectrogram separation for audio event classification with unknown devices," Journal of Web Engineering, vol. 21, no. 2, pp. 497-522, 2022, doi: 10.13052/jwe1540-9589.21216.
[21] L. C. Reghunath and R. Rajan, "Transformer-based ensemble method for multiple predominant instruments recognition in polyphonic music," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2022, no. 1, pp. 11, 2022, doi: 10.1186/s13636-022-00245-8.
[22] V. K. Singh, K. Sharma, and S. N. Sur, "Acoustic scene classification using dynamic time warping technique based on short time Fourier transform and discrete wavelet transforms," Circuits, Systems, and Signal Processing, vol. 44, no. 3, pp. 1887-1913, 2025, doi: 10.1007/s00034-024-02895-9.
[23] V. K. Singh, K. Sharma, and S. N. Sur, "Impact of various continuous wavelet transforms for acoustic scene classification with DCASE dataset," Signal, Image and Video Processing, vol. 19, no. 6, pp. 440, 2025, doi: 10.1007/s11760-025-04031-9.
[24] R. De Prisco, A. Guarino, D. Malandrino, and R. Zaccagnino, "Induced emotion-based music recommendation through reinforcement learning," Applied Sciences, vol. 12, no. 21, Art. no. 11209, 2022, doi: 10.3390/app122111209.