Combining CRNN Modeling to Dynamically Align the Relationship Between Music Rhythm and Plot Twists in Different Film Genres

Shungong Chi[1], Haiyu Liang[2]
[1]College of Arts, Sichuan University, Chengdu 610207, Sichuan, China
[2]Jinjiang College, Sichuan University, Meishan 620860, Sichuan, China
Corresponding author: Haiyu Liang (e-mail: 17882369156@163.com)

ABSTRACT This paper addresses the challenge of quantitatively modeling the dynamic alignment between musical rhythm and plot turning points in films by proposing V^2A-AlignNet, a genre-aware cross-modal deep learning framework. As intelligent multimedia perception increasingly relies on advanced signal processing and multimodal information fusion techniques that are conceptually relevant to electromagnetic sensing and communication systems, accurate temporal alignment has become an important research topic. The proposed model adopts a dual-stream architecture integrating VideoMAE for long-range spatiotemporal video representation and a CRNN for extracting both local and global rhythmic characteristics from Mel spectrograms. A cross-modal attention module constructs a shared semantic space to generate alignment saliency sequences and similarity matrices, while a genre-conditioning mechanism enables adaptive modeling for different film categories. Experiments conducted on 120 films spanning six genres (800 clips) demonstrate that V^2A-AlignNet achieves superior performance in turning-point detection, alignment pattern classification, and genre recognition compared with representative baseline methods. Ablation studies further verify the effectiveness of each component, and visualization results reveal distinctive genre-specific alignment behaviors. The proposed framework provides a computational basis for audio-visual temporal relationship analysis and offers valuable insights for multimodal signal interpretation and intelligent information processing in advanced electromagnetic sensing and communication-related applications.

INDEX TERMS cross-modal learning, CRNN, audio-visual alignment, multimodal signal processing, electromagnetic sensing

I. INTRODUCTION

As a core multimodal art form, film's emotional narrative effectiveness relies heavily on the precise coordination of auditory and visual channels. Among these, the dynamic interaction between musical rhythm and plot structure is a key artistic means for directors to shape emotions, guide the audience, and strengthen narrative transitions [1,2]. However, this interaction has long relied primarily on the intuition and experience of creators, lacking a calculable theoretical model for systematic quantitative analysis and interpretation [3,4]. Existing computational methods mostly focus on macro-level emotional matching or static feature associations, and still lack an effective modeling framework for the core relationship of "rhythm-transition", which is complex and dynamic at the micro-temporal level [5,6]. This bottleneck restricts a deeper understanding of film narrative grammar and also limits the development of AI-assisted creative tools. In the field of cross-modal analysis of video and audio, early research mainly relied on hand-designed features and shallow models, calculating the correlation between optical flow and audio energy, or using dynamic time warping algorithms to align the temporal sequence of motion and rhythm [7,8]. Although these methods are inspiring, their representational capabilities are limited, and they are difficult to capture high-level semantic information [9]. With the development of deep learning, researchers have begun to use pre-trained models such as I3D (Inflated 3D Convolution) and VGGish to extract features from video and audio, respectively, and fuse them through splicing, pooling, and other methods to complete tasks such as event detection or sentiment classification [10,11]. Although these methods have improved the understanding of content semantics, their fusion strategies are often relatively superficial and fail to specifically model the fine temporal dependencies between audio and video [12]. More advanced cross-modal models,

such as CLIP (Contrastive Language-Image Pre-training) architecture [13,14], have learned rich joint representations on massive network data through contrastive learning, but their focus is more on global semantic alignment rather than local, dynamic temporal interaction patterns in media content. Studies specifically on the relationship between sound and image in film have largely started from film theory, attempting to computationally verify hypotheses such as "how music enhances the narrative". Some works have attempted to detect climax scenes in films and analyze their statistical correlation with musical features [15]; others have focused on the synchronicity of music and camera movement, or analyzed narrative rhythm through the alternation of dialogue and music [16,17]. However, these studies generally simplify the dynamic process of "plot twists" into discrete labels or scenes, failing to examine its coupling relationship with the evolution of musical rhythm from the perspective of continuous temporal sequence [18]. More importantly, most methods assume the existence of a universal sound-image alignment pattern, while ignoring the essential differences in narrative conventions and emotional arousal strategies between different types of films—such as horror films and romance films—and their potential type-specific alignment rules have not yet been fully explored and verified [19,20]. In summary, previous studies have significant limitations in the depth of features, the precision of temporal modeling, and the dimensions of typological analysis. To overcome the aforementioned limitations, this paper proposes a novel cross-modal deep learning framework, V^{2}A-AlignNet (Video-to-Audio Alignment Network), aiming to systematically solve the dynamic alignment problem between musical rhythm and plot twists. The main contributions of this research are threefold. First, methodologically, an end-to-end neural network is constructed, fusing VideoMAE (Video Masked Autoencoder), which excels at capturing long-range spatiotemporal context, with CRNN (Convolutional Recurrent Neural Network), which specializes in temporal pattern modeling, to achieve high-level abstraction and refined extraction of audio-visual features. Second, in model design, a dedicated cross-modal dynamic alignment module is applied, capable of actively learning and outputting a continuous alignment saliency sequence and corresponding alignment pattern labels, thereby achieving a precise and interpretable quantitative description of the core question of "when and how to align". Finally, in terms of theoretical findings, by using film genres as a conditional information embedding model, the unique audio-visual dynamic alignment patterns in different film genres are empirically revealed and compared on large-scale data, providing new analytical dimensions and theoretical insights for computational film studies.

II. METHODS

A. OVERALL FRAMEWORK OVERVIEW

The V^{2}A-AlignNet proposed in this paper is a type-aware cross-modal dynamic alignment network. Its overall architecture is shown in Figure 1. The core objective is to learn the complex alignment relationship between music rhythm and plot twists from film data.

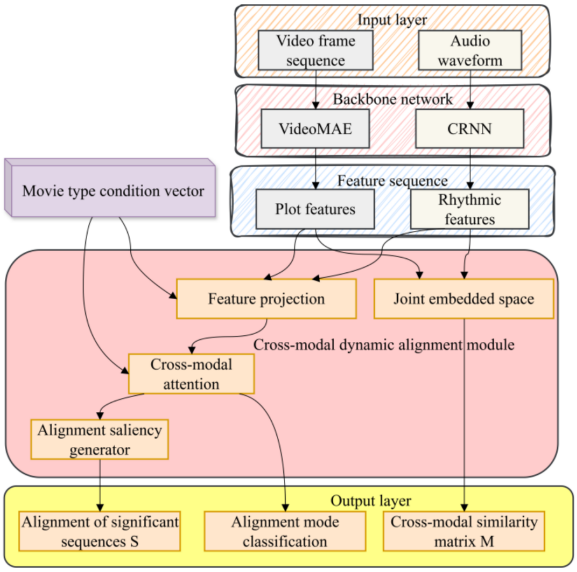

FIGURE 1. Overall architecture of V^{2}A-AlignNet

Figure 1 illustrates the overall architecture of V^{2}A-AlignNet, which employs a dual-stream encoder-fusion design. First, the video and audio streams are processed by independent backbone networks: the video frame sequence is input to the Vision Transformer-based VideoMAE encoder, which, through its powerful spatiotemporal context modeling capabilities, outputs a feature sequence containing plot development information; simultaneously, the audio waveform is converted into a Mel spectrogram and input into a CRNN network, where its CNN (Convolutional Neural Network) front end extracts local spectral patterns, and then its RNN (Recurrent Neural Network) back end captures the long-term temporal dynamics of rhythm, ultimately outputting a music rhythm feature sequence. Subsequently, the feature sequences from both modalities, along with the film genre conditional vector, are fed into the core cross-modal dynamic alignment module. In this module, the features are first projected onto a common subspace, and then deeply interact through a cross-modal attention mechanism to calculate their correlation. The output of this interaction process, on the one hand, generates alignment saliency sequences S that directly characterize the strength of the audio-visual association, and is used to classify specific alignment patterns; on the other hand, the projected features are used to construct a cross-modal similarity matrix M to visualize the temporal alignment structure. Ultimately, the model's synchronous outputs S, M, and alignment mode labels together constitute a complete quantitative description of the dynamic alignment relationship.

B. VIDEO STREAM PLOT FEATURE ENCODING

In video stream processing, VideoMAE is used as the core encoder [21,22]. Its core lies in the self-supervised pre-training paradigm of mask reconstruction. This process forces the model to learn rich spatiotemporal context information in video data to complete the reconstruction task, thereby enabling it to capture coherent plot dynamics from low-level actions to high-level semantics. It is very suitable for detecting plot twists that depend on contextual anticipation [23,24]. The feature extraction process of VideoMAE is shown in Figure 2.

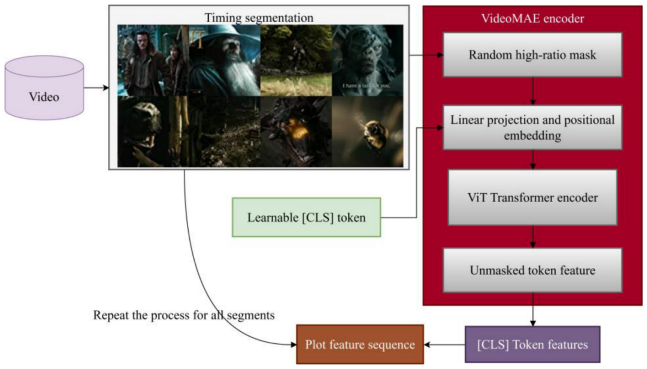

FIGURE 2. Feature extraction process of VideoMAE

The film clip in Figure 2 is taken from: https://www.kaggle.com/datasets/michau96/images-from-hobbit-movies?select=The+Desolation+of+Smaug. Given an input video $\mathcal{V} = \{V_1, V_2, \ldots, V_T\}$, it is first divided into a series of non-overlapping segments where each segment V_t contains F frames. Each segment is then fed independently into the encoder. For a video segment $V_t \in \mathbb{R}^{F\times N\times(P^2\cdot C)}$, each frame is uniformly divided into N image patches, where P is the patch size, and C is the number of channels. During the pre-training phase, VideoMAE randomly masks a very high proportion of spatiotemporal patches, only passing the visible patches through a linear projection layer with position embedding, and then inputting them along with a learnable [CLS] token into the standard Vision Transformer encoder.

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{\text{pos}} \tag{1}$$

In Formula 1, E is a projection matrix, and E_{pos} is a position embedding. The Transformer encoder consists of alternating L-layer multi-head self-attention and multi-layer perceptron modules:

$$z_l' = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = \text{MLP}(\text{LN}(z_l')) + z_l' \tag{3}$$

Finally, the state of the [CLS] token output by the last Transformer layer is used as a compact global plot representation v_t of the video segment. This process is repeated for all segments to obtain the feature sequence of the entire video, that is, $F_v = \{v_1, v_2, \ldots, v_T\}$, which carries the spatiotemporal semantic evolution of the film's narrative.

C. AUDIO STREAM RHYTHM FEATURE ENCODING

To extract features that reflect the dynamics of musical rhythm from audio signals, CRNN is used, whose structure can effectively take into account both local spectral patterns and long-term temporal dynamics [25,26]. The CNN part slides on the Mel spectrogram through its convolution kernel, which is good at capturing fine spectral patterns such as beat onset and timbre texture in local areas; while the RNN part receives the feature sequence extracted by the CNN and uses its internal state memory to model the long-term evolution trend of rhythm, thus completely depicting the temporal structure of musical rhythm from micro to macro [27,28]. The complete process of audio feature extraction is shown in Figure 3.

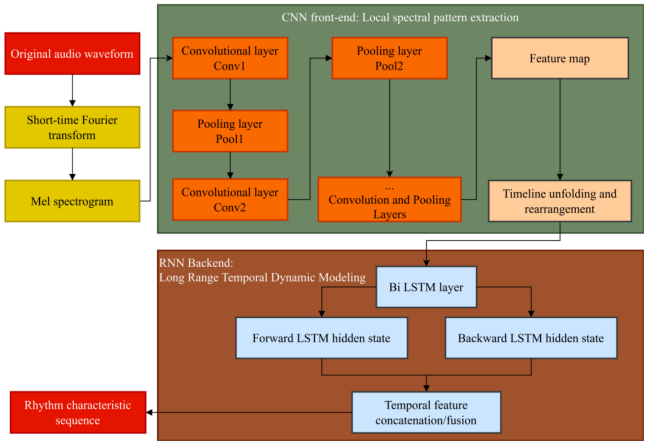

FIGURE 3. Audio feature extraction process

Figure 3 shows the entire processing process: firstly, the audio waveform is transformed into a time-frequency representation - the Mel spectrogram - through short-time Fourier transform and Mel filter bank. The CRNN network extracts local spectral patterns, including dynamic features such as rhythm, timbre, and energy, through its CNN front-end, and then models its long-term temporal evolution through the Bi LSTM back-end, ultimately outputting a comprehensive feature sequence that represents the dynamics of music. The original audio waveform is first converted into a time-frequency representation—a Mel spectrogram $M \in \mathbb{R}^{T\times F}$, where T is the number of time frames, and F is the number of Mel filter banks [29,30]. This spectrogram is then fed into the CNN front end. The front end of a CNN consists of multiple layers of convolutional and pooling operations stacked together. Let the convolutional operation of the l-th layer be:

$$H_l = \text{ReLU}(W_l * H_{l-1} + b_l) \tag{4}$$

In Formula 4, $*$ represents the convolution operation; W_l and b_l represent the weights and biases of the convolutional layer; H_{l-1} is the input feature map. Convolutional layers are typically followed by pooling layers (such as max pooling) to apply translation invariance and reduce dimensionality. After multiple convolutional-pooling layers,

the output is flattened into a sequence of feature vectors $C = \{c_1, c_2, \ldots, c_T\}$, where each vector c_t encodes the local spectral pattern corresponding to the time step t. The RNN back end uses a bidirectional long short-term memory (Bi-LSTM) network to process the sequence C. For each time step t, the Bi-LSTM aggregates the information from the forward and backward propagation:

$$\overrightarrow{h}_t = \mathrm{LSTM}(c_t, \overrightarrow{h}_{t-1}) \tag{5}$$

$$\overleftarrow{h}_t = \mathrm{LSTM}(c_t, \overleftarrow{h}_{t+1}) \tag{6}$$

Ultimately, the rhythmic features of each time step are composed of the concatenation of the forward and backward hidden states:

$$a_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{7}$$

The output of the entire audio stream is a rhythmic feature sequence $F_a = \{a_1, a_2, \ldots, a_T\}$, which integrates the local beat information of the music with the global rhythmic evolution dynamics, providing rich auditory representations for subsequent cross-modal alignment.

D. CROSS-MODAL DYNAMIC ALIGNMENT MODULE

The cross-modal dynamic alignment module is the core innovation of this study, which aims to model the dynamic interaction between video feature sequences F_v and audio feature sequences F_a. This module achieves accurate quantification of audio-visual alignment relationship through attention mechanism and contrastive learning [31,32]. The core working principle of this module lies in establishing a dynamic association between video and audio features through a cross-modal attention mechanism. Specifically, the module first projects the video and audio feature sequences onto a common semantic space, and then calculates cross-modal attention weights using video features as queries and audio features as keys. These weights reflect the correlation strength between each video segment and all audio segments. By aggregating the weighted audio context information, a video representation fused with cross-modal information is generated. By analyzing the variance of the attention weight distribution, the module can quantify the significance of audio-visual alignment at each moment and identify specific alignment patterns such as synchronization, anticipation, or continuation based on the fused features. Heterogeneous modal features are projected into a common semantic space through linear transformation:

$$v_i' = W_v v_i + b_v, \quad a_j' = W_a a_j + b_a \tag{8}$$

In Formula 8, $W_v \in \mathbb{R}^{d_p \times d_v}$ and $W_a \in \mathbb{R}^{d_p \times d_a}$ are the learnable projection matrices; b_v and $b_a \in \mathbb{R}^{d_p}$ are the bias terms; d_p is the projection dimension. A cross-modal similarity matrix $M \in \mathbb{R}^{T \times T}$ is constructed based on the projected features:

$$M_{ij} = \frac{v_i'^T a_j'}{\|v_i'\| \|a_j'\|} \tag{9}$$

In Formula 9, M_{ij} represents the cosine similarity between the i-th video segment and the j-th audio segment in the public semantic space. A query-key-value cross-attention architecture is adopted [33,34]. Video features are used as queries, and audio features are used as keys. The projection of the query, key, and value is calculated as follows:

$$q_i = v_i' W_Q, \quad k_j = a_j' W_K, \quad v_j = a_j' W_V \tag{10}$$

In Formula 10, W_Q, W_K, and $W_V \in \mathbb{R}^{d_p \times d_k}$ are the projection matrices, and d_k is the attention dimension. The attention weight calculation is:

$$e_{ij} = \frac{q_i k_j^T}{\sqrt{d_k}}, \quad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \tag{11}$$

The context vector generation is:

$$c_i = \sum_{j=1}^{T} \alpha_{ij} v_j \tag{12}$$

The alignment significance score for each time step is obtained by calculating the variance of the row vectors of the attention weight matrix:

$$s_i = \frac{1}{T} \sum_{j=1}^{T} (\alpha_{ij} - \bar{\alpha}_i)^2 \tag{13}$$

In Formula 13, $\bar{\alpha}_i = \frac{1}{T} \sum_{j=1}^{T} \alpha_{ij}$. The alignment significance score aims to quantify the degree of alignment concentration between video segment i and all audio segments by calculating the variance of the row vectors in the attention weight matrix. From the perspective of information theory and attention mechanism, if the attention distribution is relatively concentrated, it indicates that the video clip only has strong correlation with a few specific audio clips, reflecting that the audio and video are highly synchronized or closely coupled in semantics or rhythm at this time. Therefore, it can be used as an effective proxy indicator for "alignment strength". On the contrary, low variance implies scattered attention, corresponding to weak cross modal correlations or unclear temporal alignment. The original video features are concatenated with the context vector before classification:

$$h_i = \mathrm{ReLU}(W_h[v_i; c_i] + b_h) \tag{14}$$

$$y_i = \mathrm{Softmax}(W_y h_i + b_y) \tag{15}$$

In Formula 15, $y_i \in \mathbb{R}^3$ corresponds to the probability distribution of three alignment modes: synchronization, prediction, and continuation.

E. FILM GENRE CONDITIONALIZATION AND OUTPUT LAYER

To achieve type-aware modeling, film type information is injected as a conditional signal into each stage of the network. The type embedding representation is:

$$e_g = W_g g \tag{16}$$

In Formula 16, $g \in \{0,1\}^{N_g}$ is a one-hot type label, $N_g = 6$, and $W_g \in \mathbb{R}^{d_g \times N_g}$ is an embedding matrix. The feature extraction process is modulated using the FiLM (Feature-wise Linear Modulation) mechanism. For the output feature $H_v^l \in \mathbb{R}^{B \times T \times d_v}$ of the l-th layer of the video encoder, there are:

$$\gamma_v^l = W_{\gamma v} e_g + b_{\gamma v}, \quad \beta_v^l = W_{\beta v} e_g + b_{\beta v} \tag{17}$$

$$\mathrm{FiLM}(H_v^l) = \gamma_v^l \odot H_v^l + \beta_v^l \tag{18}$$

In Formula 18, $\odot$ indicates element-wise multiplication, and the modulation parameters $\gamma_v^l, \beta_v^l \in \mathbb{R}^{d_v}$ are generated based on the type characteristics. In the alignment module, type information is integrated into the attention calculation:

$$q_i = [v_i'; e_g] W_Q', \quad k_j = [a_j'; e_g] W_K' \tag{19}$$

Film genre conditional information is integrated into the model during the feature extraction stage via the FiLM mechanism, specifically acting on the intermediate layer features of VideoMAE and CRNN to achieve typological modulation of video spatiotemporal semantics and audio rhythm representation. Simultaneously, the genre embedding vector is concatenated into the query and key of the cross-modal attention module, directly participating in the calculation of alignment weights. This dual fusion strategy ensures that genre information both influences low-level feature extraction and guides high-level alignment relationship modeling, thereby achieving true genre-aware dynamic alignment. The current model adopts a static type condition vector and performs global modulation at the feature level through the FiLM mechanism, assuming that the film type alignment style remains consistent within the same segment. Although this method can effectively capture the alignment preferences of different movie genres as a whole, this article acknowledges its theoretical limitations in handling dynamic style transitions within segments, such as action scenes that suddenly appear in drama films. Future work can explore dynamic type conditioning mechanisms, such as predicting type weights based on local content or introducing temporal aware type embeddings, to more finely model the evolution of type alignment strategies in the time dimension.

F. LOSS FUNCTION

The model employs a multi-task learning framework, and the total loss function is:

$$L = L_{\mathrm{align}} + \lambda_1 L_{\mathrm{mode}} + \lambda_2 L_{\mathrm{genre}} \tag{20}$$

Alignment loss combines contrastive learning and regression loss:

$$\mathcal{L}_{\mathrm{contrast}} = -\frac{1}{T} \sum_{i=1}^{T} \log \frac{\exp(M_{ii}/\tau)}{\sum_{j=1}^{T} \exp(M_{ij}/\tau)} \tag{21}$$

In Formula 21, τ is a temperature hyperparameter that encourages audio and video features at the same moment to move closer together in the joint space. Regression loss uses mean squared error:

$$L_{\mathrm{reg}} = \frac{1}{T} \sum_{i=1}^{T} \|s_i - s_i^*\|_2^2 \tag{22}$$

In Formula 22, s_i^* represents the true alignment saliency generated based on the labeled plot turning points. The total alignment loss is:

$$L_{\mathrm{align}} = L_{\mathrm{contrast}} + L_{\mathrm{reg}} \tag{23}$$

Alignment pattern classification uses cross-entropy loss:

$$\mathcal{L}_{\mathrm{mode}} = -\frac{1}{T} \sum_{i=1}^{T} \sum_{c=1}^{3} y_{ic}^* \log(y_{ic}) \tag{24}$$

In Formula 24, y_{ic}^* is the one-hot encoding of the true alignment mode. The film genre classification loss is:

$$\mathcal{L}_{\mathrm{genre}} = -\sum_{c=1}^{N_g} g_c^* \log(\hat{g}_c) \tag{25}$$

In Formula 25, $\hat{g}_c$ represents the probability distribution of the type predicted by the model. The key hyperparameter settings of the model are shown in Table 1.

TABLE 1. Key hyperparameter settings of the model

Parameter category	Parameter name	Value	Instructions
Optimize configuration	Optimizer	AdamW	Use weight decay
Optimize configuration	Initial learning rate	1e-4	Cosine annealing scheduling
Optimize configuration	Batch size	16	Based on memory
Optimize configuration	Weight decay	0.01	L2 regularization coefficient
Loss weight	λ_1	0.5	Pattern classification weight
Loss weight	λ_2	0.2	Type classification weight
Dimension configuration	Video feature dimension	768	VideoMAE output
Dimension configuration	Audio feature dimension	512	CRNN output
Dimension configuration	Projection dimension	256	Dimension of Public Space
Dimension configuration	Attention dimension	64	Each dimension
Architecture parameters	Number of attention heads	8	Multi head attention
Architecture parameters	Temperature coefficient	0.1	Compare learning temperature
Architecture parameters	ViT (Vision Transformer)	ViT-Base	12 layer Transformer
Architecture parameters	Type embedding dimension	128	Conditional vector dimension
Training setup	training cycle	50	Early stop strategy

III. EXPERIMENTS

A. DATASET AND EVALUATION METRICS

In order to systematically investigate the audio-visual alignment relationships across multiple genres of films, this paper creates a comprehensive multi-genre film dataset. The dataset consists of 120 films released between the years 2000 and 2023. Each genre is made up of 20 films in 6 primary genres, including action, horror, romance, suspense, comedy, and drama. All films have been collected in 1080p resolution or higher, with a standard audio sampling rate of 44.1kHz. This assures a certain quality of consistency across the dataset. During data pre-processing, the opening sequences and end credits are removed, keeping only the relevant content for analysis. Each film is broken down into 5-10 minute segments for analyzing. Therefore, there are a total of 800 valid segments in the dataset that provide a wealth of data for training and testing models. The processing method of dividing the film into 5-10 minute segments in this study is mainly based on practical considerations such as training efficiency, computational resource limitations, and existing annotation granularity. Although longer narrative arcs may span longer periods of time, segmented design can still capture plot twists and rhythm alignment patterns within local boundaries, and establish effective contextual connections within segments through VideoMAE's temporal modeling capabilities. Longer segments or full film modeling may theoretically better capture the narrative development across paragraphs, which is also an important limitation and future improvement direction of this study. At the current stage, segmented analysis can provide effective and reproducible quantitative insights for typified audio visual alignment, and lay the methodological foundation for subsequent modeling that extends to longer temporal contexts. Detailed statistics of the dataset are shown in Table 2.

TABLE 2. Dataset statistics

Film genre	Number of films	Number of fragments	Total duration (h)	Average fragment duration (s)	Total turning points
Action	20	135	18.2	485.3 ± 45.2	2,415
Horror	20	142	19.5	494.7 ± 52.1	2,836
Romance	20	128	17.3	486.9 ± 38.7	1,792
Suspense	20	139	18.9	489.5 ± 43.6	2,645
Comedy	20	131	17.8	489.1 ± 41.3	2,098
Drama	20	125	17.1	493.2 ± 47.5	1,875
Total	120	800	108.8	489.8 ± 44.7	13,661

The division of the dataset is based on the film ratings and includes a training set (84 films), a validation set (18 films), and a test set (18 films) (which avoids data leakage by ensuring that different clips from the same film do not exist in different sets). Plot turning points are indicated independently by three trained annotators. Turning points are thought of as interesting moments that cause a significant change in plot direction, character state, or emotional tone, and can include (but are not limited to) the following examples: sudden events; major discoveries or revelations; emotional intensity shifts; major pace shifts in the narrative. Annotation consistency is assessed using the Fleiss' Kappa coefficient, which yields a final $\kappa = 0.78$, showing good consistency. The timing difference Δt is derived directly from the time separation between the peak of the audio energy envelope and the visual inflection point. The annotation of alignment patterns is reliant on the precise timing of the audio event and its temporal relationship with the visual inflection point. Three patterns are established based on the audio-visual timing: synchronous (musical change and visual inflection occur simultaneously, $|\Delta t| < 1$ second); predictive (musical change occurs in advance of visual inflection, $\Delta t \geq 1$ second), and continuous (musical change occurs after visual inflection, $\Delta t \leq -1$ second). Annotations use professional audio analysis software to extract rhythmical feature peaks and compare them to the visual inflection timestamps recorded in the video annotations, thus assuring the objective and consistency of pattern determinations. A multi-dimensional evaluation system is constructed to provide an analysis of model performance. For align point detection, the average precision is used to analyze the accuracy of inflection point detection, and computed using the below formula:

$$AP = \int_0^1 p(r)dr \tag{26}$$

Alignment mode classification uses macro-average F1 scores to treat all categories equally:

$$F1_{\text{macro}} = \frac{1}{3}\sum_{c=1}^{3}\frac{2P_cR_c}{P_c + R_c} \tag{27}$$

In Formula 27, P_c and R_c represent the precision and recall of category c, respectively. Film genre classification uses overall accuracy to evaluate genre recognition performance:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} I(\hat{g}_i = g_i)}{N} \tag{28}$$

Cross-modal retrieval: the quality of the joint embedding space is evaluated, and the reported retrieval hit rate is:

$$R@K = \frac{RS}{TS} \tag{29}$$

In Formula 29, RS represents the correct retrieval in the topK, and TS represents the total number of queries. R@1 and R@5 are calculated for both the video-to-audio (V→A) and audio-to-video (A→V) directions.

B. COMPARISON OF MODELS AND IMPLEMENTATION DETAILS

To comprehensively evaluate the performance of the proposed V^{2}A-AlignNet, this paper selects representative baseline models for comparison, covering traditional temporal alignment methods, classic multimodal fusion models, and cutting-edge cross-modal pre-training architectures. Specifically, these include: the DTW (Dynamic Time Warping) method based on dynamic time warping; traditional machine learning models using handcrafted features and SVM (Support Vector Machine) classifiers [35-36]; fusion

models that use I3D and VGGish to extract video and audio features respectively and then concatenate them; deep networks that combine SlowFast and LSTM (Long Short-Term Memory) for temporal modeling; a visual-audio contrastive learning model based on CLIP-ViT and CLAP (Contrastive Language-Audio Pre-training); Perceiver IO (Perceiver Input-Output architecture) that uses a unified Transformer to process multimodal inputs; the X-CLIP model designed specifically for video-language tasks. These baselines reflect the mainstream technical approaches for current audio-visual alignment tasks from different perspectives. For fair comparison, they are all reproduced or adapted under the same dataset and evaluation protocol. This research uses the PyTorch deep learning framework to carry out experiments. All models are trained in parallel on eight NVIDIA A100 80GB GPUs, and mixed precision is used to speed up computations and mitigate memory usage. To assess the comprehensiveness and stability of the V^2A-AlignNet training process, Figure 4 presents the loss and performance convergence curves of the model over 50 training epochs.

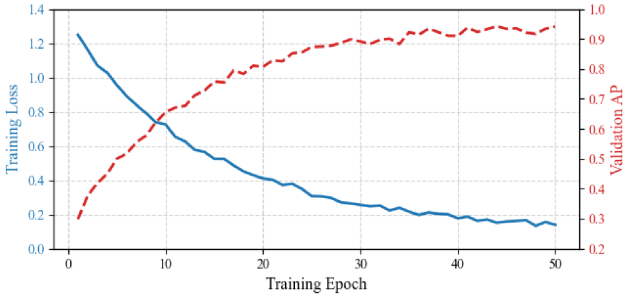

FIGURE 4. Loss and performance convergence

Overall, the training loss steadily decreases and stabilizes with more epochs; the validation set AP (Average Precision) also steadily increases and converges at a high level, confirming that the model does not overfit and the optimization process is effective. The loss curve initially decreases quite rapidly for the first 15 epochs, indicating that the model can quickly learn effective audio-visual alignment features; the curvature of the loss in subsequent epochs indicates the stability of the training process. The validation set AP curve shows an approximately monotonic increasing trend and then enters a relative plateau around 35 epochs, indicating that the model can generalize well. In summary, guided by the multi-task loss function design, V^2A-AlignNet is capable of simultaneously focusing on alignment saliency regression, pattern recognition classification, and type recognition objectives with sound optimization characteristics and trustworthy performance convergence behaviour.

IV. RESULTS

A. COMPARISON OF QUANTITATIVE RESULTS

To systematically evaluate model performance, Figure 5 compares the AP, F1 score, and type classification accuracy of each method on the test set, comprehensively measuring their cross-modal alignment capabilities.

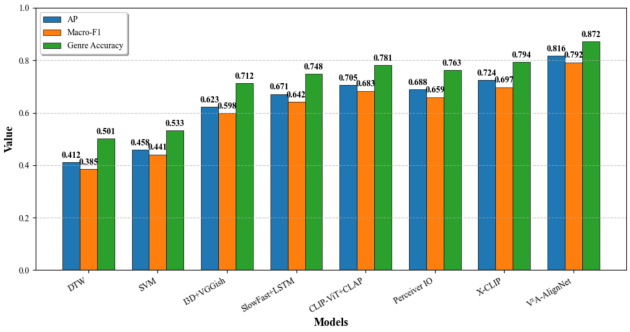

FIGURE 5. Comparison of quantitative results

Quantitative findings indicate that V^2A-AlignNet surpasses all baseline models in AP, macro-average F1 score, and type classification accuracy, exhibiting comprehensive performance improvements. For example, the proposed model attains an AP of 0.816 and F1 score of 0.792, exceeding prior models; it also obtains the highest type classification accuracy (0.872), demonstrating the validity of the type conditionation mechanism employed. The performance benefits of V^2A-AlignNet arise from three aspects: first, the VideoMAE encoder establishes a superior spatiotemporal representation capacity acquired via mask reconstruction pre-training that surpasses conventional video encoders such as I3D; second, the CRNN architecture can simultaneously capture both the local spectral patterns and long-term temporal dynamics of musical rhythms that outperforms static audio features such as VGGish; third, the cross-modal dynamic alignment module explicitly models the temporal dependencies existing between audio and video content, featuring an attention mechanism for the modeling process, while also enriching the representation using type conditionation; whereas baseline models use basic strategies for fusing audio and visual features such as concatenation or pooling, which cannot effectively capture such sophisticated dependencies. In summary, V^2A-AlignNet achieves optimization in three domains: feature extraction, temporal modeling, and cross-modal fusion, thus obtaining superior overall performance in a multi-dimensional evaluation approach.

B. ABLATION EXPERIMENTS

To verify the effectiveness of each module of V^2A-AlignNet, the contribution of key components to the model performance is evaluated through ablation experiments. The results are shown in Table 3.

TABLE 3. Comparison of ablation test performance

Model	AP	R@1 (V→A)	R@5 (V→A)	R@1 (A→V)	R@5 (A→V)
(a)	0.816	0.452	0.783	0.438	0.769
(b)	0.742	0.398	0.714	0.385	0.702
(c)	0.758	0.411	0.728	0.396	0.721
(d)	0.693	0.352	0.665	0.341	0.648
(e)	0.779	0.426	0.752	0.412	0.741

Notes: (a) Full model; (b) No VideoMAE (I3D alternative); (c) No CRNN (VGGish alternative); (d) No cross-modal attention (spoofing alternative); (e) No type conditionalization.

The ablation studies indicate that the full model performs the best across all outcome measures, which confirms that every design choice matters. For example, substituting VideoMAE with I3D causes AP to decrease by 0.074 and R@1(V→A) to decrease by 0.054, indicating that VideoMAE provides an important spatiotemporal context modeling capability obtained from pre-training its mask reconstruction. Another example is removing CRNN with VGGish, which causes AP to decrease by 0.058 and R@5(A→V) to decrease by 0.048, indicating CRNN's relative advantages for capturing longer dependencies for musical rhythm. Certainly, the largest impact occurs when the cross-modal attention mechanism is removed - replacing it with simple concatenation - which leads to a substantial performance drop, where AP decreases by 0.123. The cross-modal attention mechanism is key in modeling the fine-grained temporal alignment of audio and video signals. Removing the type condition also reduces performance with AP decreasing by 0.037. For the cross-modal retrieval task, it also decreases performance simultaneously, demonstrating that applying film type information allows the model to help learn type specific alignment patterns. Overall each component substantively contributes to final performance, with the cross-modal attention module having the most impact.

C. VISUALIZATION OF DYNAMIC ALIGNMENT RELATIONSHIPS

To provide a deeper understanding of the dynamic alignment mechanism between audio and video, Figure 6 presents typical examples of salient sequences in action films and cross-modal similarity matrices in horror films.

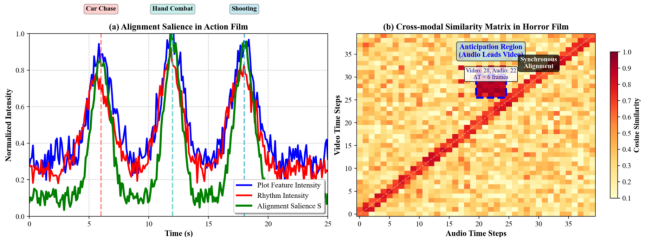

FIGURE 6. Dynamic alignment relationship

In Figure 6(a), the dynamic alignment of a car chase scene in an action film is analyzed. The feature intensity curves depicting plot events reveal that 6, 12, and 18 seconds peaks in the intensity coincide with the three important action scenes taking place in front of viewers: car chase, hand-to-hand combat, and shooting, respectively. The music rhythm features intensities begin to rise just prior to these visual transition points following a highly precise synchronous reinforcement model pattern. The predicted alignment saliency sequence from the model, S, captures all three important temporal points. This precise chronological mapping indicates V^{2}A-AlignNet is sensitive in regard to its perception of the audiovisual coordination mechanism in action films. The music, through the consistency of strong low-frequency impact sound effects with the action sequences, requires the viewer to have the greatest sensory impact when analyzing emotional intensity of audiovisual coordination and implicates strength surrounding the predicted quantification of change of audiovisual intensity. Figure 6(b) shows the cross-modal similarity matrix for a typical "sudden scare" scene in a horror film. The most noticeable structure in the matrix is the clear bright area that appears in the upper left diagonal (at around video time step 28 and audio time step 22); this demonstrates a clear anticipation effect where audio precedes video (in time). The finding is that, around audio time step 22, the string vibrato gradually builds in intensity, and a low-frequency hum starts to overwhelm the scenes with background tension. The visual scare only appears later (around video time step 28). This is an audio-first alignment strategy, that is, a narrative grammar unique to horror films, that inflates the audience's anticipated anxiety through the emotional build of music, thereby enhancing the psychological impact of the sudden scare. The off-diagonal bright spot in the similarity matrix clearly demonstrates that V^{2}A-AlignNet can differentiate patterns of audio-visual temporal relationships.

D. FILM GENRE PATTERNS

To explore the audio-visual alignment patterns of different film genres, Figure 7 compares the average alignment significance patterns of horror, action, and romance films on the normalized time axis.

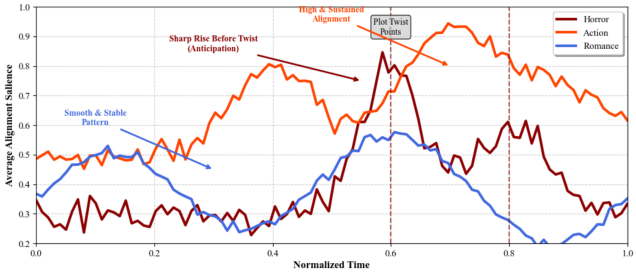

FIGURE 7. Mean alignment significance pattern of horror, action, and romance films

There are notable differences in audio-visual alignment trends across these three film genres, reflecting distinct narrative strategies and emotional arousal mechanisms. Horror films exhibit a "sharp rise" pattern, with alignment significance significantly increasing before plot turning points (at

normalized times 0.6 and 0.8), consistent with the genre's use of music to build tension and amplify shock effects. Action films display a "high and sustained" pattern, maintaining consistently high alignment significance throughout the timeline (baseline around 0.5, peaking above 0.9), which aligns with the need for continuous audio-visual intensification to sustain viewer excitement. Romance films follow a "smooth and stable" pattern, characterized by gentle fluctuations and a mild mid-plot rise, corresponding to the genre's emphasis on gradual emotional buildup rather than dramatic climaxes. These systematic patterns support the theoretical premise that film genre is a key moderating variable in audio-visual alignment and offer a quantitative foundation for genre-adaptive film analysis and generation. To quantify the differences in alignment strategies among different film genres, Figure 8 shows the distribution frequency of three alignment modes—synchronization, anticipation, and continuation—across six film genres.

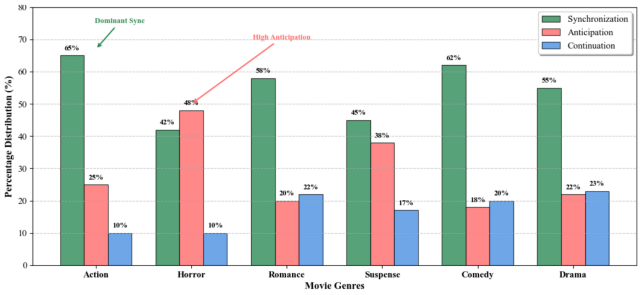

FIGURE 8. Alignment pattern distribution

The analysis of Figure 8 indicates that audio-visual timing strategies vary greatly by film genre. Horror films exhibit the greatest prevalence of the anticipation pattern (48%) compared to other genres, which aligns with the genre's inherent narrative organization of tension through music. Suspense films rely on the anticipation effect (38%) to establish audience's foresight and expectations. Action films are characterized by the synchronous pattern (65%), which is a reflection of action films' creative characteristic of reinforcing impact with synchronized audio-visual presentation of action. Comedies and romance films make up slightly less than action films with 62% and 58% occurrence of the synchronous pattern, thus indicating these films utilize audio-visual synchronization for amplifying emotional resonance or comedic punchline effect. Drama films boarder somewhere in the middle of the apparatus variations with a slightly balanced distribution (synchronous 55%, anticipation 22%, and continuation 23%), indicating that drama films respond the needs of a complex narrative. The continuation pattern accounts for a relatively high proportion in romance and drama films (22-23%), indicating that these films assign a greater importance on extended emotional resonance through music and extending thematic expression. These distributions not only confirm the theoretical assumption that genre is an important determining factor of audio-visual alignment strategies, but also offers significant quantitative contributions to account for the emotional regulation and creative formation of respective genres.

V. CONCLUSION

The paper presents V^{2}A-AlignNet, a cross-modal dynamic alignment network that integrates VideoMAE and CRNN. The model captures the spatiotemporal semantic evolution of the video and the rhythmic behavior of the audio in a dual-stream coding framework. Through a cross-modal attention mechanism, the model builds a joint semantic space, which produces continuous alignment saliency sequences and similarity matrices to effectively identify "when and how alignment occurs". In addition, it includes a type conditionalization mechanism, which adapts film genre information to identify audio-visual interaction patterns in different genres. The model achieves state-of-the-art performance on a large-scale dataset of six film genres for inflection point detection, alignment pattern recognition, and genre classification. Ablation studies indicate that VideoMAE, CRNN, cross-modal attention, and type conditionalization, each, is responsible for improving performance. This research builds an end-to-end deep learning architecture for dynamic audio-visual alignment, proposes a quantifiable and interpretable alignment saliency modeling approach, and systematically reveals the audio-visual alignment patterns of different film genres with large-scale data. This paper advances the development of computational film studies in cross-modal temporal modeling, providing theoretical support and technical pathways for AI-based film creation and analysis.

AUTHOR CONTRIBUTIONS

Conceptualization –Shungong Chi and Haiyu Liang; methodology – Shungong Chi and Haiyu Liang; investigation – Shungong Chi and Haiyu Liang; writing-original draft preparation – Shungong Chi and Haiyu Liang. 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.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] N. Elyamany, "A chronotopic approach to identity performance in musical numbers: a choreo-musical case study of 'rewrite the stars' and 'this is me'," Visual Communication, vol. 22, no. 2, pp. 278-296, 2023, doi: 10.1177/1470357220974069.

[2] A. K. Herget, "On music's potential to convey meaning in film: A systematic review of empirical evidence," Psychology of Music, vol. 49, no. 1, pp. 21-49, 2021, doi: 10.1177/0305735619835019.

[3] T. Behrouzi, R. Toosi, and M. A. Akhaee, "Multimodal movie genre classification using recurrent neural network," Multimedia Tools and Applications, vol. 82, no. 4, pp. 5763-5784, 2023, doi: 10.1007/s11042-022-13418-6.

[4] A. K. Herget, "Well-known and unknown music as an emotionalizing carrier of meaning in film," Media Psychology, vol. 24, no. 3, pp. 385-412, 2021, doi: 10.1080/15213269.2020.1713164.

[5] W. Q. Liu, M. X. Lin, H. B. Huang, C. Y. Ma, Y. Song, W. M. Dong, et al., "Emotion-Aware Music Driven Movie Montage," Journal of Computer Science and Technology, vol. 38, no. 3, pp. 540-553, 2023, doi: 10.1007/s11390-023-3064-6.

[6] M. J. Lucia-Mulas, P. Revuelta-Sanz, B. Ruiz-Mezcua, and I. Gonzalez-Carrasco, "Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises," Applied Intelligence, vol. 53, no. 22, pp. 27096-27109, 2023, doi: 10.1007/s10489-023-04967-w.

[7] S. G. Cunha and M. Lee, "Emotional Video to Audio Transformation Using Deep Recurrent Neural Networks and a Neuro-Fuzzy System," Mathematical Problems in Engineering, vol. 2020, no. 1, 8478527, 2020, doi: 10.1155/2020/8478527.

[8] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Music video emotion classification using slow–fast audio–video network and unsupervised feature representation," Scientific Reports, vol. 11, no. 1, 19834, 2021, doi: 10.1038/s41598-021-98856-2.

[9] J. Zhang, X. Wen, and M. Whang, "Recognition of emotion according to the physical elements of the video," Sensors, vol. 20, no. 3, 649, 2020, doi: 10.3390/s20030649.

[10] A. Sharma, K. Sharma, and A. Kumar, "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion," Neural computing and applications, vol. 35, no. 31, pp. 22935-22948, 2023, doi: 10.1007/s00521-022-06913-2.

[11] J. Tian and Y. She, "A visual–audio-based emotion recognition system integrating dimensional analysis," IEEE Transactions on Computational Social Systems, vol. 10, no. 6, pp. 3273-3282, 2022, doi: 10.1109/TCSS.2022.3200060.

[12] T. W. Kim and K. C. Kwak, "Speech emotion recognition using deep learning transfer models and explainable techniques," Applied Sciences, vol. 14, no. 4, 1553, 2024, doi: 10.3390/app14041553.

[13] S. Deng, L. Wu, G. Shi, L. H. Xing, W. J. Hu, and H. Zhang, "Simple but powerful, a language-supervised method for image emotion classification," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3317-3331, 2022, doi: 10.1109/taffc.2022.3225049.

[14] G. Shi, S. Deng, B. Wang, C. Feng, Y. Zhuang, and X. M. Wang, "One for all: A unified generative framework for image emotion classification," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 8, pp. 7057-7068, 2023, doi: 10.1109/TCSVT.2023.3341840.

[15] E. A. Iyilikci, A. Demirel, F. Isik, and O. Iyilkci, "How do sound and color features affect self-report emotional experience in response to film clips?," Current Psychology, vol. 43, no. 11, pp. 10185-10216, 2024, doi: 10.1007/s12144-023-05127-6.

[16] F. Liu, D. L. Chen, R. Z. Zhou, S. Yang, and F. Xu, "Self-supervised music motion synchronization learning for music-driven conducting motion generation," Journal of Computer Science and Technology, vol. 37, no. 3, pp. 539-558, 2022, doi: 10.1007/s11390-022-2030-z.

[17] S. Landry and M. Jeon, "Interactive sonification strategies for the motion and emotion of dance performances," Journal on Multimodal User Interfaces, vol. 14, no. 2, pp. 167-186, 2020, doi: 10.1007/s12193-020-00321-3.

[18] G. Bente, K. Kryston, N. T. Jahn, and R. Schmalzle, "Building blocks of suspense: subjective and physiological effects of narrative content and film music," Humanities and Social Sciences Communications, vol. 9, no. 1, pp. 1-13, 2022, doi: 10.1057/s41599-022-01461-5.

[19] E. M. Khan, M. S. Mukta, M. E. Ali, and J. Mahmud, "Predicting users' movie preference and rating behavior from personality and values," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 10, no. 3, pp. 1-25, 2020, doi: 10.1145/3338244.

[20] J. D. Bradbury and R. E. Guadagno, "Documentary narrative visualization: Features and modes of documentary film in narrative visualization," Information Visualization, vol. 19, no. 4, pp. 339-352, 2020, doi: 10.1177/1473871620925071.

[21] B. N. Alkahtani, N. Alsraisri, and A. J. Almalki, "Achieving State-of-the-art Accuracy in Arabic Sign Language Recognition with VideoMAE and Data Augmentation," Journal of Disability Research, vol. 4, no. 5, 20250661, 2025, doi: 10.57197/JDR-2025-0661.

[22] A. Berroukham, M. Lahraichi, and K. Housni, "Self-Supervised Method for Risky Situation Detection in Road Traffic Sequences Using Video Masked Autoencoder," International Journal of Advanced Computer Science & Applications, vol. 16, no. 6, pp. 559-559, 2025, doi: 10.14569/ijacsa.2025.0160654.

[23] M. Suresha, S. Kuppa, and D. S. Raghukumar, "A study on deep learning spatiotemporal models and feature extraction techniques for video understanding," International Journal of Multimedia Information Retrieval, vol. 9, no. 2, pp. 81-101, 2020, doi: 10.1007/s13735-019-00190-x.

[24] M. S. Islam, S. Sultana, R. U. Kumar, and J. A. Mahmud, "A review on video classification with methods, findings, performance, challenges, limitations and future work," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 6, no. 2, pp. 47-57, 2020, doi: 10.26555/jiteki.v6i2.18978.

[25] Y. Mao, G. Zhong, H. Wang, and K. Huang, "Music-CRN: An efficient content-based music classification and recommendation network," Cognitive Computation, vol. 14, no. 6, pp. 2306-2316, 2022, doi: 10.1007/s12559-022-10039-x.

[26] A. Bansal and N. K. Garg, "Robust technique for environmental sound classification using convolutional recurrent neural network," Multimedia Tools and Applications, vol. 83, no. 18, pp. 54755-54772, 2024, doi: 10.1007/s11042-023-17066-2.

[27] J. Y. Kwak and Y. J. Chung, "Audio Event Detection Based on Attention CRNN," The Journal of the Korea institute of electronic communication sciences, vol. 15, no. 3, pp. 465-472, 2020, doi: 10.13067/JKIECS.2020.15.3.465.

[28] M. P. Fatah, "CRNN Algorithm and MFCC Feature Extraction in Classifying Hijaiyah Letter Pronunciation: A Systematic Literature Review," Khazanah Journal of Religion and Technology, vol. 3, no. 1, pp. 31-35, 2025, doi: 10.15575/kjrt.v3i1.1555.

[29] 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.

[30] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," Journal of Information and Optimization Sciences, vol. 44, no. 6, pp. 1057-1074, 2023, doi: 10.47974/jios-1431.

[31] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, et al., "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," ACM Transactions on Multimedia Computing Communications, and Applications (TOMM), vol. 17, no. 4, pp. 1-23, 2021, doi: 10.1145/3451390.

[32] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image–text matching," IEEE transactions on neural networks and learning systems, vol. 31, no. 12, pp. 5412-5425, 2020, doi: 10.1109/TNNLS.2020.2967597.

[33] L. Li, T. Jin, W. Lin, H. Jiang, W. W. Pan, and J. Wang, "Multi-granularity relational attention network for audio-visual question answering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 8, pp. 7080-7094, 2023, doi: 10.1109/TCSVT.2023.3264524.

[34] L. Ye, M. Rochan, Z. Liu, X. Q. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3719-3732, 2021, doi: 10.1109/TPAMI.2021.3054384.

[35] P. Wei, F. He, L. Li, and J. Li, "Research on sound classification based on SVM," Neural Computing and Applications, vol. 32, no. 6, pp. 1593-1607, 2020, doi: 10.1007/s00521-019-04182-0.

[36] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. Dragoni, "Violence detection in videos by combining 3D convolutional neural networks and support vector machines," Applied Artificial Intelligence, vol. 34, no. 4, pp. 329-344, 2020, doi: 10.1080/08839514.2020.1723876.