Supplementary Material
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]Nankai University, [3]Fudan University
{jiaqi01, fanbaoyu, xucong, liulu06, jinliang, duguoguang}@ieisystem.com,
{guozhenhua zhaoyaqian}@ieisystem.com, xjhuang@fudan.edu.cn, lirg@ieisystem.com

A Dataset Details

This section provides a comprehensive overview of the **CSMV** dataset. The CSMV dataset comprises micro videos and their corresponding comments, which have been updated from February 2020 to October 2022. This extensive time range allows for the inclusion of a diverse set of content, capturing the evolution of sentiments over the course of more than two years. In total, the CSMV dataset comprises 8,210 micro videos, totaling approximately 68.83 hours of video duration, along with 107,267 related comments. The CSMV dataset defines two distinct types of labels, opinion and emotion, for analyzing the sentiment expressed in the comments towards the micro videos. By leveraging the combination of video and textual content in this dataset, researchers can examine the interaction between language expressions and visual cues in sentiment analysis.

To deepen our understanding of the CSMV dataset, we performed an analysis of the distribution of videos and related comments using specific hashtags. As depicted in Fig. 1, this distribution exhibits a rich diversity of topics in video content. This diversity has brought rich expression of sentiment in user comments, giving the CSMV dataset an advantage in comprehending the complexity of induced sentiment. Moreover, this diversity expands the application of the dataset for multimodal sentiment analysis tasks. By leveraging the CSMV dataset, researchers and practitioners can explore the interactive relationships between visual content and textual expressions by analyzing the multi-modal induced sentiment. This comprehensive understanding not only aids in sentiment analysis but also provides valuable insights into the complex interplay of modalities within human communication.

The distribution of labels in our CSMV dataset is shown in Fig. 2. In Fig. 2a, the opinion labels are distributed as follows: positive - 47%, neutral - 42%, and negative - 11%. Negative comments are clearly in the minority. The distribution of emotion labels is depicted in Fig. 2b. According to the statistical results, the top two labels with the highest proportion are joy (32%) and trust (27%). Additionally, the labels of sadness, fear, and anger have the smallest proportions, accounting for 5%, 2%, and 2%, respectively. The distribution indicates an imbalance in labels. TikTok platforms tend to filter out videos with overly negative content. It may lead to the imbalance in label distribution. More to the point, the distribution of opinions aligns with the behavior tendencies of users in comments on micro videos. Users are more inclined to write comments for positive experiences. In contrast, expressing negative opinions about videos they dislike is considered rare, and users tend to ignore such micro videos. Meanwhile, TikTok uploaders are more inclined to create and share content to attract other users to watch. Consequently, the majority of comments have a non-negative sentiment. The distribution of emotion labels follows a similar pattern. On the social media platform, users tend to express positive emotions, such as joy and trust, more frequently than negative ones. Based on

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

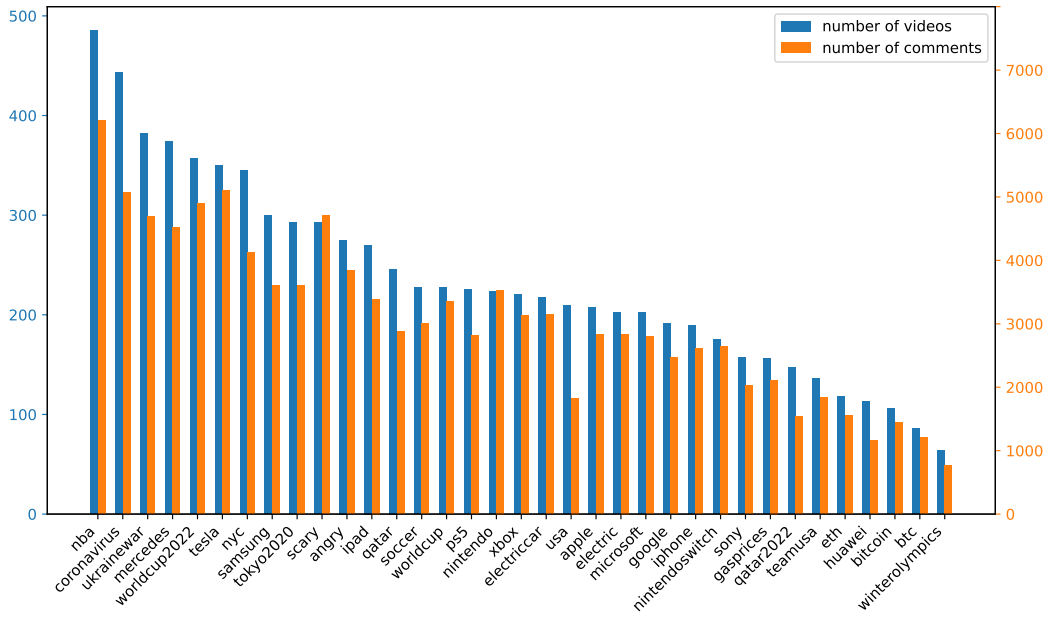
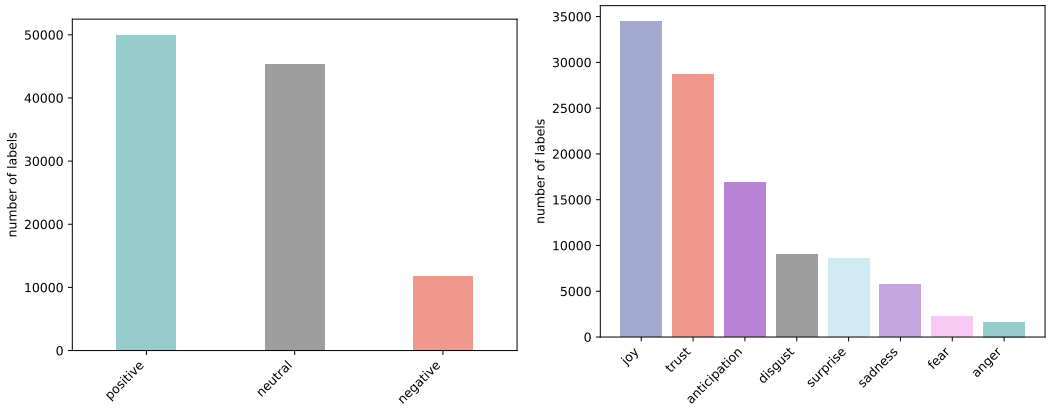


Figure 1: The distribution of the amounts of the micro video and comments under the hashtag.



(a) Opinion label distribution.

(b) Emotion label distribution.

Figure 2: The distribution of the number of labels in our CSMV dataset.

these findings, our dataset accurately represents the real distribution of human sentiment in real-world scenarios.

One primary objective of the CSMV dataset is to enhance the diversity of sentiment responses towards videos. This goal is achieved through the annotation of multiple comments for each video, ensuring a broader range of sentiment expressions within CSMV. An analysis of our dataset has been conducted, and the results are discussed in Fig. 3. The distribution clearly illustrates that the majority of our videos provide more than 10 annotated comments, while only a small proportion have 2-5 annotated comments. This observation signifies that our dataset exhibits a greater level of complexity compared to conventional multi-modal sentiment analysis tasks. By incorporating multiple comments for each video in our annotation process, we enable a more comprehensive and nuanced understanding of human sentiment. This approach facilitates the training of artificial intelligence systems to recognize and respond to diverse human experiences, as our dataset CSMV encompasses a broader spectrum of sentiment responses. The inclusion of multiple comments allows for a deeper exploration of the various opinions and emotions conveyed by individuals in response to videos. Each comment

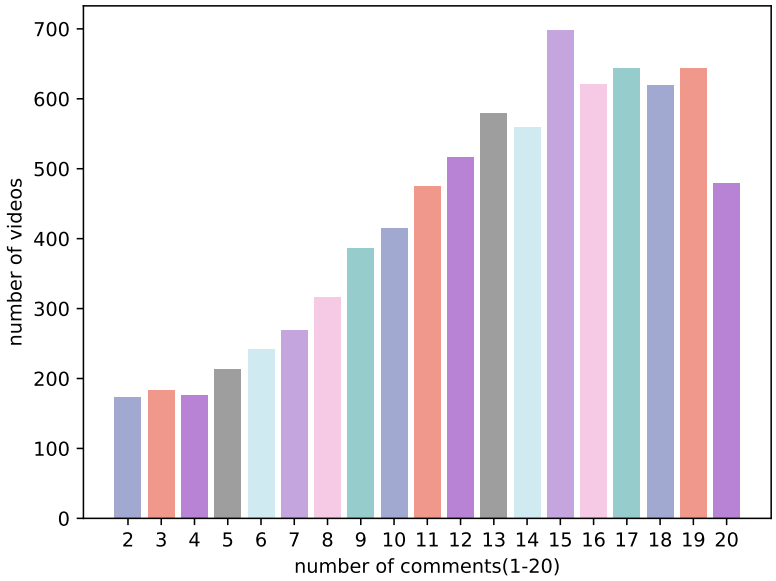


Figure 3: The distribution of the amounts of the comments under single micro video.

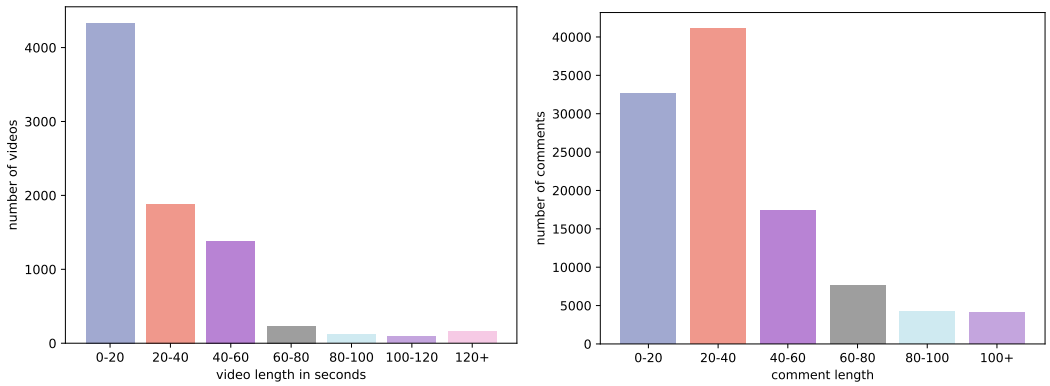
represents a distinct opinion, contributing to the rich complexity of sentiment found within CSMV. In essence, the CSMV dataset's commitment to incorporating a diverse range of sentiment responses strengthens its capacity to train AI systems, which make it capable of accurately identifying and effectively responding to the multitude of sentiment expressed by individuals in relation to videos. This comprehensive collection of sentiment responses enables researchers and developers to develop a more sentimental AI system. It could possess a more profound understanding of human sentiment and own the ability to discern and appropriately react to the intricacies of human experience.

Length information for the samples in the CSMV dataset, is mainly focus on the distribution of video duration and comment text length. The analysis, as depicted in Fig. 4 (a), indicates that the majority of videos have a duration of 60 seconds or less, with the highest proportion falling between 0 and 20 seconds. Micro video creators aim to convey impactful narratives within a limited time frame, allowing the popularity of their content on social media platforms. Likewise, comments in our dataset follow a distribution pattern similar to that of video duration. Fig. 4 (b) illustrates that the majority of comments contain 60 characters or less, with the most frequently observed length ranging from 40 to 60 characters. When users respond to a video by commenting, they may use concise text with video-specific unconventional abbreviations. Without the visual context provided by the video, understanding the intended message can be difficult. In conclusion, analyzing the video duration and comment length in our CSMV dataset emphasizes the prevalence of concise communication and short-form content in the field of social media. Utilizing brevity and unconventional abbreviations, creators and commenters endeavor to captivate audiences and engage in fast-paced online discourse. A comprehensive understanding of the underlying context and meaning of these micro videos and comments requires video context support.

B More Experiments

B.1 Optimization Hyper-parameter

Our **VC-CSA** integrates multiple modules requiring hyper-parameters configuration. To determine the setup, we explore the several key factors through experiments: (1) the layer count in Multi-scale Temporal Representation module; (2) the layer count in Consensus Transformer module corresponding to each scale of multi-scale temporal representations $\{f_v^i\}$, and (3) the count of Consensus Tokens in the Consensus Transformer. Systematic experiments with these parameters as shown in Tab. 1. The findings indicate that the VC-CSA model achieves optimal performance with 4 layer Multi-scale Temporal representation, 1 layer Consensus Transformer, and 1 Consensus Token.



(a) Video length distribution.

(b) Comment length distribution.

Figure 4: The distribution of the length of video duration and comment text.

Hyper-parameter setting	Opinion Micro F1	Opinion Macro F1	Emotion Micro F1	Emotion Macro F1
2,1,1	72.62	66.49	62.10	53.85
3,1,1	71.79	65.78	60.82	52.58
4,1,1	**73.52**	**67.51**	**62.99**	**55.18**
5,1,1	72.67	66.38	62.69	54.90
6,1,1	72.46	66.32	61.93	54.26
4,1,2	70.83	64.11	60.88	52.41
4,1,3	69.42	63.62	59.80	50.66
4,2,1	72.58	65.99	62.52	54.64
4,3,1	72.72	65.81	61.99	53.27

Table 1: The experimental results of diverse hyper-parameter settings for **VC-CSA**. In the Hyper-parameter setting column, the three parameters are the layer count in Multi-scale Temporal Representation module, the layer count in the Consensus Transformer module, and the count of Consensus Tokens.

B.2 Evluation on Large Multi-modal LLMs

Recent advancements in research on multimodal large language models have been remarkable. With the expansion of model parameters and the scale of training data, these models have demonstrated exceptional performance across a wide range of tasks. In light of this progress, we selected several prominent multimodal large models to evaluate on our MSA-CRVI task, including LLaVA (8B) [5], GLM-4V (9B) [7], ShareGPT4Video (8B) [3], and InternVL (40B) [4]. We designed a simple prompt to evaluate all four models on the test set in a zero-shot setting. The performance evaluations for these models are presented in Table 2.

The experimental results demonstrate that, in a zero-shot setting, current multimodal large language models perform poorly on this task, significantly underperforming compared to various multimodal sentiment analysis baselines. This performance gap arises from two primary factors. First, these models have not been trained on the CSMV dataset. Second, the limitations are inherent to the models themselves. One key issue is that multimodal large language models are typically trained on semantically aligned data, where different modalities share consistent meanings. In contrast, our task involves the interaction between two modalities, commentary text and video, where one modality

Models	Opinion Micro F1	Opinion Macro F1	Emotion Micro F1	Emotion Macro F1
LLaVA (8B)	20.74	19.65	7.44	6.49
GLM-4V (9B)	50.98	45.61	33.55	27.16
ShareGPT4Video (8B)	45.98	32.82	16.32	15.12
InternVL-40B (40B)	56.53	52.04	29.02	18.13

Table 2: The experimental results of multimodal large language models on the MSA-CRVI task. These models are evaluated on the test set in a zero-shot setting with a simple prompt.

(text) responds to the other (video). This divergence from the models' training objectives contributes to their suboptimal performance. Additionally, these models process video content through frame sampling, which may result in incomplete semantic representation of the video, leading to interpretive discrepancies. Future work will focus on further evaluating these models on this task, with the goal of providing a more comprehensive analysis of the challenges involved.

C The Importance of Video in MSA-CRVI Task

To demonstrate the significance of video in the MSA-CRVI task, we present the inference outcomes for several samples from the test set, as shown in Fig. 5.

A shared feature among the three samples is that the absence of video information complicates the interpretation of the intended meaning when relying solely on textual comments. This often leads to inaccurate inferences about the opinions and emotions conveyed in the comments related to the video. Even the fine-tuned RoBERTa model on the comment of the CSMV dataset exhibits notable deviations as well. However, our VC-CSA model demonstrates this capability by accurately identifying sentiment within the context of video content, proving the integration of video viewing enables a more accurate understanding of the comments.

Samples Explanation. To provide a comprehensive explanation of our task, we will offer detailed explanations for the four selected examples discussed in the main article, as shown in Fig. 6.

Example 1 in Fig. 6a highlights a common debate in the smartphone world-Android versus iOS. The video showcases a comparison between the charging speed of an Android smartphone and an iOS smartphone, aiming to demonstrate the superiority of Android over iOS. The comment "it is time to change my phone." in response to the video expresses agreement with its content and a desire to acquire a new phone. However, another comment argues that advocating Android's superiority based solely on charging speed is a one-sided argument and points out that the iOS platform has a healthier application ecosystem, stating "The app ecosystem of iOS is much healthier than Android!" This comment conveys a negative opinion towards the video content and a sense of disgust. This contrasting opinion reveals the diverse perspectives within the audience.

Moving on to Example 2 in Fig. 6b, the video focuses on smartphones and presents a visually appealing device. A comment, "please give me one," acknowledges the aesthetic appeal of the phone shown in the video and expresses a desire to possess it. This comment not only highlights the allure of visually appealing products but also emphasizes the psychological impact of such imagery on consumer behavior. It demonstrates how videos can act as persuasive tools in influencing audiences' desires.

In Fig. 6c, the video depicts Lionel Messi walking through the player tunnel and shaking hands with young fans on the opposite side. An interesting comment in response states, "I'd never wash my hand after that." This comment expresses the idea that shaking hands with Messi is such a joyful experience that they would not want to wash their hand afterward, in order to preserve that moment. It represents a positive reaction to the video content and conveys joy. This reaction illustrates how videos with positive or emotionally charged content can evoke strong emotional responses from audience, emphasizing their power to create a sense of connection and emotional engagement.

Lastly, Example 4 presented in Fig. 6d showcases a video of a painting process starting with a rough sketch and gradually completing the entire artwork. A comment exclaims, "details are insane," expressing astonishment at the level of detail in the painting process and acknowledging the outstanding intricacy of the artwork. This comment reflects a positive reaction to the video and acknowledges the presented impressive level of detail. This response not only offers an endorsement of the video's content but also serves as an acknowledgement of the skill and talent of the artist.

Through the aforementioned examples, we can reiterate the objective of our study. Our task entails inferring the opinions toward to the content of the video, as well as the emotions evoked by it. We need to consider the video as the contextual backdrop and pay attention to the intricate interaction between the comments and the video in order to ensure precise inference. These examples shed light on the multifaceted nature of videos and their ability to influence opinions and evoke emotions. Analyzing the comments alongside the videos helps us understand the diverse perspectives, desires, and reactions of the audiences. Understanding the intricate interaction between videos and comments provides valuable insights into the impact and effectiveness of video content in online discourse.

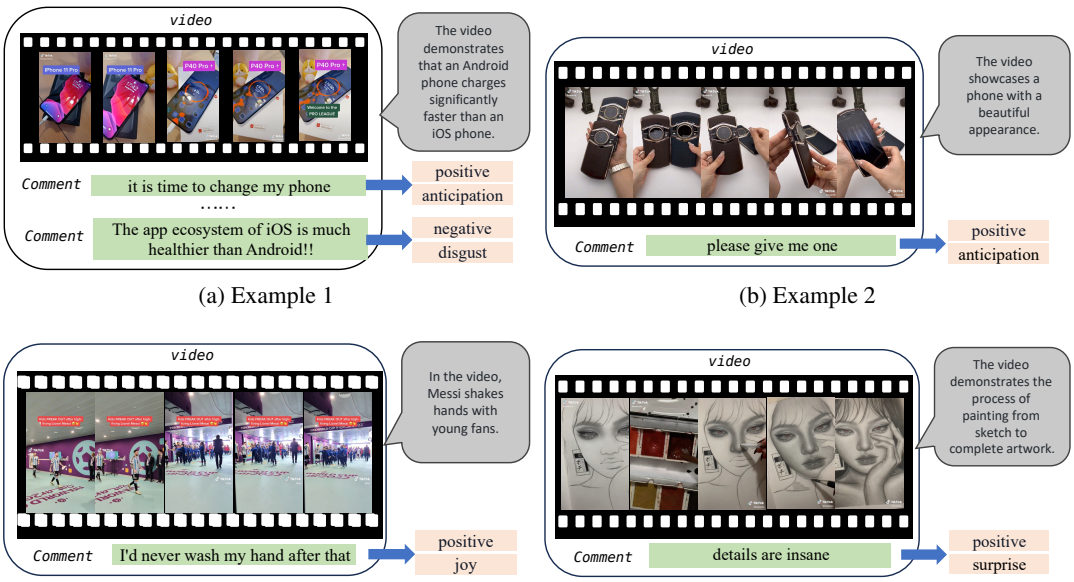
Figure 5: The figure shown the inference results of several test set samples, including the ground truth, inference results of RoBERTa (only text), and our VC-CSA method.

D Datasheet

This section provides a datasheet [6] for **CSMV**. The dataset and our source code is publicly available on Github at `https://github.com/AnonymousUserabc/MVI-MSA_DateAndSource.git`.

D.1 Motivation

• **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.** The prosperity of micro



(a) Example 1 (b) Example 2

(c) Example 3 (d) Example 4

Figure 6: Some examples in the CSMV dataset. To enhance comprehension, a concise description of the video content is presented in a gray box. This description does not serve as input.

video platforms has brought new topics to multi-modal affective computation. The comments of a micro video convey different sentiments response to it. Therefore, we propose a task to infer induced sentiment of comment with understanding the context of the micro video. In light of this, we create a new dataset called **CSMV** (**C**omment **S**entiment toward to **M**icro **V**ideo) to study this task, which includes more than 100k manually annotated comments. The key to the task is to construct the complex induced sentiment interplay between video and comment.

• **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The dataset was created by researchers from IEIT SYSTEMS Co., Ltd., Nankai University and Fudan University.

• **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.** The IEIT SYSTEMS Co., Ltd. provided compute support for collecting the dataset and performing experments. Nankai University and Fudan University provided research support and guidance for the baseline design and the main experiments in the paper.

D.2 Composition

• **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.** The dataset comprises micro video features(visual features which generated by video pre-trained models, including I3D [2], R(2+1)D [10] and VideoMAEv2 [11]) and their associated comments text. Each comment is annotated for opinions and emotions which induced by micro video. Meanwhile, we provide the URLS of the micro video webpage, allowing other researchers to access the original content through web links. The combination of comment and video provides a rich resource for analyzing and understanding the relationship between video content and viewer reactions, particularly in terms of emotional and opinion-based responses.

• **How many instances are there in total (of each type, if appropriate)?** In total, the CSMV dataset comprises 8,210 micro videos, totaling approximately 68.83 hours of video duration, along with 107,267 related comments. The CSMV dataset defines two distinct types of labels, opinion and emotion, for analyzing the sentiment expressed in the comments towards the micro videos.

• What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. The CSMA dataset include the micro video feature and the related comment text. For the micro video, we publishes only the visual features extracted from micro videos using the vidoe pre-trained, including I3D [2], R(2+1)D [10] and VideoMAEv2 [11] models instead of the raw video. For the comment text, we removed any personal information within textual comments (e.g., usernames, emails, phone numbers). Meanwhile, we provide the URLS of the micro video webpage, allowing other researchers to access the original content through web links.

• Is there a label or target associated with each instance? If so, please provide a description. The CSMA dataset labeled opinion and emotion for each comment.

• Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. No.

• Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. No.

• Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. We randomly split our dataset into training, development (dev), and testing sets using a ratio of 7:1:2. We provide these splits, and researchers can also re-split the dataset according to their specific needs.

• Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. We devised a data annotation workflow to ensure annotator quality and reduce individual subjective biases in the annotations.

• Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? It is self-contained.

• Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. No.

• Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. No.

D.3 Collection Process

• How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The data for each instance was acquired from TikTok, including micro video and associated comments. We employ hashtags to collect raw data. Hashtags were manually selected to cover diverse topics like policy, business, sports, and technology.

• What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? We developed a simple software with python to obtain the raw data from TikTok website.

• Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? We employed 30 human annotators to manually label comments. We compensated approximately $\$16.5k$ based on the final number of annotations.

• Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. The CSMV dataset comprises micro videos and their corresponding comments, which have been updated from February 2020 to October 2022.

• Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. Concerning personal privacy, the CSMV dataset would not publish the original videos. Instead, it publishes only the visual features extracted from micro videos using the video pre-trained models. Additionally, the comments solely preserve the text, removing all user-related information. Both the code and data are publicly accessible under the CC BY-NC-SA 4.0 license, intended for academic and non-commercial use.

D.4 Preprocessing/cleaning/labeling

• Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. 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.

• Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. We retained the raw micro videos and comments. If necessary, we will also publish the raw URLs corresponding to each video to facilitate use by other researchers.

D.5 Distribution

• Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. No.

• How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? We will release our dataset via GitHub.

• When will the dataset be distributed? We will release our complete dataset once the paper is confirmed for publication. For now, we have already published a portion of the dataset on GitHub.

• Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. Both the code and data are publicly accessible under the CC BY-NC-SA 4.0 license, intended for academic and non-commercial use.

• Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

• Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

D.6 Maintenance

• Who will be supporting/hosting/maintaining the dataset? This dataset will be maintained by all of authors.

• How can the owner/curator/manager of the dataset be contacted (e.g., email address)? You can contact the dataset authors via the following email: fanbaoyu@ieisystem.com, jiaqi01@ieisystem. Alternatively, you can submit an issue on GitHub.

• Is there an erratum? If so, please provide a link or other access point. No erratum for now.

• Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? We will continuously update our dataset. Our

planned future updates include: 1. Extracting the feature of the raw video frames using ResNet [8]. 2. Incorporating more video visual features, including VideoMAE [9, 12]. 3. Extracting audio signals feature with MFCC and Wav2vec [1]. 4. Further expanding the dataset scale. We intend to implement these updates gradually after the formal publication of the paper.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[2] 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.

[3] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions, 2024.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

[5] Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. LLaVA-MORE: Enhancing Visual Instruction Tuning with LLaMA 3.1, 2024.

[6] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2021.

[7] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc., 2022.

[10] 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.

[11] 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.

[12] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023.