Representation Learning through Multimodal Attention and Time-Sync Comments for Affective Video Analysis

Although temporal patterns inherent in visual and audio signals are crucial for affective video content analyses, they have not been thoroughly explored yet. In this paper, we propose a novel temporal-aware multimodal (TAM) method to fully capture the temporal information. Specifically, we design a cross-temporal multimodal fusion module that applies attention-based fusion to different modalities within and across video segments. As a result, it fully captures the temporal relations between different modalities. A single emotion label lacks supervision for learning representations’ segments, making temporal pattern mining difficult. We leverage time-synchronized comments (TSC) as auxiliary supervision, since these comments are easily accessible and contain rich emotional cues. Two TSC-based self-supervised tasks are designed: the first aims to predict the emotional words in a TSC from video representation and TSC contextual semantics, and the second predicts the segment in which the TSC appears by calculating the correlation between video representation and TSC embedding. These self-supervised tasks are used to pre-train the cross-temporal multimodal fusion module on a large-scale video-TSC data set, which is crawled from the web without labeling costs. These self-supervised pre-training tasks prompt the fusion module to perform representation learning on segments including TSC, thus capturing more temporal affective patterns. Experimental results on three benchmark data sets show that the proposed fusion module achieves state-of-the-art results in affective video content analyses. Ablation studies verify that after TSC-based pre-training, the fusion module learns more segments’ affective patterns and achieves better performance.



The left is a Titanic clip at 1min56s; the right shows the time-sync comments appearing between 1min51s and 1min55s. Several viewers commented on feelings such as romance, awe, love, and freedom.

# Introduction

Affective video content analyses aim to predict the emotions viewers may feel when watching videos. The development of video-sharing websites has led to the proliferation of videos on social websites such as YouTube and Bilibili. Managing this vast number of videos is a challenge. Classifying videos by their induced emotion is a good solution with multiple benefits. First, video retrieval through emotional keywords is easy and accurate. Website managers can also utilize the induced emotions of the videos to enhance the recommendation system and improve viewers’ experience. Thirdly, understanding affective video content can help video generators make more appealing videos. Affective video content analyses could also be used to detect extremely emotional videos, allowing government regulators to take preventive measures quickly.

While computer vision methods have become more advanced, affective video content analyses still face several challenges. First, the visual and audio information need to be fused efficiently, since both provoke the viewer’s emotions. Existing methods mainly adopt decision-level or feature-level fusion to integrate visual and audio signals. The former directly combines analysis results from visual and audio features, ignoring dependencies across modalities. The latter typically captures the temporal features within each modality and fuses them into a joint feature. However, since most feature-level fusion methods capture temporal and cross-modal dependencies separately, they don’t fully capture the temporal dependencies between visual and audio signals. This paper introduce a cross-temporal multimodal fusion module to solve this issue. The proposed fusion module applies the self-attention operation to different modalities within each video segment as well as across segments. Consequently, all temporal dependencies across visual and audio modalities are exploited for affective video content analyses.

Current one-label-per-video supervision is insufficient, since affective states vary within different segments of the same video. For instance, a surprise video could end with a happy atmosphere setting. This kind of affective transition can be used as a temporal pattern for affective content understanding. However, the supervision provided by a single emotional label is too limited to describe the affective states of all video segments, making it difficult to mine temporal patterns. We address this by introducing time-sync comments (TSC) as auxiliary supervision to pre-train the cross-temporal multimodal fusion module. Time-sync comments are brief timestamped viewer comments containing emotional feelings. As shown in Fig. [1](#fig:example), the TSC in this famous movie clip express multiple emotions aroused by this scene: romance, awe, love, and freedom. TSC can provide emotional and temporal cues for affective video content analyses. We design two TSC-based self-supervised tasks: emotional word prediction and appearing time prediction. The former uses the video representation extracted by the fusion module and TSC contextual semantics to predict the emotional words in TSC, and the latter predicts which TSC goes with which segment by calculating the similarities between video representation and TSC embedding. These two tasks make full use of TSC semantics and temporal cues to enhance segment-level representation learning. Videos with intensive TSC are easily accessible on the Internet. The proposed fusion module is pre-trained on a large-scale video data set collected from the web, requiring no manual annotation. The TSC are discarded during affective video analysis inference.

The temporal-aware multimodal method is evaluated on three popular benchmark data sets: VE-8, YF-E6, and LIRIS-ACCEDE. Experimental results demonstrate that the fusion module achieves state-of-the-art results, extracting more discriminative representation after TSC-based pre-training.

The contributions of the proposed temporal-aware multimodal (TAM) method are as follows:

* We design a cross-temporal multimodal fusion module to learn the temporal dependencies between visual and audio signals for affective video content analyses.
* We propose two TSC-based self-supervised pre-training tasks to enhance segment-level representation learning.
* We conduct extensive experiments on three affective video content data sets to demonstrate the effectiveness of the proposed method.

# Related Work

## Affective Video Content Analyses

Multimodal fusion of affective video content analyses can be divided into two types: decision-level fusion and feature-level fusion. Decision-level fusion combines the results from different classifiers, ignoring the dependencies between visual and audio features. For example, Acar et al. learned visual and audio features separately and then employed three multiclass support vector machines to obtain affective predictions. Feature-level fusion combines visual and audio features and feeds them jointly to a classifier or regressor. For example, Xu et al. selected each modal feature based on emotional concepts, and summed these selected features into a joint feature for emotion classification. Qiu et al. simply concatenated action and scene features as a whole and then input them into a dual attention network to focus on emotion-related frames. Wei et al. also concatenated object and scene features, with a focus on estimating the affective saliency value of frames. Zhao et al. integrated spatial, channel-wise, and temporal attention into a visual extractor, and temporal attention into an audio extractor, then concatenated the output visual and audio features into a joint emotional feature. However, these methods use a simple fusion strategy, summing or concatenating the visual and audio features while ignoring inherent interactions among them.

Some feature-level fusion methods use complex strategies to mine the dependencies between visual and audio signals. For instance, Gan et al. used a deep regression Bayesian network to capture the high-order dependencies between low-level visual and audio features, ignoring temporal patterns. Qi et al. used an attention mechanism to aggregate the temporal features in each modality, then aligned the visual and audio features by jointly mapping them into a common space. Mittal et al. used an LSTM encoder to learn the temporal features and a co-attention mechanism to calculate correlation scores between the pairwise modalities. They then weighted and summed these multimodal temporal features using correlation scores. Gao et al. proposed a synchronous modal-temporal attention block to capture the visual and audio relations within each moment, then used LSTM to learn the temporal dependencies within each modality. However, these methods either discard temporal relations or learn temporal and cross-modal dependencies separately, thereby ignoring inherent dependencies between temporal elements of visual and audio signals, which are essential for effective multimodal fusion in videos.

To mine these dependencies, we design a cross-temporal multimodal fusion module that employs self-attention to learn the pairwise modalities’ relations within each video segment and across different segments. This module simultaneously learns relations between all the video segments of different modalities, so it can fully capture the temporal dependencies between visual and audio signals.

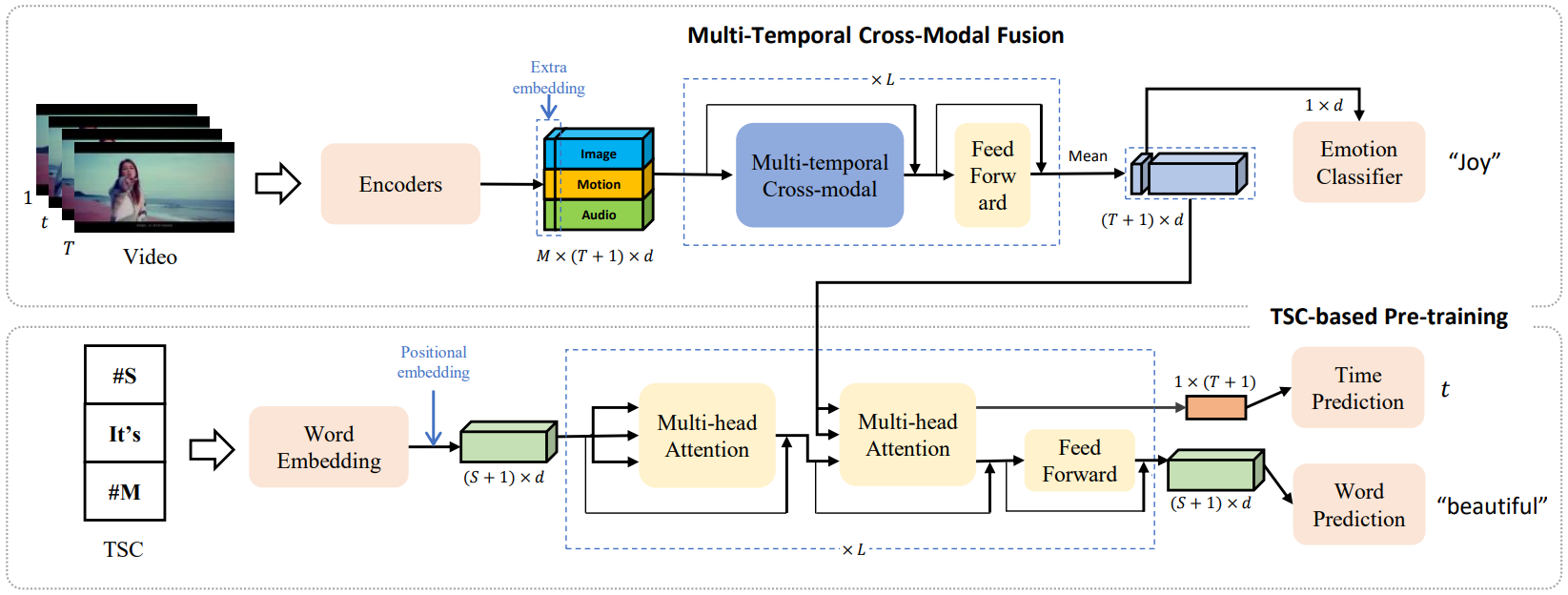
## Video Analyses with TSC

TSC has great research significance for video understanding. It can be used for video tagging, video description, and video recommendation, among other things. TSC are viewers’ real-time emotional expressions to video content. Their semantics and temporal cues indicate the induced emotion of the video and the moment of the emotion burst. Hence, TSC has great potential for affective video content analyses. However, this potential has not been successfully explored. To the best of our knowledge, only Li et al. leveraged TSC for video emotion recognition. They first utilized canonical correlation analysis to maximize the mutual information between visual and TSC textual features, and then used LSTM to separately capture temporal dependencies within visual and TSC modalities. The output features of visual and TSC are concatenated as a joint feature for video emotion recognition. However, their method requires emotion-labeled video-TSC data during training, and the expensive labeling costs limit the number of training samples. Moreover, this method needs to simultaneously input video and TSC in the inference stage, limiting its application in video analyses.

We design two self-supervised pre-training tasks to mine the semantic and temporal cues from TSCs to enhance segment-level representation learning. These tasks are designed to force the cross-temporal multimodal fusion module to learn the representation of segments in which TSCs appear. Specifically, the emotional word prediction task uses video representation and TSC contextual semantics to fill the masked TSC with emotional words. The appearing time prediction task computes the similarities between video representation and TSC embedding to predict the segment in which the TSC appears. We collect a large-scale video-TSC data set to pre-train the cross-temporal multimodal fusion module without manual annotation. We also fine tune the pre-trained fusion module on affective video data sets for affective video analyses. TSC are not needed in the fine-tuning and inference stages.

# Problem Statement

Suppose we have an affective video data set and a video-TSC data set . The contains videos with affective labels . The contains videos with TSC sets. is the TSC set of the -th video. is the appearing time of the -th TSC. is the maximum number of words in the TSC. is the number of TSCs in . Our goal is to pre-train a network in a self-supervised manner on the video-TSC data set , then fine tune this network with the affective labels on the affective video data set . Only videos are inputted into the network to predict the affective labels during inference.



[fig:framework]

Figure 2: The overall structure of the temporal-aware multimodal method. Our method can be divided into two parts: cross-temporal multimodal fusion and TSC-based pre-training. Image, motion, and audio features are extracted from three video encoders. The fused feature can be directly used for emotion classification or regression.

# Methodology

In this section, we introduce the proposed temporal-aware multimodal (TAM) method in detail. As shown in Figure [[fig:framework]](#fig:framework), the proposed TAM method consists of two parts: cross-temporal multimodal fusion and TSC-based pre-training. The former helps capture the temporal relations between different modalities, while the latter designs two self-supervised tasks, forcing the backbone to fully capture the affective pattern throughout video segments.

As shown in the upper part of Fig. [[fig:framework]](#fig:framework), each video is divided into segments for affective video content analysis. For the -th video segment, we use video encoders to extract the image, motion, and audio features . The features of segments and modalities are stacked as the video features , where represents the feature dimension. Finally, a fusion module merges the features of different modalities and segments in into , and then inputs into a classifier or regressor to predict .

## Cross-Temporal Multimodal Fusion

For efficient affective video content analyses, the fusion module needs to fully capture the correlation between the features of different modalities and temporal segments. A recent multi-head self-attention operation in Transformer achieves outstanding performance in various vision tasks by learning long-range dependencies between sequences. Inspired by this, we adapt a variant of the self-attention operation to learn both the temporal and modal dependencies to fully explore the video information.

Specifically, our cross-temporal multimodal fusion module consists of an inner-modal attention and a cross-modal attention, as shown in Fig. [2](#fig:attention). First, similar to ViT , we pretend an extra learnable embedding to the -th modal sequence for learning the comprehensive video feature, and add a position embedding to each modal sequence to retain temporal information as follows:

where represents the -th modality feature and the -th segment feature, respectively, and is the matrix concatenation along the first dimension. Then, the extracted features of each modality are mapped into query, key, and value domains:

where are learnable parameter matrices for the -th modality and , , and are the query, key, and value matrices respectively.

The inner-modal attention encodes the temporal correlations between the features within the same modality by matching their query and key matrices. The inner-modal fusion results are calculated as:

where is the transposed key matrix and is the inner-modal fusion result of the -th modal features. Following the vanilla self-attention , is used as a normalization factor to keep from the region with an extremely small gradient.

Cross-modal attention aims to capture the temporal correlation between pairwise features of different modalities. We match the query matrix of the -th modality with the key matrices of all other modalities to learn the correlation weights, and then average the weighted features from different modalities to get cross-modal fusion results, as follows:

where is the transposed key matrix and is the cross-modal fusion result of the -th modal features.

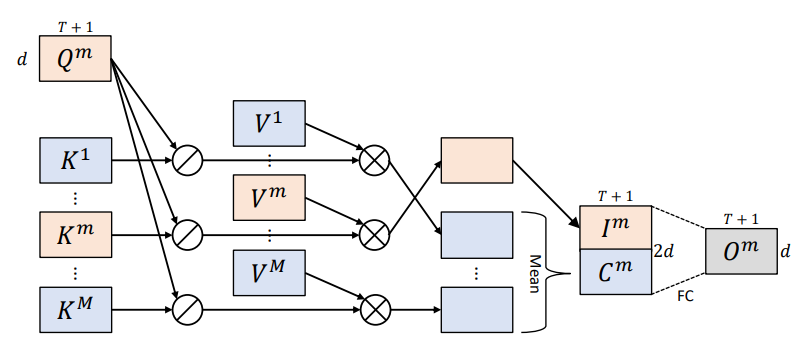
The multi-temporal cross-modal fused features are obtained by concatenating the inner-modal attention and cross-modal attention as follows:

where the represents matrix concatenation along the last dimension, is a learnable parameter matrix, and is the fused feature. We fully capture the long-range dependencies between the features of different modalities and temporal segments by stacking the proposed cross-temporal multimodal fusion (CTM) module times, along with feed-forward network (FFN):

where . is the layer normalization.

Finally, we apply the mean operation along the modal dimension to obtain the affective features and apply a classifier or regressor on the first temporal vector to predict the emotion result for the video:

The loss is calculated as:



The cross-temporal multimodal fusion module consists of cross-modal and inner-modal attentions. is matrix multiplication. FC represents the fully connected layer.

[fig:attention]

## TSC-Based Pre-Training

We predict the emotion label of the video by fusing the features of different modalities and temporal segments. The cross-temporal multimodal fusion module is supervised by the target emotion label during training. However, the supervision provided by a single emotional label is insufficient to describe the affective states throughout video segments, making temporal pattern mining difficult. To address this problem, we design two TSC-based self-supervised tasks to pre-train the fusion module for improved mining of temporal affective patterns.

The procedure of the TSC-based self-supervised pre-training is shown in the lower part of Fig. [[fig:framework]](#fig:framework). Specifically, we first prepend a special word to each TSC in order to learn the correlation between TSC semantics and the video feature. The prepended word sequence is converted into a TSC embedding via a word embedding layer:

where is a positional embedding. Then this TSC embedding and the video feature extracted by the cross-temporal multimodal fusion module are input into a TSC decoder. Like the transformer decoder, the TSC decoder is stacked by identical decoding blocks, each of which consists of two multi-head self-attention (MSA) layers and a feed-forward network (FFN):

where . At the -th decoding block, the attention output and the intermediate weight are used for further prediction. The intermediate attention weight is calculated as:

where is the query of the special word and is the key mapping parameter matrix in the second MSA. The intermediate attention weight represents the correlation between the TSC semantics and video segment features.

We utilize the extracted and to design two TSC-based self-supervised pre-training tasks: emotional word prediction and appearing time prediction. Details about these two tasks are explained below.

**Emotional Word Prediction** The masked language model (MLM) is highly successful at learning sentence semantics. Inspired by this, we design an emotional word prediction task based on MLM. This task uses the guidance of the video features and the features from other non-masked words to predict the masked word in each TSC. Unlike the random masking strategy of the original MLM, we focus on emotional content learning and only mask emotional words in each TSC. Specifically, before entering the word embedding layer, the TSC sequence is masked by replacing the emotional words with a special word :

Then the masked sentence is input into the TSC decoder claimed in Eq. [[eqtsc]](#eqtsc) - [[final]](#final) to get the attention output . Finally, a multilayer perceptron regressor is applied to predict the masked words from the attention output as follows:

where is a parameter matrix, is the total number of emotional words in the affective lexicon, and are the predicted words in probability vector format. The loss of emotional word prediction is:

where is the cross-entropy loss function, is the prediction of the -th masked word, and is the ground truth of the -th masked word in one-hot format.

**Appearing Time Prediction** Temporal level video representation learning is enhanced by aligning the TSC embedding with the video segment feature through the appearing time prediction task. This task aims to predict the appearing temporal segment of the TSC. Along with the sentence inputted to the TSC decoder claimed in Eq. [[eqtsc]](#eqtsc) - [[final]](#final), the intermediate weight at the final decoding block is extracted for further prediction. The weight represents the correlation between the TSC and the features of the video segments, as shown in Eq. [[attn]](#attn). Therefore, we take as the time prediction result . Compared to the ground truth appearing segment , the loss of appearing time prediction is as shown in Eq. [[tpeq]](#tpeq).

**Summary** The emotional word prediction task utilizes the abundant emotional cues in TSC as auxiliary supervision for multimodal representation learning, while the appearing time prediction task forces this learning to focus on the appearing temporal segment of TSC. Consequently, the video encoders and fusion modules learn to fully capture affective information throughout the temporal segments.

The total loss of the TSC-based pre-training is formulated as:

where are two tradeoff parameters that weigh the importance of these two tasks.

# Experiments

This section first introduces the three public benchmark data sets and the self-collected video-TSC data set. Then we perform several ablation studies to verify the effectiveness of our design. Lastly, we compare the proposed method to other state-of-the-art methods.

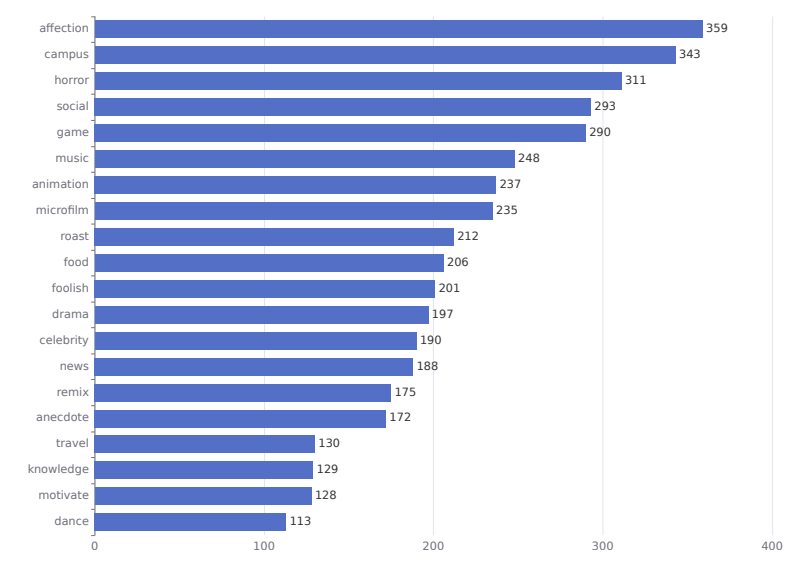
## Data Sets

The VideoEmotion-8 (VE-8) data set is collected from YouTube and Flickr, and it contains a total of 1,101 videos with an average duration of 107 seconds. These videos are labeled with one of eight emotions: *anger, anticipation, disgust, fear, joy, sadness, surprise,* or *trust*. Each category contains a minimum of 100 videos. We follow the common experimental setup, randomly splitting the data set for ten runs with 2/3 training and 1/3 testing, and report the average results of the ten runs.

The YouTube/Flickr-EkmanSix(YF-6) data set consists of 1,637 videos collected from YouTube and Flickr. The average duration is 112 seconds. The videos are labeled with one of six basic emotion categories: *anger, disgust, fear, joy, sadness,* or *surprise*. There are at least 221 videos in each category. We use the public splitting of 819 videos for training and 818 for testing.

The LIRIS-ACCEDE data set is the largest data set for video affective analysis, consisting of 9,800 videos extracted from 160 movies. The videos last between 8 and 12 seconds. There are two tasks based on this data set. In the MediaEval2015 affective impact of movies task, a total of 10,900 (1,100 additional) videos are used for classification. These videos are split into 6,144 videos for training and 4,756 videos for testing. For each video, the ground truth consists of the arousal class (calm-neutral-active) and the valence class (negative-neutral-positive). MediaEval2016 emotional impact of movies is a regression task and includes 11,000 videos (1,200 additional) split into 9,800 training videos and 1,200 testing videos. Each video has the absolute affective scores of valence and arousal.

The Video-TSC data set is collected from the Chinese video website Bilibili.[[1]](#footnote-1) We crawled 7,000 videos containing intensive TSC from the life, short film, and popularity sections published between January 2018 and December 2021. Most of the TSC are brief Chinese sentences with around 10 words. The distribution of the top twenty video tags is shown in Fig. [3](#fig:tags), and reveals a wide variety of video content. We divide these videos into segments to facilitate analyses. Specifically, we use the -means algorithm to cluster the appearing time of the TSC, with each cluster corresponding to the video segment to be split. For each video, we set as the video duration divided by 30 so that each video segment is around 30 seconds. We also consider the equilibrium of evoked emotions in these videos. Using the Chinese sentiment analysis model Ernie, the sentiment score (ranging from 0 to 1, negative to positive) of each TSC is predicted and the average of all the TSC sentiment scores in the video segment are used as the final sentiment score. These sentiment scores are only used to pick out an equal number of positive and negative videos, and are not used in the training stages. We selected the 8,000 most negative and 8,000 most positive video segments, with a total average duration of 27.6 seconds. There are a total of 163,411,023 TSCs in these video segments.



Distribution of the Top 20 Tags of the Crawled Videos

[fig:tags]

## Implementation Details

**Multimodal Video Encoders** Video features are extracted using three video encoders: CLIP-enhanced ViT for image features, ResNet3D for motion features, and VGGish for audio features. The three models are pre-trained on large-scale data sets via image text aligning, action recognition, and audio classification tasks. Specifically, we choose Base Vision Transformer with 32 layers (ViT-B/32) pre-trained by CLIP as the image encoder. The input image is pixels, and the output is a 512-dimensional feature. The ResNet3d needs 16 consecutive 112 x 112 frames as the input and outputs a 2048-dimensional feature. The audio is first converted to the Mel-frequency cepstral coefficient (MFCC), then input into the VGGish model to obtain a 128-dimensional feature. These outputs are input into fully connected layers to be separately mapped to the same 768-dimensional space. The parameters of these three encoders do not participate in gradient optimization.

**Multimodal Fusion** In this phase, the effectiveness of the proposed cross-temporal multimodal fusion module is verified by direct training on the affective video data set. For each video, we randomly sample consecutive video segments, each containing 16 frames. All frames and audio MFCCs corresponding to the segments are directly input into the motion and audio encoders. In each segment, one frame is selected and input into the image encoder. Models are optimized using an Adam optimizer with a learning rate of and weight decay of . The models are trained with batch size 8 for 100 epochs and for classification tasks on VE-8 and YF-6. For MediaEval2015 and MediaEval2016 tasks, the batch size is 16 and the training epoch is 20, and . For each task, we randomly split 15% of the training set off as a validation set to choose the best hyperparameters. We compare our cross-temporal multimodal fusion strategy to three other fusion strategies:

* **Simple concatenation** This fusion module is stacked with 12 identical layers. Each layer consists of an eight-head self-attention, a feed-forward network, and the residual connection around every two sublayers. The image, motion, and audio features are separately input into three parallel fusion modules. Then the first outputs of these three modules are concatenated as the fused feature. This simple concatenation strategy uses the attention mechanism to focus on temporal relations within each modality, but disregards the interactions between different modalities.
* **Divide attention** This fusion module is formed by inserting an eight-head self-attention between the two sublayers of the simple concatenation fusion module. First, the temporal features of each modality are input into the first self-attention. Then, the multimodal output features of each segment are input into the second self-attention. Thus the temporal relations and cross-modal dependencies are separately and consecutively captured.
* **Joint attention** This fusion module is the same as the simple concatenation fusion module. First, all temporal features of the three modalities are concatenated as a joint sequence. Then the joint sequence is input into the fusion module for full pairwise attention between segments and modalities. This fusion strategy mixes up the temporal and cross-modal relations.

**TSC-Based Pre-Training** The TSCs are filtered using the Chinese emotion dictionary labeled with seven emotions. The dictionary contains 27,466 words, and only TSCs containing one or more of these words are reserved. Removing the repeated TSCs per second yields 891,400 TSCs for video-TSC pre-training. Each TSC is tokenized by the Chinese tool Jieba[[2]](#footnote-2). In the pre-training phase, we randomly divide the video-TSC data set into training and testing sets at a ratio of 4:1. An Adam optimizer is used and the total training epoch is 200; the initial learning rate is 0.001, and it decays by 0.1 every 50 epochs.

## Ablation Study

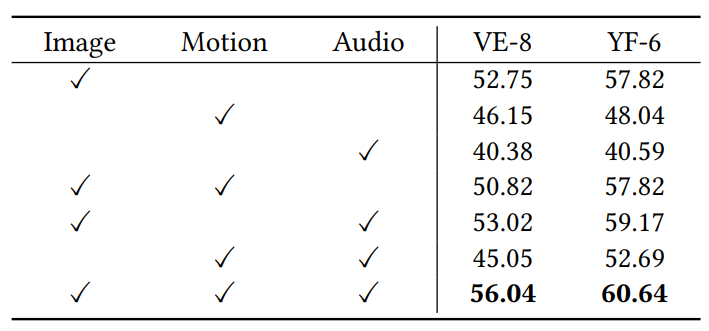
**Different Combinations of Modalities** We perform an ablation study featuring different combinations of the modal features to show the necessity of each. Table [1](#tab:Comparison2) shows the video emotion recognition accuracy of different combinations of modalities. Combining three modalities achieves the best performance of 56.04% and 60.64% on VE-8 and YF-6. The smallest gap between all three modalities combined and the other combinations is 3.02% and 1.47%. These results verify that all three modalities contribute to the performance of the model, and each is essential to design an effective multimodal fusion module.

**Different Multimodal Fusion Design** To validate the effectiveness of the proposed cross-temporal multimodal (CTM) fusion module, we perform an ablation study on different multimodal fusion designs. Specifically, the fusion strategies mentioned in Sec. [[other]](#other) are set as the baselines and compared to the CTM module on VE-8 and YF-6 data sets. Results are shown in the upper part of Table [2](#tab:Comparison1).

From the table, we can make two major observations. First, our CTM fusion module achieves the highest emotion classification accuracy on both VE-8 and YF-6. It outperforms the simple concatenation and divide attention fusion modules by a large margin. This is because these modules ignore the temporal dependencies between different modalities. Secondly, by capturing all relations between video segments and modalities, joint attention fusion achieves admirable performance as well. However, our CTM fusion module still has a 1.92% and 0.36% advantage over it. This is likely because the joint attention fusion module employs the full pairwise attention on all temporal and modal features, and the redundancy of visual and audio information weakens its capability.

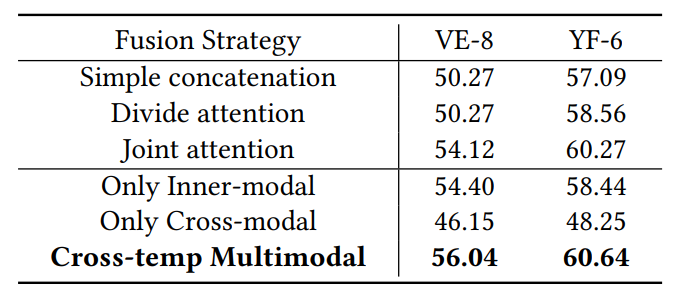
[tab:Comparison2]

Video Emotion Recognition Accuracy (%) with Different Modalities



[tab:Comparison1]

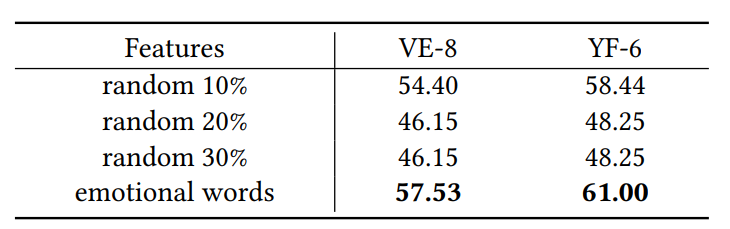
Video Emotion Recognition Accuracy (%) with Different Multimodal Fusion Modules

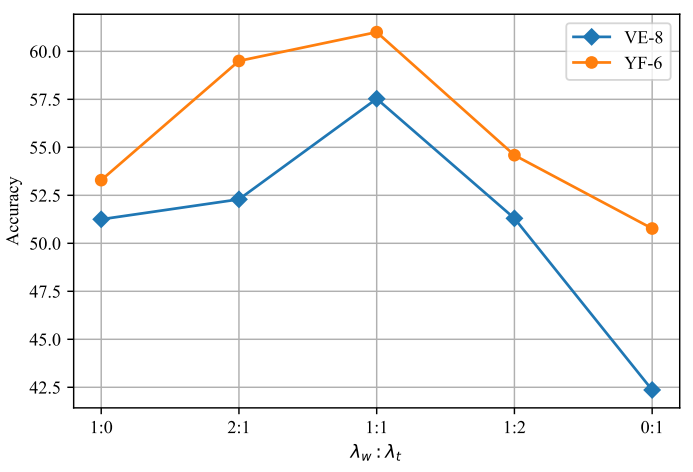


**Inner- and Cross-Modal Attention** Ablation experiments are carried out on the VE-8 and YF-6 data sets to analyze the contributions of different attentions of our fusion module. The lower part of Table [2](#tab:Comparison1) shows the results of inner-modal attention only, cross-modal attention only, and the combination of these attentions. The inner-modal attention strategy aims to mine the temporal dependencies within each modality, and the cross-modal one aims to capture the temporal relations across different modalities. Combining these two attentions yields the highest accuracies of 56.04% and 60.64% on VE-8 and YF-6, which is 1.64% and 2.2% higher than the inner-modal attention, as well as 9.89% and 12.39% higher than the cross-modal attention. Although the cross-modal attention is not competitive compared to the inner-modal attention, it complements inner-modal dependencies. This verifies that both inner-modal and cross-modal attentions contribute to the performance of the cross-temporal multimodal fusion module.

[tab:Comparison6]

Video Emotion Recognition Accuracy (%) with Different Masking Strategies





Experimental results with different parameter ratios

[fig:ratio]

**Different TSC Masking Strategies** To quantitatively verify the rationality of the emotional word prediction task, we compare the experimental results of different TSC masking strategies by randomly masking 10%, 20%, or 30% of the words, and by masking all emotional words in each TSC.

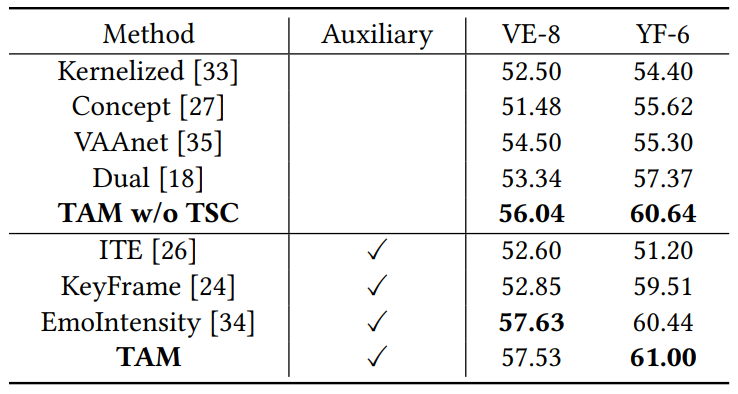
As shown in Table [3](#tab:Comparison6), as the number of masked words increases, the pre-trained representation performs more poorly on video emotion recognition tasks. That is because text semantics are reduced, so the pre-training task must emphasize text semantic understanding. The emotional word masking strategy yields the best performance on video emotion recognition, indicating that our word prediction task is emotion oriented and effective in emotional representation learning.

**Weighting the Pre-training Tasks** As seen in Eq. [[eq: final loss]](#eq: final loss), the importance of the two TSC-based pre-training tasks (emotional word prediction and appearing time prediction) are weighted by two parameters and . We perform an ablation study on different ratios of and ; experimental results are shown in Fig. [4](#fig:ratio).

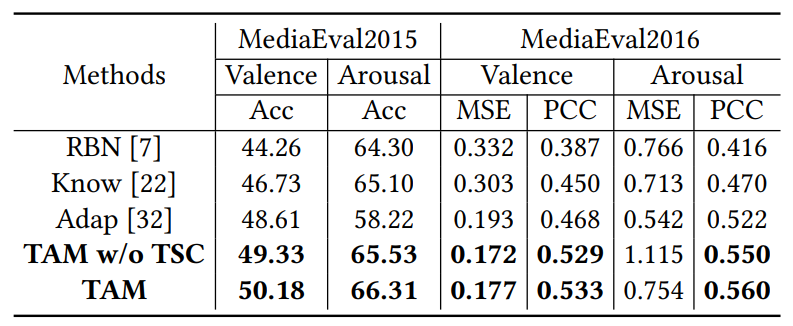
This figure reveals two trends. First, the performance increases first and then decreases as the proportion of the second loss grows. It reaches a peak when : is 1:1, and the emotional word and appearing time prediction tasks contribute equally to affective video representation learning. Secondly, there is a performance gap between the model with a single pre-training task and the model with both tasks. This is because the two TSC-based pre-training tasks have different focuses. The emotional word prediction task provides auxiliary supervision for multimodal representation learning, while the appearing time prediction task forces the learning to focus on the appearing temporal segment of TSC.

[tab:Comparison3]

Video Emotion Recognition Accuracy (%) Comparison with Other Methods. Auxiliary indicates if other data sets are used for training. TAM w/o TSC indicates the proposed TAM method without TSC-based pre-training.



LIRIS-ACCEDE Performance Compared to Other Methods. TAM w/o TSC stands for the proposed TAM method without TSC-based pre-training.



## Comparisons to State-of-the-Art Methods

We compare our results to top methods on two user-generated video data sets, i.e., VE-8 and YF-6, as well as the movie data set LIRIS-ACCEDE. The emotion recognition methods for user-generated videos include Kernelized, Concept, VAAnet, Dual, ITE, KeyFrame, and EmoIntensity. The affective analysis methods for movies include RBN, Know, and Adap. These methods mainly perform affective analysis on one type of video. However, both user-generated videos and movies contain common affective patterns, so experiments are conducted on both kinds of video.

**Affective video analyses without TSC-based pre-training** The top of Table [4](#tab:Comparison3) indicates that the proposed method performs well on the VE-8 and YF-6 data sets. The proposed method achieves 56.04% and 60.64% accuracy respectively, which is 1.54% and 3.27% higher than state-of-the-art methods lacking auxiliary data. Table [5](#tab:Comparison4) shows the effectiveness of our method on the affective movie analysis. Specifically, on the MediaEval2015 classification task, our method achieves an accuracy of 49.33% for valence and 65.53% for arousal, which is 0.72% and 0.43% higher than other methods. On the MediaEval2016 regression task, the MSE is 0.172 for valence and 1.115 for arousal, and the PCC is 0.529 for valence and 0.550 for arousal. Comprehensively, our multimodal attention model outperforms nearly all state-of-the-art methods on every task, including some methods with auxiliary data. Our method surpasses the simple fusion and cross-modal fusion methods, showing the effectiveness of our cross-temporal multimodal fusion module on affective video content analyses.

**Affective video analyses with TSC-based pre-training** As shown in Table [4](#tab:Comparison3), our temporal-aware multimodal method achieves 57.53% and 61.00% accuracy on VE-8 and YF-6. These scores are 1.49% and 0.36% higher after TSC-based pre-training. It also shows improvement on both the MediaEval2015 classification task and the MediaEval2016 regression task after TSC-based pre-training, as shown in Table [5](#tab:Comparison4). These results demonstrate that TSC can effectively enhance representation learning for affective video content analyses. Our pre-training framework also does not require additional emotional labels, while other methods transfer knowledge from emotion-labeled data sets.

# Conclusion

In this work, we propose an effective temporal-aware multimodal method for affective video content analyses. This TAM method consists of cross-temporal multimodal fusion and TSC-based pre-training. The cross-temporal multimodal fusion module employs attention-based fusion to different modalities within each segment and across different segments. As a result, it fully captures the temporal dependencies between visual and audio signals. The experimental results show that the fusion module can mine more temporal patterns across all modalities for affective video content analyses. Moreover, we leverage TSC to pre-train the fusion module for temporal-level representation learning, since TSC are easily accessible and contain affective cues. Two self-supervised tasks are used to pre-train the fusion module. The emotional word prediction task predicts the emotional words in a TSC under the guidance of video representation and TSC semantics. The appearing time prediction task aims to predict when the TSC appears by calculating the similarities between video representation and TSC embedding. These pre-training tasks successfully mine TSC emotional cues and use those cues as temporal-level supervision for representation learning. Comparison experiments verify that the cross-temporal multimodal fusion module can learn more discriminative representation after TSC-based pre-training.

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)