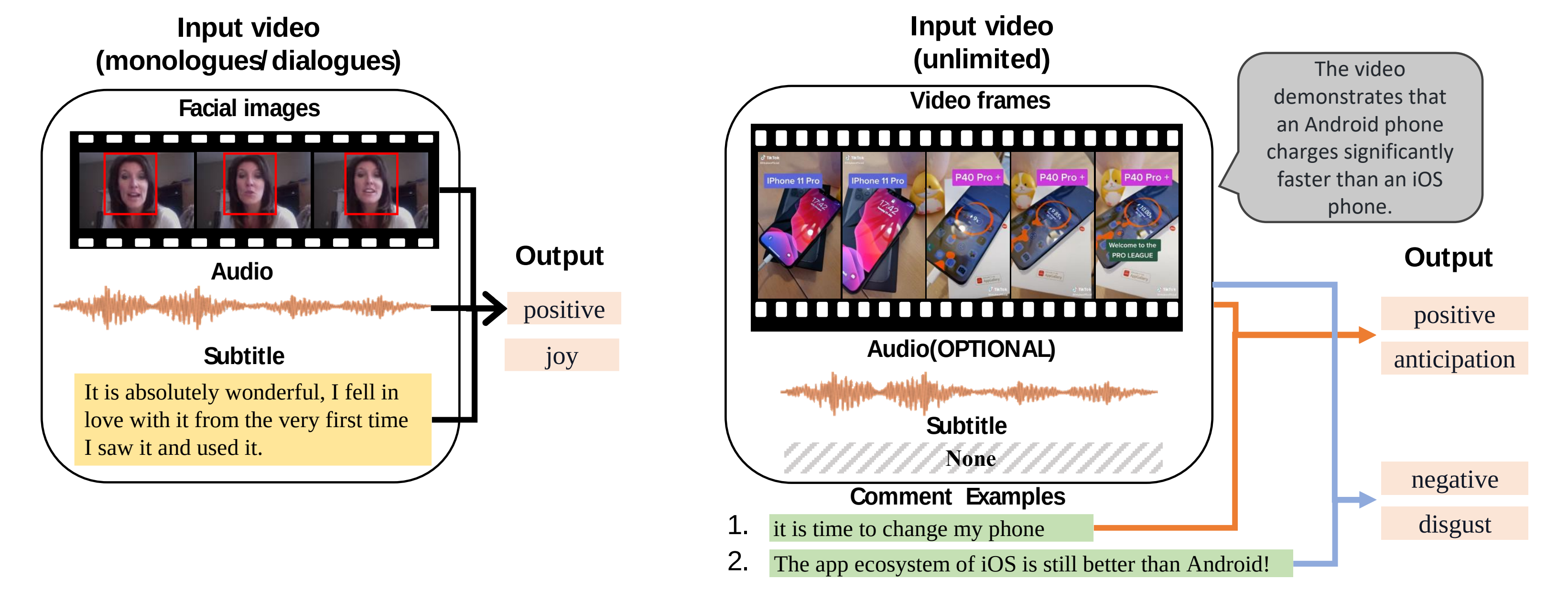
Infer Induced Sentiment of Comment Response to Video: A New Task, Dataset and Baseline

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. [3]School of Computer Science, Fudan University.

1 New Task

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.


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.

2 Dataset

We have developed a dataset to support the **MSA-CRVI** task, called Comment Sentiment toward Micro Video (**CSMV**), collected from TikTok. **CSMV** comprises micro videos and associated comments, which is annotated for opinions and emotions. The opinion indicates the user's attitude towards the micro video in comment. The emotion illustrates the emotional reaction in a comment evoked by the micro video.

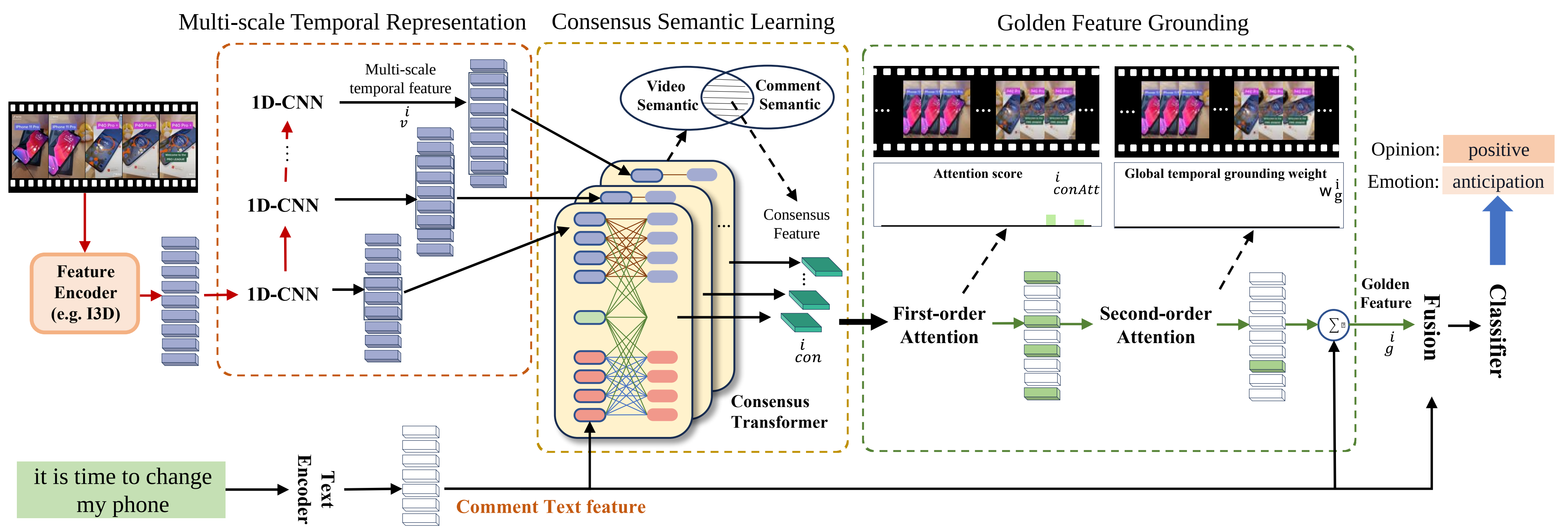
Table 1: The annotation guidelines for labeling comments on micro videos.

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.

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

3 Method

We propose a strong baseline method, named Video Content-aware Comment Sentiment Analysis (**VC-CSA**) to address these challenges by designing three key modules: Multiscale Temporal Representation, Consensus Semantic Learning and Golden Feature Grounding.



4 Experiments

We select representative sentiment analysis methods for comparison, including methods that primarily utilize textual input, such as BERT and RoBERTa, and several typical traditional multi-modal sentiment analysis methods: TBJE, SELF-MM, MISA, MMIM and CubeMLP. We use I3D, R(2+1)D and VideoMAEv2 as encoder features of video.

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**

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.

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**

To address the possible limit for the generalizability of our findings, we conduct additional experiments using a smaller dataset collected from YouTube, a widely used video platform.

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**

Dataset available on https://github.com/IEIT-AGI/MSA-CRVI