Infer Induced Sentiment of Comment Response to Video: A New Task, Dataset and Baseline

Qi Jia[1] **Baoyu Fan**[2,1*] **Cong Xu**[1] **Lu Liu**[1] **Liang Jin**[1] **Guoguang Du**[1] **Zhenhua Guo**[1]

Yaqian Zhao[1] **Xuanjing Huang**[3] **Rengang Li**[1]
[1]IEIT SYSTEMS Co., Ltd. [2]College of Computer Science, Nankai University, Tianjin, China,
[3]School of Computer Science, Fudan University
{jiaqi01, fanbaoyu, xucong, liulu06, jinliang, duguoguang}@ieisystem.com,
{guozhenhua zhaoyaqian}@ieisystem.com, xjhuang@fudan.edu.cn, lirg@ieisystem.com

Abstract

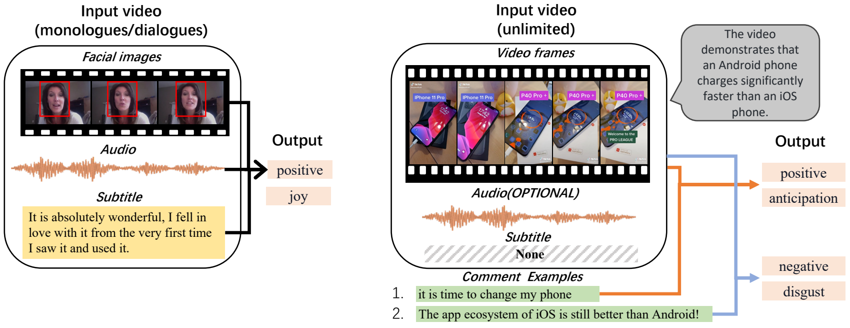
Existing video multi-modal sentiment analysis mainly focuses on the sentiment expression of people within the video, yet often neglects the induced sentiment of viewers while watching the videos. Induced sentiment of viewers is essential for inferring the public response to videos and has broad application in analyzing public societal sentiment, effectiveness of advertising and other areas. The micro videos and the related comments provide a rich application scenario for viewers' induced sentiment analysis. In light of this, we introduces a novel research task, **M**ulti-modal **S**entiment **A**nalysis for **C**omment **R**esponse of **V**ideo **I**nduced(**MSA-CRVI**), aims to infer opinions and emotions according to comments response to micro video. Meanwhile, we manually annotate a dataset named **C**omment **S**entiment toward to **M**icro **V**ideo (**CSMV**) to support this research. It is the largest video multi-modal sentiment dataset in terms of scale and video duration to our knowledge, containing $107,267$ comments and $8,210$ micro videos with a video duration of 68.83 hours. To infer the induced sentiment of comment should leverage the video content, we propose the **V**ideo **C**ontent-aware **C**omment **S**entiment **A**nalysis (**VC-CSA**) method as a baseline to address the challenges inherent in this new task. Extensive experiments demonstrate that our method is showing significant improvements over other established baselines. We make the dataset and source code publicly available at `https://github.com/IEIT-AGI/MSA-CRVI`.

1 Introduction

Video multi-modal sentiment analysis, a captivating and challenging research field, has exhibited rapid advancements with a variety of benchmarks proposed in recent years [31, 4, 21, 51, 8, 34]. These benchmarks aim to understand the opinions or emotions of speakers in monologues or dialogues, as depicted in Fig. 1a. They consider the combined input from visual, audio, and subtitle text at the utterance level, which maintain the same semantic(*"It is absolutely wonderful, I fell in love with it from the very first time I saw it and used it."*). Current methods infer the speaker's opinions or emotions by examining elements such as sequential images (e.g., facial expressions, smiles, gazes), audio cues (e.g., tones, pauses, pitch), and transcribed text from spoken words [13].

Nevertheless, current research has primarily centered on the sentiments of the people in the video, paying less attention to viewers' induced sentiment while watching the video. People create and upload micro videos, and viewers contribute comments as responses to the micro video [5]. These comments often reveal sentiments which are induced by the video. Analyzing the viewers' induced sentiments of the video based on comments is significant for developing comprehensive applications

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.


Input video
(monologues/dialogues)
Facial images
Audio
Subtitle
It is absolutely wonderful, I fell in love with it from the very first time I saw it and used it.
Output
positive
joy
Input video
(unlimited)
Video frames
The video demonstrates that an Android phone charges significantly faster than an iOS phone.
Audio(OPTIONAL)
Subtitle
None
Comment Examples
1. it is time to change my phone
2. The app ecosystem of iOS is still better than Android!
Output
positive
anticipation
negative
disgust

(a) Existing Multi-modal Sentiment Analysis

(b) Multi-modal Sentiment Analysis in Comment of Video-Induced (Ours)

Figure 1: Figure (a) describes the setting of traditional multi-modal sentiment analysis, which aims to determine the speaker's sentiment based on the given multi-modal information. Figure (b) illustrates the example of our proposed task. Two comments are highlighted in the figure and hold different induced sentiments toward the related video. For easy comprehension, a description of the video content is presented in a gray box. This description does not serve as input.

such as the analysis of public social sentiment, evaluating advertising effectiveness [47, 37, 24, 35, 25, 14, 46]. We consider the video-induced sentiment analysis as a new paradigm of multi-modal sentiment analysis. In contrast to the existing multi-modal sentiment analysis, a single micro video may yield multitude comments, and each comment may express different sentiments about the video content. As in the example in Fig. 1b, the video shows that an android phone charges faster than an iOS phone, with the aim of illustrating the advantage of android phone. Comment 1 expresses a willingness to switch phone, thus agree with the related video. But the comment 2 offers disagreement by praising the better app ecosystem of iOS. Obviously, just relying on text to infer comments' opinions and emotions toward the video is often inaccurate as both comments exhibit preferences for different smartphones. For precise sentiment inference from comments, it is essential to integrate the video content. Current approaches tend to treat comment sentiment analysis as a simple NLP task and neglect the semantic connection between videos and comments [48, 44, 33, 10, 2, 27].

Taking into account this, we introduce a new task termed **M**ulti-modal **S**entiment **A**nalysis for **C**omment **R**esponse of **V**ideo **I**nduced(**MSA-CRVI**). This task focuses on understanding the induced sentiment of the video, as conveyed through viewers' comments. MSA-CRVI incorporates both the textual comment and the associated video as inputs. Unlike existing video multi-modal sentiment analysis, the MSA-CRVI task presents unique challenges within this innovative paradigm. Firstly, it is challenging to ground the associated video content with each comment. A single video yields a multitude comments that emphasize diverse aspects, necessitating grounding the relevant video contents for each comment's sentiment analysis. Since comments are responses to the video rather than mere textual descriptions, it becomes challenging to directly grounding the video content with the comment. Secondly, it is challenging to model the correlation between comments and their corresponding micro videos due to its temporal complexity. Comments could focus on different temporal-granularity content within the video. Meanwhile, comment may be grounding multiple segment across various video timeline. This implies the necessity for carefully encoding video temporal features and precisely processing the grounding video information.

We have developed a dataset to support the MSA-CRVI task, called **C**omment **S**entiment toward **M**icro **V**ideo (**CSMV**), collected from TikTok, a popular micro video social media platform. CSMV comprises micro videos and associated comments, each of which is annotated for opinions and emotions. Furthermore, we propose a strong baseline method, named **V**ideo **C**ontent-aware **C**omment **S**entiment **A**nalysis (**VC-CSA**) to address these challenges by designing three key modules: Multi-scale Temporal Representation, Consensus Semantic Learning and Golden Feature Grounding. Comprehensive experiments have validated that our method significantly outperforms established baselines. The data and source code are released at `https://github.com/IEIT-AGI/MSA-CRVI`.

Our main contributions including (1) We introduce the **MSA-CRVI** task with a novel setting in multi-modal sentiment analysis. This task involves inferring the induced sentiment according to the comments toward micro-video. (2) To support this task, we have created a dataset named

CSMV, comprising manually annotated opinions/emotions on comments and related videos. To our knowledge, CSMV is the largest dataset of its kind in terms of scale and video duration. (3) As an initial exploration of the task, we present the **VC-CSA** method, which focus on understanding the correlation between comments and micro-videos to infer the opinions and emotions induced by the videos. (4) The extensive experiments demonstrated that VC-CSA outperforms other state-of-the-art multi-modal sentiment analysis methods on the CSMV dataset. We also highlight the critical role of video in the MSA-CRVI task.

2 Related work

Multi-modal sentiment analysis has gained substantial attention and driven the rapid expansion of multi-modal applications. Over the years, a variety of multi-modal datasets have emerged, typically categorized based on the presentation of videos. One category comprises datasets featuring monologue-style videos, such as MOUD [31], OMG-Emotion [4], CH-SIMS [49, 21], CMU-MOSEI [51] and so on. Another category encompasses dialogue-style video datasets such as IEMOCAP [8], MELD [34], often derived from movies and TV shows. Existing multi-modal sentiment analysis datasets impose stringent presentation constraints, but micro-videos are much more diverse in content and format than monologue and dialogue.

Many approaches have been proposed for the multi-modal sentiment analysis. Md Shad Akhtar et al. [1] introduced a context-level inter-modal attention framework aiming to infer the sentiment expressed by the speaker utterance. Delbrouck et al. [11] proposed a transformer-based joint-encoding method employing cross-modal attention mechanisms to capture inter-modality interactions. Their experimental verification revealed the limited role of the visual modality in reasoning within existing multi-modal sentiment analysis. Subsequent studies, including MMIM [15], Self-MM [50], and MISA [16], further support this perspective. Obviously, researchers focus on fusing signals from various modalities to extract complementary information that shares the same semantic meaning. However, these relevant approaches are confined by the current setting which is same as the constraints of the benchmark. Unlike previous research, the VI-MSA task in this paper provides a more complex analysis situation between the video and the textual comment. Additionally, the video themes exhibit greater diversity and free style.

Induced emotion analysis, distinct from perceiving emotion conveyed by content creators, pertains to analyzing emotional reactions induced from content consumers [18, 42]. Presently, there is a growing interest in comprehending the patterns of emotion induced by video [6], since it has a wide range of applications in various perspectives [41, 38, 30, 3]. Currently, researchers mainly focus induced emotion of viewer responses to movies (e.g., DEAP [19], COGNIMUSE [52], LIRIS-ACCEDE [6]). They capture viewers' physiological features to analyze the induced emotion, like EEG and facial videos [19, 36]. Typically, these datasets offer continuous numerical labels for arousal and valence. Due to these characteristics, these datasets are expensive in construction cost and severely restricted for application. In comparison, micro videos are largely created in a freestyle, and the related comments are often easy to obtain and directly reflect induced sentiment.

Several attempts have been made to infer the induced sentiment of video. Benini et al. [7] argued that similar connotations in movie scenes could evoke identical emotional responses. They proposed a method to develop a construct for affective description for movies based on their connotative properties. Tian et al. [42] emphasized the difference between perceived and induced sentiment in the viewer. They employed an LSTM-based model to recognize induced sentiment from the viewers' physiological features, showcasing the effect of integrating multiple modalities, including external information like affective cues in movies. Muszyński et al. [28] further investigated the correlation between dialogue and aesthetic features in inducing sentiment in movies. They introduced an innovative multi-modal model for predicting induced sentiment. Liu et al. [23] employed EEG signals to real-time infer induced sentiment in audiences while watching movies. Their study centered on the widely used LIRIS-ACCEDE database [6] in recent research on induced sentiment in movies. These studies aim to advance the movie art research and support filmmakers in creating emotionally engaging content, where features beyond video are employed to infer induced sentiment. These methodologies mainly rely on the viewers' physiological features, which significantly differ from the video-induced multi-modal sentiment analysis task that utilizes textual comments and videos.

3 Dataset

3.1 Data collection

TikTok is one of the most popular micro video social media platforms and has already attracted significant attention from numerous researchers [17, 26, 20, 40]. Users spontaneously create micro videos and contribute related comments as responses on TikTok. These videos encompass diverse topics (e.g., sports, politics, technology), reflecting human experiences, thereby providing a substantial amount of valuable data for our proposed task MSA-CRVI. The metadata of micro videos on TikTok includes hashtags denoting video topics and the number of likes on comments, facilitating raw data processing [5].

We employ hashtags to collect raw data from Tiktok. Hashtags are formed spontaneously by users creating micro videos, reflect current trends on social media platforms. To enhance data diversity, we set many hashtags with different topics and no restrictions on micro video representation format. A set of hashtags encompassing diverse topics like policy, business, sports, and technology is manually selected. For ensuring the quality of micro videos and comments, micro videos with less than $1,000$ comments are excluded. Then, we sort comments for each micro video based on the number of likes and select the top 20 English comments for annotation. Furthermore, a series of pre-processing steps is undertaken to prevent personal information leakage. Initially, we delete the metadata about the creators of micro videos and comments. Subsequently, any personal information within textual comments (e.g., usernames, emails, phone numbers) is removed. Lastly, instead of raw video data, micro video features generated via the pre-trained visual model are published, including I3D [9], R(2+1)D [43] and VideoMAEv2 [45]. We also provide the webpage URLS of the micro video, allowing other researchers to access the original content through web links. The same features will serve to evaluate our proposed method.

3.2 Data annotation

Table 1: The annotation guidelines for labeling comments on micro videos.

Task	Label	Description
Opinion	positive	Hold a positive attitude towards the content of the video, agree with the information presented in the video, consider the video to be accurate, and experience a sense of comfort induced by the video.
	negative	Hold a negative attitude towards the content of the video, disagree with the information presented in the video, consider there to be errors in the video, and feel uncomfortable because of the video.
	neutral	Hold no clear bias towards the content of the video; provide objective statements without any particular leaning; make comments that are associations triggered by the video rather than expressing a specific attitude; make comments that are not directly related to the content of the video.
Emotion	fear	Fear, terror, apprehension evoked by the video, including reactions of being startled by watching the video, etc.
	disgust	Disgust, dislike, boredom for video content, uninterested in video.
	anger	Rage, anger, annoyance cause by the video.
	sadness	Feel sadness, grief within the video. Catch pensiveness in video.
	joy	Feel happy, joyful, or serenity in heart because of video, including teasing and laughing at the content of the video
	trust	Trust, or feel admiration, or express a convinced attitude towards the content of the video.
	anticipation	Looking forward to, sparking curiosity about, or expressing anticipation cause of the video.
	surprise	The content of the video is surprising, amazed, or shocked more than expected.

We employed 30 human annotators to manually label comments, defining two different types of labels: opinion and emotion, which were derived from previous multimodal sentiment analysis studies [51, 34]. The opinion label indicates the user's attitude towards the micro video in comment. This can encompass agreement with or expression of feelings towards the video, ranging from positive, negative, to neutral. Specifically, the neutral label signifies an absence of clear opinion or views unrelated to the video. The emotion label illustrates the emotional reaction in a comment evoked by the micro video. We employ the Plutchik wheel [32] to define eight categories: joy, disgust, surprise, sadness, trust, fear, anger, and anticipation. These categories encompass a wide range of emotional directions, each illuminated into three levels from mild to intense, effectively capturing human emotional expressions.

Table 2: The statistical information of datasets include video induced emotion, multi-modal sentiment analysis, and the proposed dataset CSMA, from top to bottom.

Dataset	Scale	Video Duration	Video Representation
DEAP [19]	120	2 hours	music
COGNIMUSE [52]	50	3.5 hours	movie
LIRIS-ACCEDE [6]	$9,800$	27 hours	movie
MOUD [31]	400	1 hour	monologue
OMG-Emotion [4]	$2,400$	1 hour	monologue
CH-SIMS2.0 [21]	$4,402$	4.43 hours	monologue
CMU-MOSEI [51]	$23,453$	65.9 hours	monologue
IEMOCAP [8]	$7,433$	12 hours	dialogue
MELD [34]	$13,000$	13 hours	dialogue
CSMV	**107,267**	**68.83 hours**	**unlimited**

We devised a data annotation workflow to ensure annotator quality and reduce individual subjective biases in the annotations. Detailed guidelines and processes for annotating our CSMV dataset are outlined below.

Annotation guidelines. Initially, we establish comprehensive annotation guidelines that precisely define the criteria for data labeling (referenced in Tab. 1). These guidelines provide explicit instructions for identifying and annotating various elements within the dataset.

Pre-annotation phase. Prior to formal annotation, we create a small dataset with ground truth to select annotators. We request them to undergo three rounds of annotation. After each round, we review the results across all annotators. In the final round, annotators with an annotation accuracy that exceeds 90% are chosen for the formal annotation phase.

Formal annotation phase. The raw data comprises both comments and related micro videos. We allocate the raw data to each annotator on a hashtag level. Annotators are tasked with simultaneously labeling opinion and emotion for each comment. Besides, we emphasize the comment that is difficult to understand should be skipped. Throughout this phase, we maintain communication and offer support to annotators. Meanwhile, regular meetings and feedback sessions facilitate continuous improvement and maintain high-quality annotations.

Cross-validation. Assessing opinions and emotions often involves subjective judgment. To ensure annotation consistency and minimize bias, we implement a three-fold cross-validation among annotators. We assign annotation tasks to individual annotators based on hashtags, and each annotator is independently responsible for a specific hashtag. After the initial annotation, we randomly sample 20% of the labeled data from each hashtag and exchange them among two other annotators for correction. These annotators assessed whether they agreed with the annotations. Subsequently, the consistency is calculated based on the validation outcomes. If the consistency rate of the label is less than 90% of the sampled data, the original annotator is required to review the entire dataset for that hashtag. This process would be repeated until a 90% consistency rate was achieved for the entire data set. Our final annotation consistency rate is 94.89%, indicating high-quality annotations. The cross-validation process ensures the consistency of the label. It also provides a reference for human performance for this task.

Finally, we construct **CSMV** dataset comprising $107,267$ comments and $8,210$ micro videos collected from 35 hashtags, totaling a video duration of 68.83 hours.

3.3 Comparison of dataset statistics

Tab. 2 presents a comparison between CSMV and current multi-modal sentiment analysis datasets. Provides details on the scale, duration, and content of the video in each data set. In terms of scale, CSMV stands out with a substantial sample count of 107,267. Meanwhile, the video duration of CSMV is 68.83 hours, offering notably extensive video content. This indicates that CSMV provide the a relatively large scale both in scale and video duration. Furthermore, a key distinction lies in that the video representation in CSMV is unlimited. Comparatively, the existing video induced emotion focus on the movie, and the existing multi-modal dataset focus on the sentiment of speaker in the video. These datasets have limitations in conveying visual information and expression. Conversely, our proposed CSMV comprises a broader and more diverse range of video representation, potentially introducing additional complexities and challenges in sentiment analysis. More statistics pertaining to our CSMV dataset are available in the supplementary materials.

3.4 Ethics

Our research is conducted entirely for academic purposes, in line with TikTok's privacy policy, which permits independent research under specific criteria. Numerous researchers have used TikTok's public data, releasing related datasets under the same principles [17, 26, 20, 40]. Concerning personal privacy, our CSMV dataset would not publish the original videos. Instead, it publishes only the visual features extracted from micro videos using the pre-trained video models including I3D [9], R(2+1)D [43] and VideoMAEv2 [45]. Additionally, the comments solely preserve the text, removing all user-related information. Meanwhile, we provide the URLS of the micro video webpage, allowing other researchers to access the original content through web links. Both the code and data are publicly accessible under the CC BY-NC-SA 4.0 license, intended for academic and non-commercial use.

4 Method

To infer the comment's induced sentiment toward to related micro video, we propose a novel method called **V**ideo **C**ontent-aware **C**omment **S**entiment **A**nalysis (**VC-CSA**). It takes a comment and the related micro video as input to infer the opinion and emotion toward to the video which expressed through the comment. Fig. 2 is the architecture of the framework. The feature encoder is a video pre-trained model(e.g., I3D [9]), which encodes the micro video into a set of vector representations as original temporal visual features input. The comment text is encoded by a RoBERTa [22] language pre-trained model to extract text features from the comment. Our proposed method consists of three principal modules: Multi-scale Temporal Representation, Consensus Semantic Learning, and Golden Feature Grounding. We integrate the multi-scale video golden feature with the textual comment with a fusion module and utilize a $Softmax$ classifier to infer opinions and emotions. For training, we apply cross-entropy loss to each classification head and aggregate these losses for optimization.

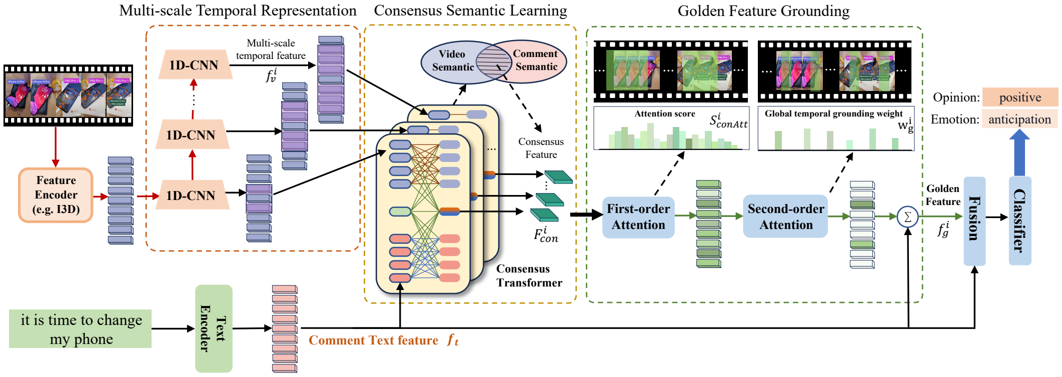


Figure 2: The architecture of **V**ideo **C**ontent-aware **C**omment **S**entiment **A**nalysis (**VC-CSA**). We mainly design Multi-scale Temporal Representation, Consensus Semantic Learning and Golden Feature Grounding modules to address the new challenges of the proposed task.

4.1 Multi-scale Temporal Representation

Each video has the potential to provoke a multitude of comments from viewers. These comments may address specific segments or the entirety of the video story. For instance, in a video where a dog chases a cat and ultimately collides with a door, one comment, '*That hurts*,' refers to the end of video. In contrast, another comment, '*Dogs do not like cats*,' reflect the general theme of the video. Consequently, it is crucial to encode the semantic features of the video across various temporal scales to facilitate the correlation between the video and comments spanning different different time ranges.

The **Multi-scale Temporal Representation** module is design to capture the visual features from the video in various temporal scales. This is accomplished by stacking multiple 1D Convolutional Neural

Networks (CNNs) and employing the $ReLU$ activation function between layers. Each 1D-CNN operates with a kernel size of three and a stride of one, traversing the temporal dimension of the video's visual features input. As the number of layers increases, the network progressively expands broader temporal contexts within the video's visual features. Consequently, we obtain hierarchy of multi-scale temporal representations, denoted as $\{f_v^i \in \mathbb{R}^{v_l \times d_v}\}$ of the video visual information, capturing a spectrum from finer to coarser granularities, where v_l represents the length of the video, d_v denotes the dimension of the visual representations, and i indexes the layers.. This approach is instrumental in analyzing the video's contextual representation difference over various time spans.

4.2 Consensus Semantic Learning

The comments, being responses to a video, do not describe the video content directly, creating a semantic gap between the video and the comments. This distinction hinder directly grounding relevant video content based on the comment text. To address this challenge, we introduce **Consensus Semantic Learning**, a module designed to deeply model semantic correlation between the video and the comments, facilitating more effective video content grounding.

The foundation for effectively grounding video content lies in building a well-defined query to bridge the semantic gap between the comments and videos. Therefore, we introduce a video-comment consensus transformer to capture the shared semantic occurrences between comment and video. As illustrated in Fig. 2, this transformer is a special structure derived from transformer encoder block. It processes video feature f_v^i, text feature f_t, and several trainable Consensus Tokens as input. Consensus token representations $f_{con} \in \mathbb{R}^{v_l \times d^T}$, are randomly initialized and become trainable in training phase, d^T denotes the dimension of the consensus transformer. With in the construction, the attention connection between the video and text features are masked, allowing for information exchange solely through the Consensus Token. Consequently, the consensus tokens serve as mediums, sharing the semantic between the video and the comment. We take the consensus tokens representation output F_{con}^i in Eq. 1 as a consensus feature to reduce the semantic gap like a bridge.

$$F_{con}^i = ConsensusTansformer_{cons}([f_v^i; f_{con}; f_t]) \tag{1}$$

In Eq. 1, $f_t \in \mathbb{R}^{l_t \times d_t}$ denotes the comment text feature, where l_t is the length of text, d_t is the dimension. Importantly, the consensus transformer block is parameter-independent for each temporal scale of $\{f_v^i\}$. This methodology is applied across all multi-scale temporal features.

4.3 Golden Feature Grounding

To accurately interpret the sentiment of comment related to a video, it is important to ground the video content referenced by the comment. Given the temporal nature of video, continuous frames often exhibit high similarity, leading to redundant information during the grounding phase. To control this, We design **Golden Feature Grounding** module, which comprises a two-steps approach to compute grounding weight. In the first-order grounding, we employ a multi-head attention mechanism. This mechanism utilize the consensus token representation F_{con}^i as the query, the video visual feature f_v^i as the key and the value. The process of calculation is illustrated as Eq. 2:

$$S_{conAtt}^i = MultiHeadAttention(F_{con}, f_v^i, f_v^i) \tag{2}$$

Here, the attention score $S_{conAtt}^i \in \mathbb{R}^{head \times v_l}$ reflects the comment attention across the video temporal, with $head$ representing the number of attention heads.

Viewers' attention to video content varies over time, often focusing on multiple segments simultaneously. This attention score S_{conAtt}^i may exhibit smoothness due to the similarity for adjacent temporal segments, resulting from the the temporal nature of video. Consequently, shorter segments could be overshadowed by longer ones if the attention score S_{conAtt}^i is use directly to obtain video information relevant to the comment. Address this, we design a second-order grounding to filter out the redundancies and obtain the golden feature that represents the essence of video relevant to the comment. We take the multi-head attention score S_{conAtt}^i as an indicator of temporal attention trends across distinct vector space. This score is input to a memory module to analyze the trend of attention scores in the temporal direction. The memory module is same as the the cell state of LSTM. Then, the representation is processed through a $ReLU$ function to obtain the global temporal grounding weight $W_g^i \in \mathbb{R}^{v_l}$.

$$W_g^i = ReLU(cellState(LSTM(S_{conAtt}^i)) \tag{3}$$

This grounding weight is then multiply to the video features f_v^i to produce the features $f_g^i \in \mathbb{R}^{d_v}$ along the temporal axis, which we regard as the golden features related to the comment. The calculation process is shown as Eq. 4:

$$f_g^i = \sum W_g^i f_v^i \tag{4}$$

4.4 Fusion and Classifier

After the steps outlined above, we introduce a fusion module designed to integrate video features across various temporal scales into the comment feature, thereby enriching the interaction between comment text and video data. To facilitate this, we employ a multi-view attention mechanism, wherein the comment text token feature denoted as f_t^j as the query, and the video golden features in multi-scales, represent by $\{f_g^i\}$ as key and value. This approach specifically targets capturing the interactions at the token level between the comment and the video, where j corresponds to the index of the token within the comment.

$$AttnScale_j^i = Attention(f_t^j, \{f_g^i\}) \tag{5}$$

$$F_g^j = \sum_i AttnScale_j^i f_g^i \tag{6}$$

Subsequently, we concatenate the features $\{F_g^j\} \in \mathbb{R}^{l_t \times d_v}$ and text features f_t in token-level along the feature dimension to generated the video context-aware comment semantic feature $f_s \in \mathbb{R}^{l_t \times (d_v + d_t)}$. This feature f_s is processed through a layer of multi-head self-attention and a pooling mechanism to get the final context-aware comment semantic feature representation F_s for sentiment analysis.

$$F_s = MaxPool(MultiHeadSelfAttention([\{F_g^j\}; f_t])) \tag{7}$$

This fusion strategy aims to incorporate both video and textual information into a unified representation. We utilize two $softmax$ functions on F_s to calculate the possibility of opinion and emotion which the comment response to video.

5 Experiments

We select representative sentiment analysis methods for comparison. Notably, our selection included methods that primarily utilize textual input, such as BERT [12] and RoBERTa [22]. We exclusively trained these models on the comment text from the CSMV dataset, facilitating an evaluation of the micro videos' impact on the MSA-CRVI task. Furthermore, we select several typical traditional multi-modal sentiment analysis methods: TBJE [11], SELF-MM [50], MISA [16], MMIM [15] and CubeMLP [39]. Our method and comparative methods are implemented on the PyTorch platform [29] and trained on 4 Nvidia Tesla V100 GPUs. We use I3D [9], R(2+1)D [43] and VideoMAEv2 [45] as encoder features of video.

For the implementation of our proposed model, we set the hidden dimensions d_v, d^T and d_t to 768. To ensure equitable comparisons, we align the training settings (e.g., loss function, batch size, learning rate strategy, etc) with all methods. To evaluate the performance of the models, we randomly split our dataset into training, development (dev), and testing sets using a ratio of 7:1:2. The dev set serves as the basis for selecting the most effective model for each method based on performance outcomes. We follow prevailing evaluation protocols to use F1-score as the primary metrics to measure the performance. Additionally, we calculate mean values from 5 random seeds for each performance metric.

5.1 Comparison Analysis

The performance metrics of each method is presented in Tab. 3 individually. It is evident that the **VC-CSA** achieve the highest F1 scores for opinion recognition and emotion recognition on all video feature encoders. It exhibits significant advantages over existing multi-modal methods in our proposed task, indicating the limitation of current approaches in addressing the distinctive challenges presented by our research. Meanwhile, the results clearly demonstrate that multi-modal approaches outperform those depending solely on text, underscoring the importance of video content in interpreting sentiments of comments.

Table 3: The experiment results of the comparison.

Models	Opinion				Emotion			
	Micro	Macro			Micro	Macro		
	F1-score	F1-score	Recall	Precision	F1-score	F1-score	Recall	Precision
BERT [12](only text)	56.42	48.52	48.14	49.31	43.34	33.64	32.98	34.59
RoBERTa [22](only text)	56.95	49.29	48.87	49.98	47.27	37.56	36.85	38.77
TBJE [11](I3D)	65.81	59.80	59.20	60.94	55.67	48.14	48.71	46.61
SELF-MM [50](I3D)	65.77	58.56	57.30	61.20	53.92	46.44	44.64	49.87
MISA [16](I3D)	72.41	66.54	65.40	68.69	57.42	49.71	48.07	52.77
MMIM [15](I3D)	65.40	58.39	59.96	57.65	52.35	43.65	42.37	45.86
CubeMLP [39](I3D)	65.60	61.51	60.82	61.16	51.87	47.31	45.07	46.16
SELF-MM [50](R(2+1)D)	64.65	58.74	57.39	60.18	53.89	42.85	42.17	43.49
MISA [16](R(2+1)D)	70.65	66.53	65.55	67.50	57.42	48.48	47.94	49.01
SELF-MM [50](VideoMAEv2)	67.18	61.47	63.10	59.96	53.57	45.41	44.66	46.16
MISA [16](VideoMAEv2)	73.00	67.07	64.58	69.75	59.69	48.72	49.50	47.39
VC-CSA(I3D)	**73.52**	**67.51**	**66.51**	**69.19**	**62.99**	**55.18**	**54.47**	**56.36**
VC-CSA(R(2+1)D)	**72.34**	**65.15**	**64.89**	**65.42**	**58.46**	**54.24**	**54.05**	**54.42**
VC-CSA(VideoMAEv2)	**74.56**	**68.90**	**67.60**	**70.25**	**63.67**	**56.18**	**55.93**	**56.42**

5.2 Ablation Study

We execute ablation studies on the three principal modules to validate the effectiveness. We adopted standard strategy instead of our custom design to assess performance difference. For the Multi-scale Temporal Representation, we use the **only single layer** and the **only last layer** CNN representation as the video feature at a single scale instead of it, respectively. For the Consensus Semantic Learning, we replace our consensus token with the **last token in original transformer** (**LT** for short) encoder block. This involved concatenating the video and comment text and using the feature from the last position as the attention query. For Golden Feature Grounding, we directly use the **attention score** S_{conAtt} (**AttnS** for short) as the grounding weight to obtain the video visual feature in relation to the comment. We conducted ablation studies using various combination methods. All experiments is based on the I3D feature encoder. The findings from the ablation studies are presented in Tab. 4. It is evident that excluding these designs from **VC-CSA** results in a decrease in evaluation metrics, highlighting their critical contribution to the method. A comparison between the ablated models and our complete model reveals an approximate $1 - 4\%$ improvement in the Micro F1 score.

Table 4: The Ablation study on our method. The Ablation Setting column is the alternative designs.

Ablation Setting	Opinion Micro F1	Opinion Macro F1	Emotion Micro F1	Emotion Macro F1
-Only single layer	72.35	65.51	62.06	54.18
-Only last layer	69.13	63.37	59.67	51.81
-LT	72.32	66.43	62.52	54.74
-AttnS	71.93	65.23	61.22	52.82
-LT, AttnS	72.11	63.28	60.85	50.07
-Only single layer, AttnS	71.66	64.52	60.96	50.85
-Only single layer, LT	72.15	65.81	61.48	51.58
-Only last layer, AttnS	70.20	63.28	57.08	48.51
-Only last layer, LT	68.90	62.89	57.04	48.80
-Only single layer, LT, AttnS	70.70	62.33	60.25	51.56
-Only last layer, LT, AttnS	68.90	62.38	57.01	48.62
VC-CSA	**73.52**	**67.51**	**62.99**	**55.18**

5.3 Evaluation On YouTube

Our CSMV is sourced from a single platform, which may limit the generalizability of our findings. To address this, we conduct additional experiments using a smaller dataset collected from YouTube, a widely used video platform. We manually annotated a subset of YouTube videos and their corresponding comments, testing our model trained on the CSMV dataset. The YouTube dataset consists of 21 videos and 138 associated comment samples. The evaluation results are shown in Tab. 5. The results indicate that our method performs well on the YouTube data, suggesting that our approach can be generalized to other video platforms. Although the YouTube test set is relatively small due to time and resource constraints, it provides initial evidence of the broader applicability of our work. More experiments and discussions are available in the supplementary materials.

Table 5: Evaluation VC-CSA(I3D) model on a small YouTube dataset.

Ablation Setting	Opinion Micro F1	Opinion Macro F1	Emotion Micro F1	Emotion Macro F1
VC-CSA(I3D)	**71.73**	**70.67**	**61.59**	**58.89**

6 Conclusion

In conclusion, this study introduces the task of multi-modal sentiment analysis in comment of video-induced (MSA-CRVI), focusing on understanding sentiment from comments related to micro-video content. To support in this task, we have developed CSMV dataset, consisting of micro videos and their annotated comments. The proposed VC-CSA method effectively infers sentiments from comments within the context of corresponding video, making a significant contribution for the novel multi-modal sentiment analysis setting. Our work still has limitations. In particular, our current dataset and baseline do not include audio features. However, incorporating audio features could enhance the understanding of sentiment in our task context. Looking forward, we aim to enlarge the dataset to expand the diversity. And we plan to release the corresponding audio features of the videos and more visual feature to further enhance the dataset's utility.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Shandong Provincial Natural Science Foundation under Grants ZR2023LZH003.

References

[1] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*, 2019.

[2] Rawan Fahad Alhujaili and Wael M. S. Yafooz. Sentiment analysis for youtube videos with user comments: Review. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 814–820, 2021.

[3] Ghadah Alqahtani and Abdulrahman Alothaim. Predicting emotions in online social networks: challenges and opportunities. *Multimedia Tools and Applications*, 81(7):9567–9605, 2022.

[4] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[5] Kristen Barta and Nazanin Andalibi. Constructing authenticity on tiktok: Social norms and social support on the "fun" platform. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[6] Yoann Baveye, Jean-Noël Bettinelli, Emmanuel Dellandréa, Liming Chen, and Christel Chamaret. A large video database for computational models of induced emotion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 13–18. IEEE, 2013.

[7] Sergio Benini, Luca Canini, and Riccardo Leonardi. A connotative space for supporting movie affective recommendation. *IEEE Transactions on Multimedia*, 13(6):1356–1370, 2011.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[10] Smitashree Choudhury and John G. Breslin. User sentiment detection: a youtube use case. In *The 21st National Conference on Artificial Intelligence and Cognitive Science*, August 2010.

[11] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[14] Benjamin Guinaudeau, Kevin Munger, and Fabio Votta. Fifteen seconds of fame: Tiktok and the supply side of social video. *Computational Communication Research*, 4(2):463–485, 2022.

[15] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, 2021.

[16] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.

[17] Heekyoung Jung and Qiyang Zhou. Learning and sharing creative skills with short videos: A case study of user behavior in tiktok and bilibili. In *Online*, 09 2019.

[18] Kari Kallinen and Niklas Ravaja. Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2):191–213, 2006.

[19] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[20] Chen Ling, Krishna P. Gummadi, and Savvas Zannettou. quot;learn the facts about covid-19quot;: Analyzing the use of warning labels on tiktok videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):554–565, Jun. 2023.

[21] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module, 2022.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from eeg signals. *IEEE Transactions on Affective Computing*, 9(4):550–562, 2017.

[24] Jessica T. Lovett, Kamran Munawar, Sharon Mohammed, and Vinay Prabhu. Radiology content on tiktok: Current use of a novel video-based social media platform and opportunities for radiology. *Current Problems in Diagnostic Radiology*, 50(2):126–131, 2021.

[25] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, page 257–266, New York, NY, USA, 2020. Association for Computing Machinery.

[26] Sepehr Mousavi, Krishna P. Gummadi, and Savvas Zannettou. Auditing algorithmic explanations of social media feeds: A case study of tiktok video explanations. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):1110–1122, May 2024.

[27] Abbi Nizar Muhammad, Saiful Bukhori, and Priza Pandunata. Sentiment analysis of positive and negative of youtube comments using naïve bayes – support vector machine (nbsvm) classifier. *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, pages 199–205, 2019.

[28] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing*, 12(1):36–52, 2019.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[30] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 860–868, 2015.

[31] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-level multimodal sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics*, 2013.

[32] ROBERT PLUTCHIK. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980.

[33] Rhitabrat Pokharel and Dixit Bhatta. Classifying youtube comments based on sentiment and type of sentence. *ArXiv*, abs/2111.01908, 2021.

[34] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.

[35] Zea Qiyang and Heekyoung Jung. Learning and sharing creative skills with short videos: A case study of user behavior in tiktok and bilibili. In *Int. Assoc. Soc. Des. Res. Conf*, number 10, pages 25–50, 2019.

[36] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[37] Lauren Southwick, Sharath C. Guntuku, Elissa V. Klinger, Emily Seltzer, Haley J. McCalpin, and Raina M. Merchant. Characterizing covid-19 content posted to tiktok: Public sentiment and response during the first phase of the covid-19 pandemic. *Journal of Adolescent Health*, 69(2):234–241, 2021.

[38] Prakash Chandra Sukhwal and Atreyi Kankanhalli. Determining containment policy impacts on public sentiment during the pandemic using social media data. *Proceedings of the National Academy of Sciences*, 119(19):e2117292119, 2022.

[39] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3722–3729, New York, NY, USA, 2022. Association for Computing Machinery.

[40] Saimourya Surabhi, Bhavik Shah, Peter Washington, Onur Cezmi Mutlu, Emilie Leblanc, Prathamesh Mohite, Arman Husic, Aaron Kline, Kaitlyn Dunlap, Maya McNealis, Bennett Liu, Nick Deveaux, Essam Sleiman, and Dennis P. Wall. Tiktok for good: Creating a diverse emotion expression database. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2495–2505, 2022.

[41] Thales Teixeira, Michel Wedel, and Rik Pieters. Emotion-induced engagement in internet video advertisements. *Journal of marketing research*, 49(2):144–159, 2012.

[42] Leimin Tian, Michal Muszynski, Catherine Lai, Johanna D Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same? In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 28–35. IEEE, 2017.

[43] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. Sentube: A corpus for sentiment analysis on youtube social media. In *International Conference on Language Resources and Evaluation*, 2014.

[45] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, June 2023.

[46] Gabriel Weimann and Natalie Masri. Research note: Spreading hate on tiktok. *Studies in Conflict & Terrorism*, 46(5):752–765, 2023.

[47] Alex J. Xu, Jacob Taylor, Tian Gao, Rada Mihalcea, Verónica Pérez-Rosas, and Stacy Loeb. Tiktok and prostate cancer: misinformation and quality of information using validated questionnaires. *BJU International*, 128, 2021.

[48] Douiji yasmina, Mousannif Hajar, and Al Moatassime Hassan. Using youtube comments for text-based emotion recognition. *Procedia Computer Science*, 83:292–299, 2016. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.

[49] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online, July 2020. Association for Computational Linguistics.

[50] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, 2021.

[51] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[52] Athanasia Zlatintsi, Petros Koutras, Georgios Evangelopoulos, Nikos Malandrakis, Niki Efthymiou, Katerina Pastra, Alexandros Potamianos, and Petros Maragos. Cognimuse: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017:1–24, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]

(c) Did you discuss any potential negative societal impacts of your work? [Yes]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]