ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230438

Comment-Aware Multi-Modal Heterogeneous Pre-Training for Humor Detection in Short-Form Videos

Yang Liu[†], Huanqin Ping[†], Dong Zhang[†;*], Qingying Sun[§], Shoushan Li[†] and Guodong Zhou[†]

[†]School of Computer Science and Technology, Soochow University, Suzhou
[§]Huaiyin Normal University, Huai'an

Abstract. Conventional humor analysis normally focuses on text, text-image pair, and even long video (e.g., monologue) scenarios. However, with the recent rise of short-form video sharing, humor detection in this scenario has not yet gained much exploration. To the best of our knowledge, there are two primary issues associated with short-form video humor detection (SVHD): 1) At present, there are no ready-made humor annotation samples in this scenario, and it takes a lot of manpower and material resources to obtain a large number of annotation samples; 2) Unlike the more typical audio and visual modalities, the titles (as opposed to simultaneous transcription in the lengthy film) and associated interactive comments in short-form videos may convey apparent humorous clues. Therefore, in this paper, we first collect and annotate a video dataset from DouYin (aka. TikTok in the world), namely DY24h, with hierarchical comments. Then, we also design a novel approach with comment-aided multi-modal heterogeneous pre-training (CMHP) to introduce comment modality in SVHD. Extensive experiments and analysis demonstrate that our CMHP beats several existing video-based approaches on DY24h, and that the comments modality further aids a better comprehension of humor. Our dataset, code and pre-trained models are available at https://github.com/yliu-cs/CMHP.

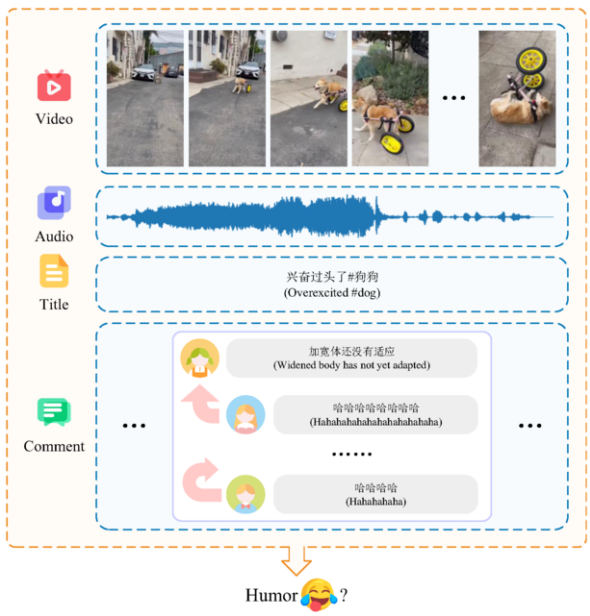


Figure 1: An example for SVHD task in the social media.

1 Introduction

As a new entertaining media, the short-form video applications like DouYin (aka. Tiktok in the world), Kwai, Snapchat, and Snack, enable users to watch, comment on, create, and share entertaining videos that last from seconds to a few minutes. Recently, it is common to watch short-form videos as micro-breaks during leisure time since it has skyrocketed in popularity globally. As we know, effectively identifying video humor can facilitate a platform to make more targeted personalized recommendations based on user preferences. Consequently, it is important to conduct short-form video humor detection (SVHD) with natural multi-modalities. [15, 34].

To this end, we argue that there are two main challenges in SVHD as follows: 1) There is a lack of humor-labeled short-form video dataset. It takes more time than usual to annotate a big scale of video data, compared to the more common textual modality. Furthermore, if we only label a tiny quantity of data, we may get subpar results. 2) Previous work in this area has often concentrated on either text, text-image pairs, or videos (with synchronised transcription, audio, and vision). While the short-form video primarily consists of audio and video modalities, it also includes a title modality that is out of sync with the time series, a large number of hierarchically interactive comments, and some metadata (such as the number of "likes" on the posted video). These detailed extra information may aid in our ability to identify humour. For example, as shown in Figure 1, although it is maybe difficult to infer the humor status from audio and video modalities, the words *excited* in the title and *haha* in the comments can provide intense humor clues.

Thanks to the great ability of pre-training applied in language and vision, we attempt to leverage a multi-modal pre-training paradigm to effectively mine the unlabeled data and utilize the different kinds of video-related information. Specifically, 1) we take the title and comments in the short-form video as different modalities and customize a single-stream model with Transformer [33] to accept multi-modality. This multi-modal architecture can not only unify the pre-training and downstream task (i.e., SVHD), but also alleviate the labeling burden by mining unlabeled data. 2) We design several heterogenous pre-training tasks, such as sequence order modeling (SOM), number of likes classification (NLC), and video-comment matching (VCM). These tasks in multi-modal pre-training could properly take the advantage of heterogeneous information in the short-form video to facilitate humor detection.

* Corresponding Author. Email: dzhang@suda.edu.cn.

Overall, the major contributions are summarized as follows:

- We built the DouYin video dataset (DY24h) with hierarchical interactive comments for the purpose of short-form video humour detection (SVHD);
- To address the challenges in SVHD on our DY24h dataset, we present a novel comment-aided multi-modal pre-training (CMHP) approach. To our best knowledge, this is the first attempt to introduce interactive commentary information from social media in video-related humor analysis;
- Extensive experiments and analysis on DY24h demonstrate the effectiveness of the comment modality and our CMHP in SVHD.

2 Related Work

2.1 Humor Detection.

Previous studies normally focus on deep models for textual humor detection [3, 5, 35]. Recently, since multi-modal expressions become more and more popular, multi-modal humor detection has gained more attention. Hasan et al. [15] and Hasan et al. [14] aim to detect the punchline humor label by multi-modal information which involves text (transcriptions), audio, and video extracted from TED talks in both context and punchline. Moreover, many existing studies [2, 6, 7, 36] detect the utterance humor in dialogue scenarios on datasets produced by TV shows from different cultures.

Unlike the above studies, we build a new video dataset from DouYin for humor detection. Furthermore, we utilize the potential interactive data with self-supervised pre-training strategy to reduce the amount of labeled data for SVHD in multi-modal scenarios. Besides, we detect the humor for the entire video instead of specific utterances or punchlines which largely depend on the preceding context.

2.2 Video Understanding.

Previous studies in video understanding normally take object tracking, action recognition, and object segmentation as the challenges for video understanding. Moreover, convolutional networks have long been the standard for backbone architectures [11, 31, 32, 37]. Recently, the great success of Vision Transformer [9] has attracted investigation of Transformer-based architectures for video understanding, several works [4, 10, 21, 30] try to use different methods to model the video content and temporal information. For example, TimeSformer [4] studies five different variants of space-time attention and suggests a factorized space-time attention for its strong speed-accuracy tradeoff.

Different from the above studies, our work starts with the representation of frames instead of patches that normally used in computer vision for pre-training. Besides, we leverage much auxiliary information that may be ignored (title, comments, likes etc.) and employ multi-modal heterogeneous pre-training to fully utilize unlabeled data. In this way, we can effectively determine the humor expressed by short-form videos with limited resources.

2.3 Video-Language Pre-Training.

Video presence is often accompanied by multi-modal data and video-language pre-training has been verified to be effective for numerous downstream tasks. Therefore, design effective pre-training tasks is crucial to the success of pre-trained models in downstream tasks. The related pre-training tasks in previous studies can be divided into 4 categories: 1) Completion Tasks: aim to reconstruct the masked tokens of input, e.g., Masked Frame Modeling (MFM) [19, 26, 41], Masked Modal Modeling (MMM) [23, 38] Masked Autoencoding (MAE) [29], etc. 2) Matching Tasks: designed to learn the alignment between different modalities, e.g., Video-Audio Matching (VAM) [29] or Video-Language Matching (VLM) [19, 38] as classical matching task originating from Next Sentence Prediction (NSP) of BERT [8]. 3) Ordering Tasks: to learn plenty semantics from each modality by recognizing the sequence order from randomly shuffled sequence, e.g., Frame Ordering Modeling (FOM) [18, 19]. 4) Contrast Tasks: contrastive learning (CL) narrows the embeddings from similar samples and distinguishes the different samples [1, 17].

Dissimilar to the above studies, we employ three heterogeneous pre-training tasks (multi-modality ordering, likes classification and video-comment matching) for video to properly leverage interactive information from social media. The three tasks do not use costly text such as transcription or caption. Instead, we adopt the text such as the title of the video, and all interactive comments from social media, which are easily accessible along with the video.

3 Methodology

Figure 2 shows the overall architecture and training strategies of our proposed approach for short-form video humor detection (SVHD).

3.1 Task Definition

Given sample $s_i = (v_i, a_i, t_i, c_i)$, where v_i denotes i-th video and a_i, t_i, c_i denotes its accompany audio, title, and comments, we aim to train a model $\mathcal{F}((v_i, a_i, t_i, c_i); \Theta) \mapsto y_i$ to detect humor of video v_i correctly by using unlabeled and few labeled data via pre-training & fine-tuning. Sample s_i is associated with a binary label $y_i \in \{0, 1\}$ to represent its humor label, where $y_i = 1$ indicates s_i is humor, and $y_i = 0$ means s_i is non-humor.

3.2 Uni-modal Embedding

We first obtain embedding representations for every modality:

Video. Raw data of video v_i can be expressed as $v_i \in \mathbb{R}^{l \times c \times h \times w}$, where l, c, h, w denotes the number of frames, number of channels, height, and width. In order to alleviate the amount of computation of huge video data, we feed v_i to a frozen pre-trained ResNet152 [16] to obtain the frame representation for every frame. Then map frames to a latent space via multi-layer perception (MLP):

$$e_i^v = \text{MLP}(\text{ResNet152}(v_i)) \tag{1}$$

where $e_i^v \in \mathbb{R}^{l \times d}$, d denotes the dimension of embedding.

Audio. As for audio modality, we expect to feed the raw waveform from audio to model. Thus, we extract the waveform a_i of the audio at f-kHz with a total duration of o seconds by torchaudio [39]. To match the sampling rate of the video, we averaged the same parts of the vector representation $a_i \in \mathbb{R}^{l \times \frac{o \cdot f}{l}}$. Then similar to video modality, mapped a_i to the same latent space via MLP:

$$e_i^a = \text{MLP}(a_i) \tag{2}$$

where $e_i^a \in \mathbb{R}^{l \times d}$.

Title. We use the tokenization strategy same as BERT [8] (Chinese version). Given a text token sequence of title t_i, in order to be able to conduct the experiment with limited resources, we use WideMLP [12] to obtain the embedding. Multi-layer perception is a simple yet effective way for categorizing text:

$$e_i^t = \text{WideMLP}(t_i) \tag{3}$$

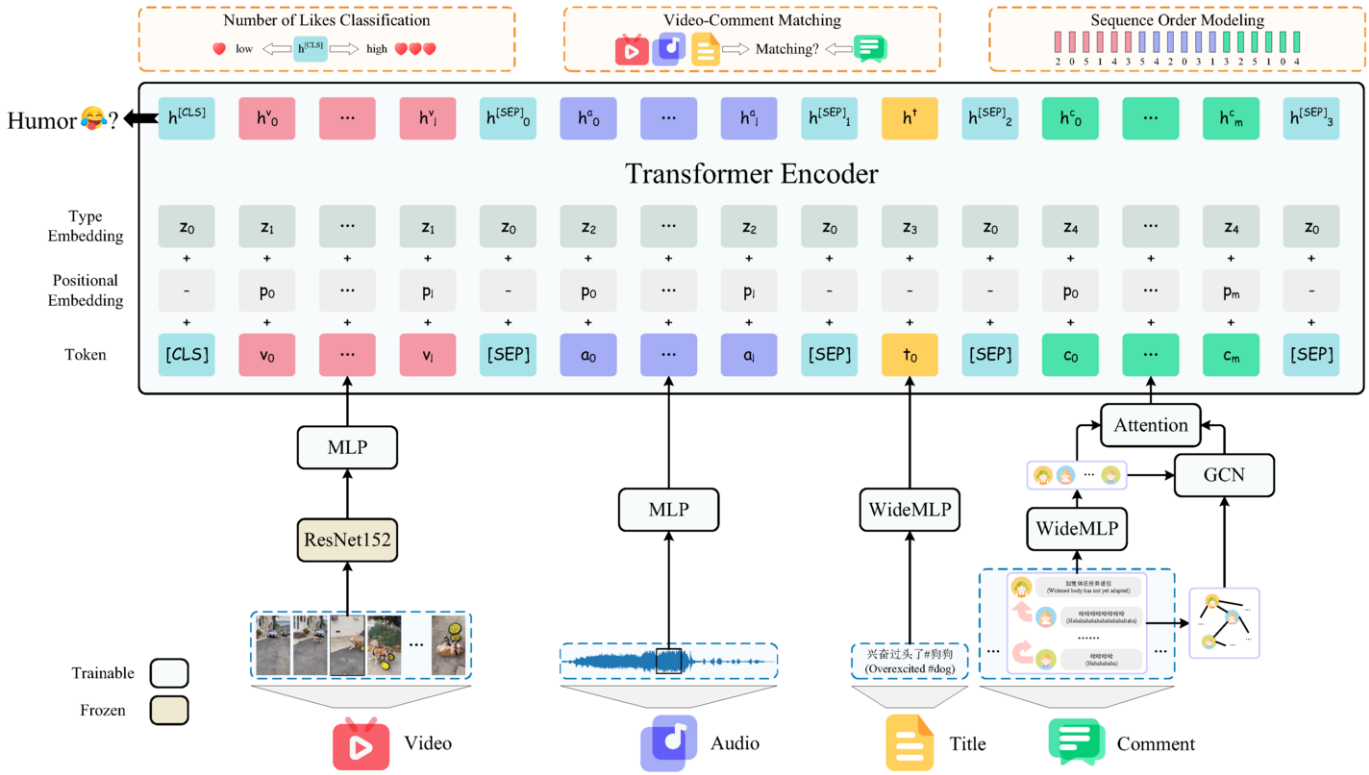


Figure 2: The architecture of our proposed pre-training approach for video humor detection task.

where $e_i^t \in \mathbb{R}^d$.

Comment. Accompanying comments c_i of sample s_i could split to 2 hierarchy: 1) every sample s_i consist m comment blocks $c_i = \{c_{i,j}\}_{j=1}^m$, where $c_{i,j}$ denotes the j-th comment block, 2) every comment block $c_{i,j}$ consist 1 comment and r replies about it $c_{i,j} = \{c_{i,j,k}\}_{k=1}^{r+1}$, where $c_{i,j,k}$ denotes k-th ($k \in [2, r+1]$) reply about the comment ($c_{i,j,1}$ denotes the comment in comment block $c_{i,j}$). Hence, the comments c_i can be expressed as $c_i = \{\{c_{i,j,k}\}_{k=1}^{r+1}\}_{j=1}^m$. Similar to the processing of the title, we use WideMLP [12] to obtain the textual representation of each comment. For every text sequence, the procedure could be stated as:

$$c'_{i,j,k} = \text{WideMLP}(c_{i,j,k}) \tag{4}$$

$$c'_{i,j} = [c'_{i,j,1}, c'_{i,j,2}, \cdots, c'_{i,j,r+1}] \tag{5}$$

where $c'_{i,j,k} \in \mathbb{R}^d$, $c'_{i,j} \in \mathbb{R}^{(r+1)\times d}$. For every comment block, we build undirected graph $\mathcal{G}_{i,j}$ according to their response relationship, the adjacency matrix could define as:

$$\mathcal{G}_{i,j} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,r+1} \\ \vdots & \ddots & \vdots \\ a_{r+1,1} & \cdots & a_{r+1,r+1} \end{bmatrix} \tag{6}$$

$$a_{u,v} = \begin{cases} 1, & \text{if} \quad \mathcal{R}(c_{i,j,u}, c_{i,j,v}) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{R}(c_{i,j,u}, c_{i,j,v}) = 1$ denotes there exist a response relationship between comment $c_{i,j,u}$ and $c_{i,j,v}$. We use Graph Convolutional Network (GCN) to extract information from the text content of the comments and the reply relationship:

$$g_{i,j} = \text{GCN}(c'_{i,j}, \mathcal{G}_{i,j}) \tag{8}$$

where $g_{i,j} \in \mathbb{R}^{(r+1)\times d}$. Inspired by Liang et al. [20], we employ a retrieval-based attention mechanism to capture the graph-oriented attention features from the comments representations $c'_{i,j}$ on the graph representation $g_{i,j}$ derived from GCN:

$$\alpha_{i,j} = \text{Softmax}(g_{i,j}^\top c'_{i,j}) \tag{9}$$

where $\alpha_{i,j}$ denotes the attention score in comment block $c_{i,j}$, $\top$ represents the matrix transposition. The final embedding of comments c_i is defined as:

$$e_{i,j}^c = \sum \alpha_{i,j}^\top g_{i,j} \tag{10}$$

$$e_i^c = [e_{i,1}^c, e_{i,2}^c, \cdots, e_{i,m}^c] \tag{11}$$

where $e_{i,j}^c \in \mathbb{R}^d$, $e_i^c \in \mathbb{R}^{m\times d}$ indicates the embedding of comment block $c_{i,j}$ and comments c_i.

3.3 *Multi-modal Encoding*

After obtaining embeddings of all modalities, we put them into sequence by concatenating: $e_i = [\text{CLS}] \oplus e_i^v \oplus [\text{SEP}] \oplus e_i^a \oplus [\text{SEP}] \oplus e_i^t \oplus [\text{SEP}] \oplus e_i^c \oplus [\text{SEP}]$, where $e_i \in \mathbb{R}^{(2l+m+6)\times d}$, $\oplus$ denotes concatenate operation.

To simplify multi-modal pre-training, we adopt a Single-Stream Transformer [33] encoder. The entire feature sequence e_i is fed into it after summing with the untrainable positional embedding p_i and trainable token type embedding z_i:

$$h_i = \text{Encoder}(e_i + p_i + z_i) \tag{12}$$

We add positional embedding for video, audio, and comment modalities, where the comment modality is sorted in descending order of likes when the input, so its positional embedding represents

the order of likes. Then, we use $h_i^{[\text{CLS}]}$, h_i^v, h_i^a, h_i^t, h_i^c respectively denotes the hidden states corresponding to [CLS], video, audio, title and comment tokens of the last layer of Transformer encoder.

3.4 Heterogeneous Pre-training

To enhance the video comprehension and presentation abilities of model, we design three heterogeneous pre-training tasks.

Sequence Order Modeling (SOM). To *increase representation capability for audio and video from recovers accurate frame order* [19], we develop sequence ordering tasks for video, audio, and comment modalities. For video and audio, we expect to correctly recover their temporal order from randomly shuffled sequences to learn the original order of randomly reordered frames. Also for comments, we anticipate that the random disordered sequence will appropriately recover the order of the number of likes on the comment content of each comment block. We believe it is more likely to *implicitly incorporate information from social media interaction commentary to assist the model grasp user-preference-based content semantics*. Therefore, models rank sequence order appropriately by understand content can identify humor. Note that, we collected data by the platform's visitor (non-logged-in) video recommendations, these data represent its mainstream preferences, so the 10 "most highly liked" comments and their replies for each sample on "mainstream" videos have a limited correlation between likes and post time. Thus, we concatenate three hidden states of various modalities, h_i^v, h_i^a and h_i^c, in order to execute sequence order using multi-layer perception and a softmax function: $\hat{y}_i^\flat = \text{Softmax}(\text{MLP}(h_i^v \oplus h_i^a \oplus h_i^c))$ where $\hat{y}_i^\flat$ represents a list of predicted sequence order. Then, the cross-entropy loss function is used to calculate the loss for the SOM task:

$$\mathcal{L}_{\text{SOM}} = -\sum_{i=1}^{n}\sum_{j=1}^{2l+m} y_{i,j}^\flat \log \hat{y}_{i,j}^\flat \tag{13}$$

where $y_i^\flat$ is the order index of the ground truth list.

Number of Likes Classification (NLC). To *make efficient usage of social media feedback data to assist model $\mathcal{F}$ in learning video information*, we design the number of likes classification task. However, we believe a task gap remains between regression and final humor detection, thus we convert the regression task to a classification task based on the data shown in Figure 3e. Based on this, we choose 100k as the dividing line between two categories of video likes: low and high. The ground truth labels $y_i^\natural \in \{0, 1\}$ for the NLC problem correspond to the values low and high. Note that as mentioned in SOM, the relationship between the number of likes for "mainstreams" videos and the time of collected is limited. We apply a multi-modal hybrid representation $h_i^{[\text{CLS}]}$ to the number of similar classifications via MLP and a softmax function: $\hat{y}_i^\natural = \text{Softmax}(\text{MLP}(h_i^{[\text{CLS}]}))$. Then, we calculate the loss for the NLC task like SOM:

$$\mathcal{L}_{\text{NLC}} = -\sum_{i=1}^{n} y_i^\natural \log \hat{y}_i^\natural \tag{14}$$

Video-Comment Matching (VCM). To *capture fine-grained alignment between video with its raw information (audio and title) and comment*, we built the binarized video-comment match task that can be trivially constructed from any media video. In VCM, comments have a 50% probability of matching with other modalities (which are the same sample) and a 50% probability of not matching (any other sample of comment modalities). The ground truth labels $y_i^\sharp \in \{0, 1\}$ for the VCM task represents matches and mismatches respectively. We adopt the multi-modal hybrid representation $h_i^{[\text{CLS}]}$ to identify video-comment matching by MLP and a softmax function: $\hat{y}_i^\sharp = \text{Softmax}(\text{MLP}(h_i^{[\text{CLS}]}))$. Same as SOM and NLC, we calculate the loss of VCM as:

$$\mathcal{L}_{\text{VCM}} = -\sum_{i=1}^{n} y_i^\sharp \log \hat{y}_i^\sharp \tag{15}$$

Total Loss. Finally, the overall heterogeneous pre-training objective of CMHP is:

$$\mathcal{L} = \mathcal{L}_{\text{SOM}} + \mathcal{L}_{\text{NLC}} + \mathcal{L}_{\text{VCM}} + \lambda||\Theta||^2 \tag{16}$$

where Θ denotes all trainable parameters of the model $\mathcal{F}$, λ represents the coefficient of L_2-regularization. We optimize parameters Θ by minimizing pre-training loss $\mathcal{L}$.

3.5 Learning Objective

We adopt multi-modal hybrid representation $h_i^{[\text{CLS}]}$ to infer the humor status in short-form video by multi-layer perception and a softmax function: $\hat{y}_i = \text{Softmax}(\text{MLP}(h_i^{[\text{CLS}]}))$. Then use the cross-entropy loss function for SVHD:

$$\mathcal{J} = -\sum_{i=1}^{n} y_i \log \hat{y}_i + \lambda||\Theta||^2 \tag{17}$$

where Θ denotes all trainable parameters of the model $\mathcal{F}$, λ represents the coefficient of L_2-regularization. During fine-tuning, we optimize parameters Θ by minimizing $\mathcal{J}$.

4 Experimentation

In this section, we describe our DY24h dataset and analyze the baselines and CMHP systematically.

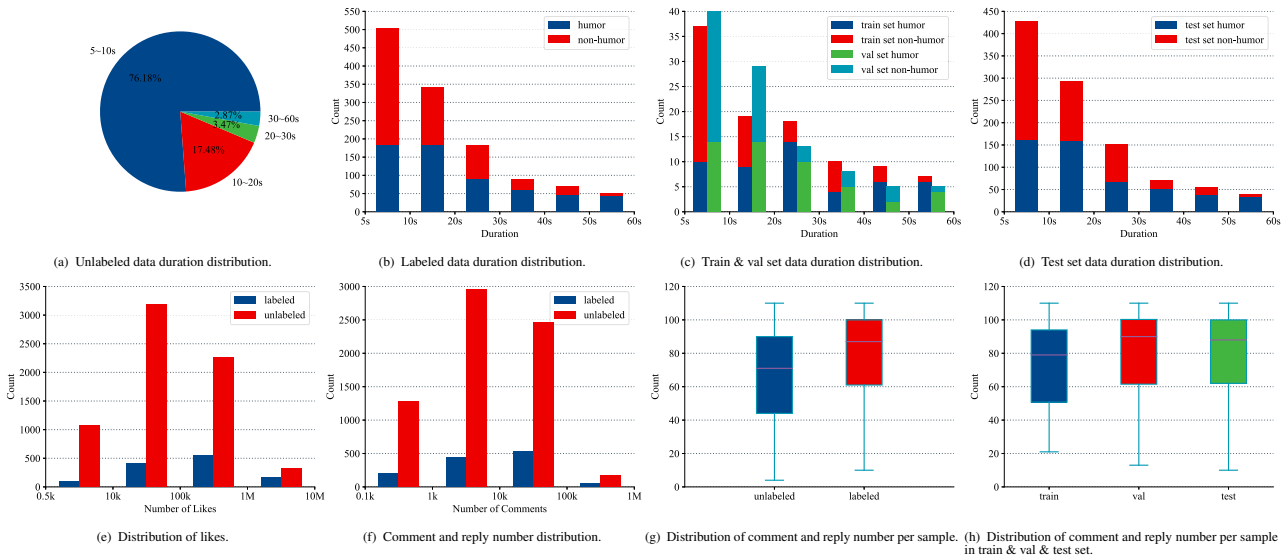
4.1 Dataset

We create one of our own to evaluate the multi-modal (short-form video) humour detection task due to a lack of publicly available datasets: we collected a total of 8090 videos (24.3 hours) within one minute from DouYin portal[1] by the platform guide strategy for visitor (non-logged-in) user, this represent the mainstream preferences of platform. For each collected sample, we crawl up to 10 "most highly liked" comment blocks, each containing its own comment content and up to 10 replies. The number of comment blocks and the number of replies under each comment block vary for each sample, and the statistics for them are shown in Figure 3g and 3h. The filtering requirement of more than 500 likes and more than 100 comments was applied to make the gathered videos more representative. Figures 3e and 3f show the sample sizes of different statistical intervals of the number of likes and comments, respectively. To prevent import data bias by browsing preferences, we regularly restart the crawler in an environment without a browsing history and login.

For unlabeled pre-training data, 6855 samples (18.3 hours) were used for the self-supervised pre-training, almost 80% of video durations are under 10 seconds, as shown in Figure 3a.

For labeled humor detection data, we employ two graduate students as annotators to manually labeled 1235 samples (6 hours) by

[1] https://www.douyin.com.



(a) Unlabeled data duration distribution. (b) Labeled data duration distribution. (c) Train & val set data duration distribution. (d) Test set data duration distribution.

(e) Distribution of likes. (f) Comment and reply number distribution. (g) Distribution of comment and reply number per sample. (h) Distribution of comment and reply number per sample in train & val & test set.

Figure 3: Overview of DY24h statistics.

Table 1: The statistics summary of our dataset.

Unlabeled	Labeled		
	1235		
6855	Train 100	Val 100	Test 1035
Humor	49	49	510
Non-Humor	51	51	525

our developed annotation toolkit based on humor classification, and the Cohen's kappa score of annotation is 0.818. Similar to unlabeled data, the amouynt of videos is inversely proportional to the duration, as revealed by Figure 3b. Figure 4 shows the annotation toolkit we devoleped, it is a general toolkit for video-based multi-modal annotation, with a complete video playback control function. When playing a video, it can also display a variety of complete accompanying information, such as the title of the video, the user of the release, the release time, the number of likes/favorites/comments, and replies of comments. Then, we randomly divided the 1235 labeled samples into training, validation, and test sets based on the labels in a balanced manner as shown in Table 1. The training and validation sets each have a few numbers of 100 samples. Different sets contain videos of varying durations evenly, as can be seen in Figure 3c and 3d.

4.2 Implementation Details

We sample 5 frame per second from video and resize frames to a spatial size of 224×224. We sample audio waveforms in sync at 16-kHz. The maximum length of text (title and comments) token sequence is 16 for each sentence, and we discard the *#Funny* and *#Funny Video* hashtags from all text content. AdamW [22] is utilized as the optimizer and the learning rate is 5e-6, 5e-5 for pre-training and fine-tuning. The coefficient of L_2-regularization λ is set to 1e-5 for all optimizer. The dimension of uni-modal embedding in our model is 768. The number of all multi-layer perception layers is set to 2, and the graph convolutional network is set to 4 layers. The number of heads in the Transformer encoder is set to 12, and the encoder is set to 12 layers. The dropout rate is set to 0.1. We implement our approach via Pytorch [25] toolkit (1.12.1+cu116), and pre-train a CMHP with

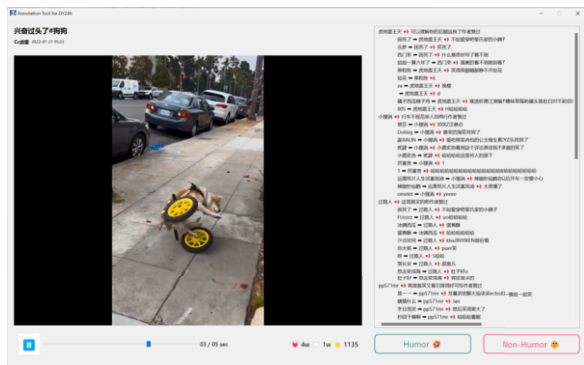


Figure 4: The interface of our annotation toolkit.

batch size 16 and epoch 300 on a single Nvidia RTX 3090 (24GB VRAM) takes less than one day.

4.3 Baselines

To put our results in perspective, we compare to several competitive baselines in multi-modal video understanding:

- TVLT [29]: a competitive multi-modal approach to video comprehension (e.g., image retrieval, video retrieval, and multi-modal sentiment analysis) employing visual-acoustic pre-training. However, since the model is specifically designed for vision and audio modality, we cannot directly implement this approach on our proposed dataset (*DY24h*) during pre-training;
- LF-VILA [28]: the state-of-the-art in long-form (similar to our short-form settings) video understanding (e,g., cross-model retrieval and long-form video classification) with video-language pre-training on their dataset. We adapt the title and comments in our task to the language modality of this model LF-VILA, then conduct humor detection. However, because this model is designed specifically for vision and language, we cannot directly input audio modality. Besides, to maximise the benefits of the released pre-trained model, we translate the title and comments into English;
- BBFN [13]: a competitive approach in emotion detection of videos without multi-modal pre-training. Although this approach concen-

Table 2: Performance comparison of several competitive baseline approaches and our CMHP for SVHD task. V, A, T, C indicates different modalities of video, audio, title, and comment accordingly. Best scores are in **bold**.

Modality	Approach	Pre-training	Acc	Macro			Weighted		
				Pre	Rec	F1	Pre	Rec	F1
V + A	TVLT [29]	DY24h	69.855	70.650	70.118	69.720	70.796	69.855	69.660
		HT100M[24] + YTT180M[40]	75.169	75.174	75.098	75.116	75.172	75.169	75.150
V + T + C	LF-VILA [28]	DY24h	75.169	75.320	75.271	75.166	75.405	75.169	75.157
		LF-VILA-8M[28]	74.493	74.529	74.495	74.484	74.530	74.493	74.484
V + A + T + C	BBFN [13]	-	74.203	75.084	74.194	73.970	75.080	74.203	73.972
	CubeMLP [27]	-	69.179	70.033	69.169	68.839	70.029	69.179	68.843
	VATT [1]	DY24h	75.459	76.148	75.686	75.389	76.300	75.459	75.350
	CMHP (Ours)	-	74.589	75.964	74.417	74.156	75.882	74.589	74.205
		DY24h	**76.039**	**76.646**	**75.924**	**75.844**	**76.593**	**76.039**	**75.875**

Table 3: Accuracy with different pre-training tasks.

Pre-Training Task	Acc
SOM	74.976
SOM + NLC	75.169
SOM + NLC + VCM	**76.039**

Table 4: Accuracy with different pre-training scales.

Scale	Acc
-	74.589
1k	74.493
4k	75.459
7k	**76.039**

trates solely on the temporal vision, audio and text, we transfer the title and comments in our task to the text modality of this model and then execute humor detection;

• CubeMLP [27]: the state-of-the-art in emotion detection of videos without pre-training. Although CubeMLP only focuses on the temporal vision, audio and text, we adapt the title and comments in our task to the text modality of this model, then conduct SVHD;

• VATT [1]: the state-of-the-art in multi-modal video understanding (e.g. video action recognition, audio event classification, image classification, and text-to-video retrieval) with visual-acoustic-language pre-training. For a fair comparison, we adapt the title and comment in our task to the language modality of this model. Moreover, we implement this approach by pre-training on our proposed dataset (*DY24h*), then conduct humor detection.

4.4 Main Results

Table 2 contrasts the experimental results of our CMHP to a number of competitive baselines. The performance of the baselines without pre-training (BBFN and CubeMLP) is inferior to the baselines with pre-training (TVLT, LF-VILA, and VATT). This suggests that, despite using all available modalities, BBFN and CubeMLP lose too much prospective knowledge from unlabeled data for humor detection. Utilising multiple modalities improves video comprehension and humor detection in the pre-training approach. By integrating all four modalities, our CMHP improves performance, surpassing all baselines on the DY24h dataset. This is primarily the result of heterogeneous pre-training, which simultaneously produces exceptional multi-modal video comprehension for SVHD.

5 Analysis

In this section, we analyze the effectiveness of the comments and the role played by pre-training in the SVHD task.

5.1 Pre-training

We used all four modalities uniformly to examine the effect of pre-training on the final humor detection results.

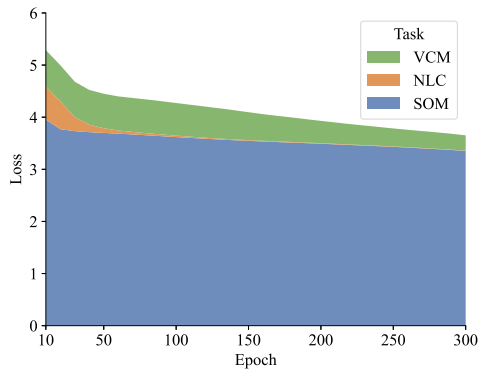
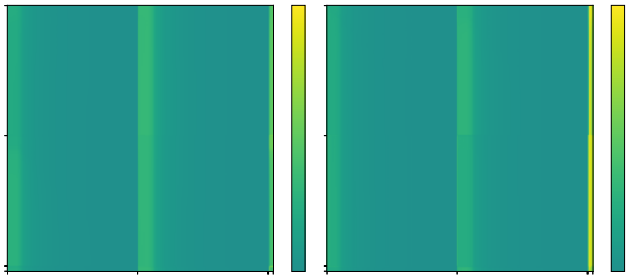


Figure 5: Loss changes during pre-training.

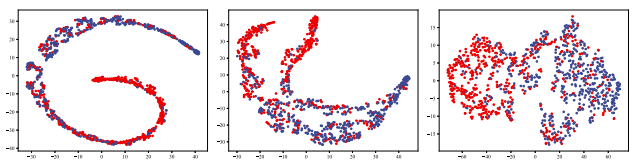
Effect of different pre-training task combinations. Here, we investigate the influence of the different pre-training task combinations by varying {SOM, SOM+NLC, SOM+NLC+VCM} and conducting experiments. Table 3 displays the performance variation resulting from our CMHP using pre-training data of varying tasks. From this table, we can see that more pre-training task corresponds to better performance, this shows that each pre-training task can bring improvements for video content understanding to the model from different perspectives. In addition, Figure 5 shows how the model keeps improving while fitting the pre-training data, we discover that the VCM loss remains constant during pre-training but begins to decrease after the NLC task is fitted. This indicates that the model can only complete the matching task if it has a certain level of multi-modal video comprehension; otherwise, it cannot be improved through this task.

Effect of different pre-training scale. Further, we investigate the influence of the pre-training scale by varying {0, 1k, 4k, 7k} samples and conducting experiments. Table 4 displays the performance variation resulting from our CMHP using pre-training data of varying scales. From this table, we can see that larger pre-training scale corresponds to better performance, so our present bottleneck in SVHD is due to insufficient unlabeled pre-training data scale. In addition, us-



(a) Without pre-training. (b) With pre-training.

Figure 6: Attention visualization of Encoder. The four axis regions from left to right (from top to bottom) represent the four modalities V, A, T and C respectively.



(a) Without pre-training. (b) With shallow pre-training. (c) With deep pre-training.

Figure 7: t-SNE visualization of multi-modal representation.

ing too few pre-training data (e.g., 1k) may introduce bias or noise, resulting in poorer performance than without pre-training. Consequently, our pre-training tasks can be readily adapted to any video in social media; this creates a favourable opportunity for more extensive general-purpose pre-training, which will become our future work.

The visualization of encoder. We provide the visualization of self-attention in the encoder with and without our proposed three heterogeneous pre-training, as illustrated in Figure 6. From this figure, we can see pre-training stage clearly enhances the degree to which the model focuses on comments data (the interval on the far right of the coordinate axes). Moreover, comments information is more involved in the interaction with other modality in the self-attention mechanism, indicating that the pre-trained model learns better connections between different modalities and can comprehend the video content more thoroughly with the aid of comments text content, in order to capture more effective humorous clues.

On the other hand, we illustrate the t-SNE visualization of the output feature representations from Encoder in our CMHP with (shallow or deep) and without our heterogeneous pre-training, as shwon in Figure 7. From this figure, we can observe that the distribution after pre-training becomes more uniform and the distinction between two humor categories is clearer, compared to the distribution without pre-training. In addition, for pre-training, deeper pre-training results in more evenly distributed and distinct multi-modal fusion feature representation, suggesting that the model could identify the humorous labels of the video across more dimensions after pre-training. Therefore, outstanding feature representation enables the model to more accurately infer video humor labels.

Validation accuracy curves Then, Figure 8 depicts the trend of increasing validation accuracy during training with and without pre-training. Based on the general and excellent video multi-modal understanding ability brought about by the pre-training, the model has a faster fitting speed in the humor detection training after the pre-training (red curve), and the accuracy of the validation set in the training stage is also improved in comparison to that without the pre-training (blue curve).

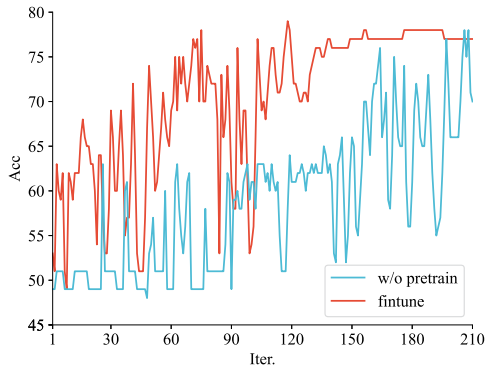


Figure 8: Accuracy curves of validation set during training.

Table 5: Accuracy with different modalities without pre-training. Since we wish to isolate the contribution of distinct modalities to the SVHD while excluding the influence of pre-training.

Modality	Acc
V	58.647
V + A	61.063
V + A + C	68.889
V + A + C + T	**74.589**

5.2 *Effect of Different Modality Combinations*

To isolate the contribution of distinct modalities to the humor detection task while excluding the influence of pre-training, we demonstrate the effectiveness of the CMHP without pre-training in the various modalities combination scenarios by comparing the performance in Table 5. This table demonstrates that CMHP performs optimally when all four modalities are employed, as each modality provides informational support for the overall humor detection perspective. Furthermore, textual modes offer the greatest performance enhancement. Specifically, the comments modality results in a nearly 8% increase, and the title modality results in a 6% increase. This demonstrates the efficacy of our innovative implementation of social media-specific interactive textual modality information.

6 Conclusion

In this paper, we focus on incorporating social media interaction comments data into the model to better understand humorous emotions in videos. To tackle this problem, we develop a novel comment-aided multi-modal heterogeneous pre-training (CMHP) approach and construct a new multi-modal dataset, DY24h, containing hierarchical comments. Through extensive experimentations and analysis, we show that our CMHP surpass more complicated or pre-trained on large-scale data approaches by making use of interactive comments that is publicly available.

Expanding our approach to encompass other cultures and task settings, as well as increasing the size of our unlabeled short-form video data, are goals for our future study.

Acknowledgements

This work was supported by NSFC grants No. 62206193, No. 62076176, and No.62006093.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, 'VATT: transformers for multimodal self-supervised learning from raw video, audio and text', in *Proc. of NeurIPS*, pp. 24206–24221, (2021).

[2] Khalid Alnajjar, Mika Hämäläinen, Jörg Tiedemann, Jorma Laaksonen, and Mikko Kurimo, 'When to laugh and how hard? A multimodal approach to detecting humor and its intensity', in *Proc. of COLING*, pp. 6875–6886, (2022).

[3] Issa Annamoradnejad, 'Colbert: Using BERT sentence embedding for humor detection', *CoRR*, (2020).

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, 'Is space-time attention all you need for video understanding?', in *Proc. of ICML*, pp. 813–824, (2021).

[5] Vladislav Blinov, Valeria Bolotova-Baranova, and Pavel Braslavski, 'Large dataset and language model fun-tuning for humor recognition', in *Proc. of ACL*, pp. 4027–4032, (2019).

[6] Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya, 'A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting', in *Proc. of COLING*, pp. 6752–6761, (2022).

[7] Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-Philippe Morency, and Soujanya Poria, 'M2H2: A multimodal multiparty hindi dataset for humor recognition in conversations', in *Proc. of ICMI*, pp. 773–777, (2021).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: pre-training of deep bidirectional transformers for language understanding', in *Proc. of NAACL*, pp. 4171–4186, (2019).

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, 'An image is worth 16x16 words: Transformers for image recognition at scale', in *Proc. of ICLR*, (2021).

[10] Christoph Feichtenhofer, haoqi fan, Yanghao Li, and Kaiming He, 'Masked autoencoders as spatiotemporal learners', in *Proc. of NeurIPS*, pp. 35946–35958, (2022).

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, 'Slowfast networks for video recognition', in *Proc. of ICCV*, pp. 6201–6210, (2019).

[12] Lukas Galke and Ansgar Scherp, 'Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP', in *Proc. of ACL*, pp. 4038–4051, (2022).

[13] Wei Han, Hui Chen, Alexander F. Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria, 'Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis', in *Proc. of ICMI*, pp. 6–15, (2021).

[14] Md. Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque, 'Humor knowledge enriched transformer for understanding multimodal humor', in *Proc. of AAAI*, pp. 12972–12980, (2021).

[15] Md. Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md. Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque, 'UR-FUNNY: A multimodal language dataset for understanding humor', in *Proc. of EMNLP*, pp. 2046–2056, (2019).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proc. of CVPR*, pp. 770–778, (2016).

[17] Jingjia Huang, Yinan Li, Jiashi Feng, Xiaoshuai Sun, and Rongrong Ji, 'Clover: Towards A unified video-language alignment and fusion model', *CoRR*, (2022).

[18] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li, 'Understanding chinese video and language via contrastive multimodal pre-training', in *Proc. of ACM MM*, pp. 2567–2576, (2021).

[19] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu, 'HERO: hierarchical encoder for video+language omni-representation pre-training', in *Proc. of EMNLP*, pp. 2046–2065, (2020).

[20] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu, 'Multi-modal sarcasm detection via cross-modal graph convolutional network', in *Proc. of ACL*, pp. 1767–1777, (2022).

[21] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, 'Video swin transformer', in *Proc. of CVPR*, pp. 3192–3201, (2022).

[22] Ilya Loshchilov and Frank Hutter, 'Decoupled weight decay regularization', in *Proc. of ICLR*, (2019).

[23] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou, 'Univilm: A unified video and language pre-training model for multimodal understanding and generation', *CoRR*, (2020).

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic, 'Howto100m: Learning a text-video embedding by watching hundred million narrated video clips', in *Proc. of ICCV*, pp. 2630–2640, (2019).

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. 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 *Proc. of NeurIPS*, pp. 8024–8035, (2019).

[26] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid, 'Videobert: A joint model for video and language representation learning', in *Proc. of ICCV*, pp. 7463–7472, (2019).

[27] 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 *Proc. of ACM MM*, pp. 3722–3729, (2022).

[28] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu, 'Long-form video-language pre-training with multimodal temporal contrastive learning', in *Proc. of NeurIPS*, pp. 38032–38045, (2022).

[29] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal, 'Tvlt: Textless vision-language transformer', in *Proc. of NeurIPS*, pp. 9617–9632, (2022).

[30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang, 'Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training', in *Proc. of NeurIPS*, pp. 10078–10093, (2022).

[31] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, 'Learning spatiotemporal features with 3d convolutional networks', in *Proc. of ICCV*, pp. 4489–4497, (2015).

[32] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, 'A closer look at spatiotemporal convolutions for action recognition', in *Proc. of CVPR*, pp. 6450–6459, (2018).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Proc. of NeurIPS*, pp. 5998–6008, (2017).

[34] Yunwen Wang, 'Humor and camera view on mobile short-form video apps influence user experience and technology-adoption intent, an example of *TikTok* (*DouYin*)', *Comput. Hum. Behav.*, 106373, (2020).

[35] Orion Weller and Kevin D. Seppi, 'Humor detection: A transformer gets the last laugh', in *Proc. of EMNLP*, pp. 3619–3623, (2019).

[36] Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu, 'MUMOR: A multimodal dataset for humor detection in conversations', in *Proc. of NLPCC*, pp. 619–627, (2021).

[37] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, 'Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification', in *Proc. of ECCV*, pp. 318–335, (2018).

[38] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer, 'VLM: task-agnostic video-language model pre-training for video understanding', in *Proc. of ACL Findings*, pp. 4227–4239, (2021).

[39] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jeff Hwang, Ji Chen, Peter Goldsborough, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, and Vincent Quenneville-Bélair, 'Torchaudio: Building blocks for audio and speech processing', in *Proc. of ICASSP*, pp. 6982–6986, (2022).

[40] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi, 'MERLOT: multimodal neural script knowledge models', in *Proc. of NeurIPS*, pp. 23634–23651, (2021).

[41] Linchao Zhu and Yi Yang, 'Actbert: Learning global-local video-text representations', in *Proc. of CVPR*, pp. 8743–8752, (2020).