VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation

Weiying Wang
School of Information
Renmin University of China
wy.wang@ruc.edu.cn

Jieting Chen
School of Information
Renmin University of China
jietingchen@ruc.edu.cn

Qin Jin*
School of Information
Renmin University of China
qjin@ruc.edu.cn

ABSTRACT

Live video interactive commenting, a.k.a. danmaku, is an emerging social feature on online video sites, which involves rich multimodal information interaction among viewers. In order to support various related research, we build a large scale video interactive comments dataset called **VideoIC**, which consists of 4951 videos spanning 557 hours and 5 million comments. Videos are collected from popular categories on the 'Bilibili' video streaming website. Comparing to other existing danmaku datasets, our VideoIC contains richer and denser comments information, with 1077 comments per video on average. High comment density and diverse video types make VideoIC a challenging corpus for various research such as automatic video comments generation. We also propose a novel model based on multimodal multitask learning for comment generation (MML-CG), which integrates multiple modalities to achieve effective comment generation and temporal relation prediction. A multitask loss function is designed to train both tasks jointly in the end-to-end manner. We conduct extensive experiments on both VideoIC and Livebot datasets. The results prove the effectiveness of our model and reveal some features of danmaku.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Computer vision**.

KEYWORDS

danmaku dataset;comments generation;multimodal interaction

ACM Reference Format:
Weiying Wang, Jieting Chen, and Qin Jin. 2020. VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413890

1 INTRODUCTION

Live video commenting, commonly known as "danmaku" or "bullet screen", is an emerging feature on online video sites, which allows viewers to post real-time comments anonymously to fly across the screen like bullets. Bilibili [1], one of the most popular online video platforms in China, has attracted 172 million active users per month in the first quarter of 2020 for its danmaku feature.

Different from traditional video comments which do not have to point to specific moment of video or interact with each other, danmaku have some unique interactive features. Firstly, danmaku involve rich multimodal information interaction. Viewers post real-time comments about the video content or respond to comments from others while watching the video. As shown in part (b) of Figure 1, viewers are heatedly discussing the ingredients of the dish in the video. The following danmaku continue focusing on this topic, referring to both the video content and surrounding comments. Secondly, danmaku present a group chatting scenario, which contains various interaction forms among viewers. For example, in part (a) viewers discuss on the same topic (in blue), while in part (b) viewers post Q&A type of comments referring to previous comments (in green). With such rich and diverse interactive multimodal information, danmaku are valuable for various research tasks. For example, with viewers' detailed comments pointing to specific video times, it can help the fine-grained video semantic comprehension. Rich opinion information in the danmaku along with specific temporal region can support fine-grained sentiment analysis etc..

High-quality danmaku dataset is therefore desired to support related research. However, this area has not been widely explored. The recent time-sync comments (TSC) dataset [33] aiming to improving user experience with danmaku information only focuses on specific virtual animation area and has very low comment density. The Livebot dataset [21] is the first public available danmaku dataset with multiple video categories. Though large in scale, Livebot has some limitations: 1) low comment density, which is not enough to present the rich interaction features of danmaku. 2) limited video types, which fails to reflect the diverse video categories in real applications.

To overcome such data limitation, we build a new danmaku dataset called 'VideoIC'[2] with higher comment density and diverse video types. To reflect the various types of videos in the real world, we collect a topic list according to the video categories of Bilibili and organize them into 6 board categories, including: *Games&Sports*, *Entertainment*, *Movie&TV*, *Science&Education*, *Daily Life* and *Culture&Art*. We then form a topic list for each category according to its popular tags and crawl relevant videos based on these topics. We finally collect 4951 videos spanning 557 hours. Along with videos, there are 5,330,393 live comments in total, with 1077 comments for each video on average. All in all, VideoIC is large in scale, with high

*Corresponding Author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413890

[1] https://www.bilibili.com/
[2] https://github.com/AIM3-RUC/VideoIC

(a) 00:31 (b) 05:33 (c) 07:10

Figure 1: An live video commenting example in VideoIC (best viewed in color).

comment density and diverse video categories, making it suitable for various research tasks related to the new media data danmaku.

Automatic video comment generation aims to generate human-like live comments, which can advance our understanding of the multimodal interactive features of danmaku. One of the reasons why danmaku is popular among young people is that it can create an atmosphere where many people watch and discuss together to eliminate loneliness. Therefore, automatic comments generation technique may help enrich live video comments and attract more viewers. The authors in Ma et al. [21] first propose such task and the Unified Transformer model. However, the model suffers some limitations. Firstly, it only utilizes limited surrounding comments, failing to track the relation among multiple comments in such a group chatting scenario. Secondly, it ignores the temporal relation across multiple modalities.

In this work, we propose a multimodal multi-task learning based comments generation framework (MML-CG), which can capture the interaction and temporal relation across multiple modalities. We conduct extensive experiments on both Livebot and VideoIC. The experimental results prove the effectiveness of MML-CG, which outperforms the state-of-the-art model with smaller model size. The experiment analysis also reveals the multimodal interactive features of danmaku.

The main contributions of this work are as follows: (1) We introduce a new large scale danmaku dataset named 'VideoIC' with high comment density and diverse video categories, which can better support related research. (2) We propose a new live video comments generation framework with multimodal multitask learning, which integrates multiple modalities to achieve effective comments generation and temporal relation prediction. (3) We conduct extensive experiments on the previous public dataset and our new dataset to verify the effectiveness of proposed model and reveal the multimodal interactive features of danmaku.

2 VIDEOIC DATASET

We build a new danmaku dataset named "VideoIC" with diverse video types and dense live comments, which can support various research tasks corresponding to the new media data. In this section, we first introduce the data collection procedure and then present the properties of the new dataset.

2.1 Data Collection

Bilibili is a popular Chinese video streaming website which features the real-time commenting function. It has become a leading culture social platform in China especially among young people. We therefore choose to build our danmaku dataset from Bilibili.

We first collect the video types on the Bilibili website and organize them into six categories: *Games&Sports*, *Entertainment*, *Movie&TV*, *Science&Education*, *Daily Life* and *Culture&Art*. We then build the topic list for each category based on its popular tags. Next, we search for videos using these topics as queries on the Bilibili website and rank videos based on the number of live comments. We download the top 1000 available videos for each category according to the rank, along with all their live comments and other related meta information such as titles and tags. After filtering out low quality and incomplete data, we finally collect 4951 videos spanning 557 hours with 5,330,393 live comments.

2.2 Data Statistics

Figure 2 shows the data distribution over the 6 video categories and the data division of our dataset. The data distribution across different video types is relatively balanced. We divide the VideoIC dataset into training, validation, and test sets according to the video duration, the number of live comments and video categories. In summary, we have 3451 videos for the training set, 500 videos for the validation set and 1000 videos for the test set.

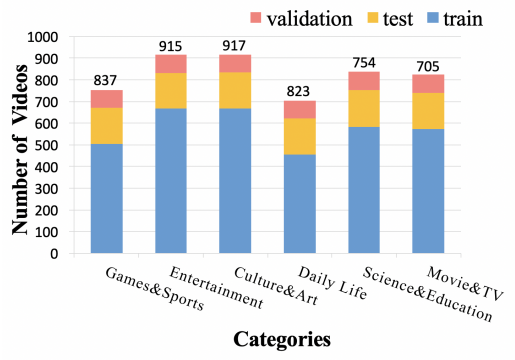

Figure 2: Statistics of video categories and dataset split.

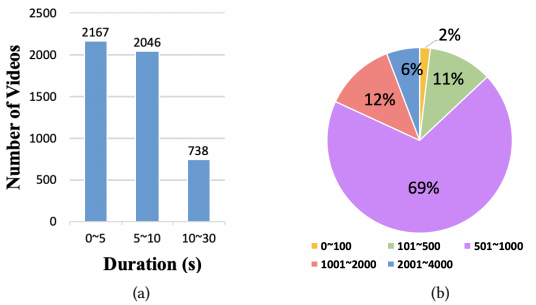

Figure 3: (a): Distribution of the video duration. (b): Distribution of the number of comments.

Figure 3 (a) shows the distribution of video duration. We divide the videos into three groups according to their lengths: short videos with duration less than 5 minutes, mid-length videos with duration between 5 to 10 minutes, long videos with duration more than 10 minutes. The ratio of the three groups is roughly 3:3:1. Figure 3 (b) shows the distribution of the number of live comments per video. As we can see, about 98% of videos have more than 100 comments and nearly 87% of videos contain more than 500 comments. In total, there are more than 5 million comments in VideoIC dataset with 1077 comments on average per video, which indicates the dense comment feature of our VideoIC dataset.

Table 1: Comments density comparison between public available danmaku datasets. Avg.C/s=Average comments per second.

Dataset	Domain	Duration(h)	Comments	Avg.C/s
TSC [33]	Animation	47,835	32,949,297	0.19
Livebot [21]	General	114	895,929	2.18
VideoIC	General	557	5,330,393	2.66

Table 1 presents the comments density comparison between danmaku datasets TSC, Livebot and our VideoIC, all of them are collected from Bilibili website. TSC [33] is domain specific and has very low comments density, which is not suitable for danmaku interaction research. Table 2 presents the detailed comparison between the broad category datasets VideoIC and Livebot. VideoIC is much larger than Livebot with respect to the number of videos (2 times larger), the total duration of videos (5 times larger) and the number of comments (6 times larger). The average number of words and characters per comment of the two datasets are similar, indicating that comments are normally short sentences on Bilibili. Our VideoIC dataset has higher comment density than Livebot, which can be demonstrated from a few aspects: the average number of comments at each time stamp in VideoIC is higher (2.7 vs. 2.2); percentage of video stream with comments in VideoIC is higher (78% vs. 61%); the percentage of videos with more than 500 comments in VideoIC is higher (87% vs 36%). Higher comments density implies more complex interactivity and dependency in danmaku, the new anonymous group chatting scenario. Table 3 presents the comparison of video categories in the two datasets. VideoIC has broader topic coverage than Livebot, indicating greater video diversity in this new dataset, which can better present the data distribution of videos in the realistic world.

In summary, the high comments density and diverse video types properties make VideoIC a challenging corpus for various investigations related to the new media data danmaku, such as automatic

Table 2: Comparison of VideoIC and Livebot datasets. % of time with comments = Total video time stamps with comments / Total video duration.

Dataset	Livebot	VideoIC
Videos	2,361	4,951
Duration (in hour)	114	557
Comments	895,929	5,330,393
Avg Character	5.42	5.39
Avg Word	9.08	9.14
Avg Comments/per Video	380	1,077
Duration w/ Comments (in second)	252,382	1,558,357
% of time w/ comments	61%	78%
% of videos w/ comments >500	36%	87%

Table 3: Topic coverage comparison between VideoIC and Livebot. G&S: Games&Sports, M&T: Movie&TV, S&E: Science&Education, C&A: Culture&Art, DL: Daily Life, Ettm: Entertainment.

Category	Livebot	VideoIC
G&S	basketball, football, fitness, NBA	basketball, football, diving, swimming, fitness, electronic sports, athletics
M&T	anime	anime, TV play, film editing, cartoon
S&E		adventure, online course, speech, nature, humanity
C&A	dance, sing, piano, drawing, origami	dance, drawing, music, custom, Chinese style
DL	pets, food, vehicle, beauty	pets, food, travel, humor, living style
Ettm	magic show	fashion, star, TV show

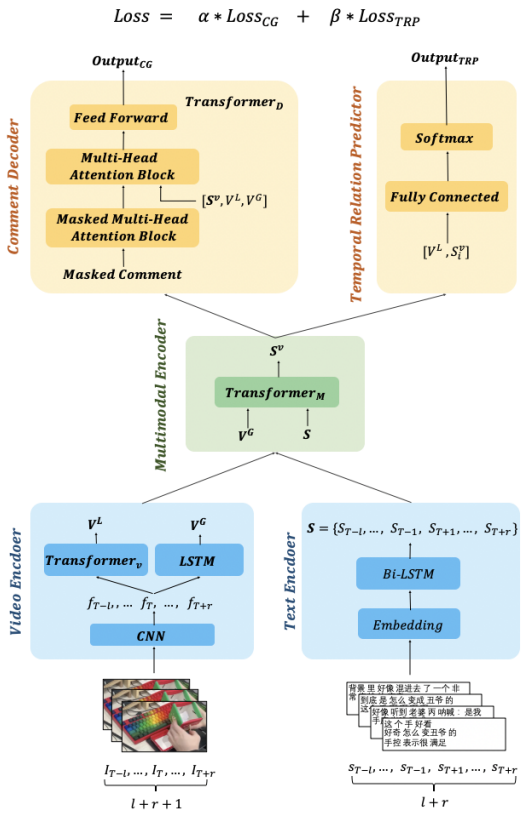

Figure 4: Multimodal Multi-task Learning based Comment Generation Framework (MML-CG), which aims to optimize the comment generation at time stamp T based on the video context $\{I_{T-l}, ..., I_{T+r}\}$ and comment context $\{s_{T-l}, ..., s_{T-1}, s_{T+1}, ..., s_{T+r}\}$.

comments generation, video classification, video highlight prediction, video recommendation etc..

3 MULTIMODAL MULTITASK LEARNING BASED COMMENTS GENERATION

Viewers post comments based on the surrounding context, including both the past and future context. The goal of live video comment generation is to generate human-like comments c at time stamp T in a given video based on the surrounding context within temporal range $[T-l, T+r]$, where l and r are the left and right context size. The surrounding context includes the video context $I_t, t \in [T-l, T+r]$ and comments context $s_t = \{s_t^1, s_t^2, ..., s_t^{n_t}\}$, $t \in [T-l, T+r]$ and $t \neq T$, as there might be multiple n_t comments at each time stamp t. We propose a Multimodal Multitask Learning framework for Comments Generation (MML-CG). The overall structure of the framework is illustrated in Figure 4. Different modality features are first extracted by different modality encoders. They are then integrated by the multimodal encoder, based on which two tasks, temporal relation prediction and comment generation, are jointly optimized in the end-to-end manner.

3.1 Multimodal Encoding

Video Encoder: The input to the video encoder includes the video frame I_T at time stamp T and its neighbor frames in the left context $I_{T-l}, ..., I_{T-1}$ and right context $I_{T+1}, ..., I_{T+r}$. As shown in Figure 4, two types of encoding representations are generated from the video encoder: the global representation V^G encodes contents of the video clip and the local representation V^L focuses on the time stamp T. Frame level feature f_t is first extracted via a pre-trained convolution neural network (CNN) [15] for each input frame I_t. Long short-term memory (LSTM) network [12] is employed to encode the semantic and temporal information of the context video clip $h_i = LSTM(f_i, h_{i-1})$ and the final hidden state is used as the global representation V^G. We also employ a transformer to encode the video information focusing on the time stamp T:

$$V^L = Transformer_v(f_T, \boldsymbol{f}, \boldsymbol{f}) \tag{1}$$

where f_T is the query, $\boldsymbol{f} = \{f_{T-l}, ..., f_{T+r}\}$ is the key and value of the $Transformer_v$. The multi-head attention block can capture the relation among frames.

Text Encoder: The input to the text encoder includes multiple comments from surrounding time stamps of T. We encode each comment s into its sentence representation S. As there might be multiple comments at certain time stamp from different viewers, each comment is encoded separately. Take the j-th comment $s_{t,j} = \{w_{t,j}^1, ..., w_{t,j}^{n_{t,j}}\}$ as an example, which appears at time t and contains $n_{t,j}$ words in total. After embedding each word through an embedding matrix W_e, we employ a bidirectional $LSTM$ to generate the comment representation:

$$e_{t,j}^k = W_e w_{t,j}^k \tag{2}$$

$$\overrightarrow{h_k} = \overrightarrow{LSTM}(e_{t,j}^k, h_{k-1}) \tag{3}$$

$$\overleftarrow{h_k} = \overleftarrow{LSTM}(e_{t,j}^k, h_{k+1}), \quad k \in [1, n_{t,j}] \tag{4}$$

We concatenate the last hidden state in two directions $h = [\overrightarrow{h}_{n_{t,j}}, \overleftarrow{h}_1]$ as the sentence representation $S_{t,j}$ for each comment $s_{t,j}$. Therefore, we get a set of sentence representation $\boldsymbol{S_t} = \{S_{t,1}, ..., S_{t,N_t}\}$ for all the comments appearing at time stamp t.

Multimodal Encoder: The multimodal encoder integrates and encodes the cross-modal temporal and semantic relation. The input to the multimodal encoder includes the set of sentence representations of comments $\boldsymbol{S}$ from the comment encoder and the global video representation V^G from the video encoder. We employ a transformer to encode the relation between video content and comments at each time stamp. Take comments at time stamp t as an example:

$$S_t^v = Transformer_M(V^G, \boldsymbol{S_t}, \boldsymbol{S_t}) \tag{5}$$

where V^G is the query and $\boldsymbol{S_t}$ is the key and value of the transformer. Through the multimodal encoder, we generate one video-aware comment feature vector S_t^v to encode all the comments that appear at the time stamp t.

3.2 Multitask Learning

The semantic and temporal relation within and across different modalities is important for comments generation. In order to enhance the encoding of such information, we introduce another temporal relation prediction task, which forces the model to emphasize the relation between comments and video content. Through multitasking learning strategy, the comment generation can be improved correspondingly.

Temporal Relation Prediction: The goal of the temporal relation prediction task is to predict the relation between the time stamp t and the target time stamp T based on the multimodal context: whether t is in the left context $[T-l, T-1]$ or the right context $[T+1, T+r]$ of the target time stamp T. We treat it as binary classification:

$$\hat{y}_t = Softmax(FC([V^L; S_t^v])) \tag{6}$$

The input feature to the relation predictor is the concatenation of the video local representation V^L of the target time stamp T from the video encoder and the video-aware comments representation S_t^v for the time stamp t from the multimodal encoder. The objective function is the cross entropy loss:

$$Loss_{TRP} = \frac{1}{l+r} \sum_t y_t log(\hat{y}_t) \tag{7}$$

where y_t is the ground truth label, which is 1 if $t < T$ in the left context, otherwise is 0 if $t > T$ in the right context.

The temporal relation prediction task can force the video-aware comments representation S_t^v to encode the semantic information related to video content and the temporal relation across different modalities. Therefore, the temporal relation prediction task can lead to better contextual representation for the generation task.

Comment Generation: The goal of the comment generation task is to generate comments for the target time stamp T based on its context. We consider multimodal context including both the comment context and the visual context. To be specific, for the target time stamp T, we have the video-aware comment contexts representations $\boldsymbol{S^v} = \{S_{T-l}^v, ..., S_{T-1}^v, S_{T+1}^v, ..., S_{T+r}^v\}$ from its left neighbor context $[T-l, T-1]$ and right neighbor context $[T+1, T+r]$, the global and local visual context V^G and V^L. We employ a transformer as the comment decoder. The probability of generating a comment sentence $c = \{w_1, ..., w_N\}$ based on the context is:

$$\begin{aligned} p_\theta(c|\boldsymbol{S^v}, V^G, V^L) = &\quad p_\theta(w_1, \ldots, w_N|\boldsymbol{S^v}, V^G, V^L) \\ = &\quad \prod_{j=1}^{N} p_\theta(w_j|w_{<j}, \boldsymbol{S^v}, V^G, V^L) \end{aligned} \tag{8}$$

$$p_\theta(w_j|w_{<j}, \boldsymbol{S^v}, V^G, V^L) = \text{Softmax}(Wy_j) \tag{9}$$

$$y_j = Transformer_C(w_j, \boldsymbol{c_j^*}, [\boldsymbol{S^v}, V^G, V^L]) \tag{10}$$

where $\boldsymbol{c_j^*}$ is the masked comment with future tokens $w_k, k > j$ masked. The transformer contains several multi-head attention blocks, which first attends to the masked comment input $\boldsymbol{c_j^*}$ and then attends to the multimodal context $[\boldsymbol{S^v}, V^G, V^L]$. We employ cross-entropy loss for comment generation in the training stage:

$$Loss_{CG} = -\sum_{c \in \mathcal{D}^{train}} \log p_\theta(c|\boldsymbol{S^v}, V^{global}, V^{local}) \tag{11}$$

Multitask Training: The temporal relation prediction and comment generation tasks are jointly trained in the end-to-end manner:

$$Loss = \boldsymbol{\alpha} Loss_{CG} + \boldsymbol{\beta} Loss_{TRP} \tag{12}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyper parameters.

Table 4: Data split for live comments generation task.

Dataset	Train	Test	Validation
Livebot	672,329	10,620	44,457
VideoIC	3,063,031	149,021	152,392

4 EXPERIMENT

We conduct experiments on two danmaku datasets: Livebot [21] and VideoIC. We form the comment generation task as text generation in the training stage and evaluate it as a ranking problem on the test set. In this section, we first introduce the experiment setup and then present the experiment results and analysis under different settings.

4.1 Experimental Setup

We pre-process the data from three modalities:

Visual: we sample one frame per second from the video and resize each frame to 224×224. Then, we extract a 2048 dimensional feature vector for each frame using a pre-trained ResNet-101 model [11].

Audio: we first extract audio stream from each video and segment it into a sequence of 1-second clips. Then we extract a 48 dimensional feature vector including mfcc, zero-crossing-rate etc. for each clip using the toolkit librosa[3] [22].

Text: we segment each comment sentence into a sequence of words using the word segmentation toolkit *jieba*[4].

A vocabulary with top 30000 most frequent words from the training set is built separately for Livebot and VideoIC datasets. We use the information within past 5 seconds and future 5 seconds as the context($l = 5, r = 5$). As there are multiple comments from different viewers at each time stamp, in the experiments, we sample at most 15 comments at each time stamp. We set the maximum length of comments as 15. We choose the hyper parameters in the joint loss function as $\alpha = 0.7$ and $\beta = 0.3$ on the validation set of Livebot and employ them on both datasets. The data split setup for each dataset is shown in Table 4.

Evaluation Setup: In the danmaku datasets, there are multiple comments from different viewers at each time stamp, which also means that proper comments at certain time stamp can be very diverse. Reference-based metrics such as BLEU [25], Meteor [8], CIDEr [28] and ROUGE [9] may not be very suitable to evaluate such diverse results. Therefore, we form the evaluation as a ranking problem similar to [21] in the test stage, which requires the model

[3] http://librosa.github.io
[4] https://pypi.org/project/jieba/

to rank a list of comments according to the their generation probabilities. We provide a list of candidate comments which consists of diverse ground-truth comments (posted by viewers at this time stamp) and some improper comments. The model's ability to rank the groundtruth comments higher (with higher generation probability) than other candidates (with lower generation probability) can reflect its ability in generating proper comments.

We thus need to construct the candidate comments list for models to rank at each target time stamp of the test video. We first pool all comments in the training set excluding the ground-truth comments for the target time stamp together as the candidate comments set Z. Then we sample improper comments in three ways from Z:

- **Plausible comments**: 20 comments most similar (cosine distance between the TF-IDF vectors) to the video titles in the candidate comments set Z.
- **Popular comments**: 20 comments randomly picked from candidate comments with the highest frequency (appears more than 100 times) in the candidate comments set Z. For example: 'What is the bgm (background music)'.
- **Random comments**: 100 comments randomly picked from the candidate comments set Z.

In this way, models are required to rank the ground-truth comments and the 140 sampled improper comments. The higher the ground-truth comments are ranked, the better the model performance.

We measure the ranking results with three types of metrics as in Das et al. [6]:

- **Recall@k**: the proportion of ground-truth comments in the top-k sorted comments, higher is better;
- **Mean Rank**: the mean rank of all ground-truth comments, lower is better;
- **Mean Reciprocal Rank**: the mean reciprocal rank of the ground-truth comments, higher is better.

4.2 Experiment Results

We conduct experiments with different setups on both Livebot and VideoIC datasets to investigate the influence of different factors on the generation performance.

4.2.1 Comparison with State-Of-The-Arts

We first compare our model with the state-of-the-art live comment generation models, including:

- **Fusional RNN**[21]: it utilizes LSTMs with attention mechanism as the encoder and decoder.
- **Unified Transformer**[21]: it employs transformers with multiple blocks in both encoder and decoder modules.

Table 5 shows the comparison results. Our proposed MML-CG model achieves the best performance on both two datasets. It is worth noting that the model size of MML-CG is smaller than that of the Unified Transformer model, which indicates that the performance improvement should come from the multimodal encoding and multitask learning instead of more model parameters.

We also conduct human evaluation on the comment generation results. The user is asked to evaluate the quality of generated comments from three aspects:

- **Fluency**: fluency of comments.
- **Relevancy**: relevancy between comments and video content.
- **Correctness**: the confidence that the comments are made by humans in the context of the video content, which can be considered as a comprehensive metric to evaluate the quality of live comments.

Table 5: Comparison between MML-CG, Fusional RNN and Unified Transformer utilizing visual and text information (V&C). R@k: Recall@k, MR: Mean Rank, MRR: Mean Reciprocal Rank.

Livebot

Model	R@1	R@5	R@10	MR	MRR
Fusional RNN	11.96	32.54	41.26	18.80	0.231
Unified Transformer	14.94	40.19	50.39	16.30	0.278
MML-CG(Ours)	**18.57**	**43.80**	**54.09**	**15.37**	**0.312**

VideoIC

Model	R@1	R@5	R@10	MR	MRR
Fusional RNN	22.32	48.03	57.11	14.70	0.343
Unified Transformer	26.34	54.66	64.37	12.66	0.390
MML-CG(Ours)	**27.50**	**56.12**	**65.68**	**12.21**	**0.402**

Table 6: Human evaluation result on the VideoIC test set.

Model	Fluency	Relevancy	Correctness
Human	4.92	4.24	4.80
Fusional RNN	4.26	2.80	3.13
Unified Transformer	4.18	3.49	3.93
MML-CG (Ours)	4.44	3.84	4.15

We hire 30 senior danmaku users to rate each comment with a score in the range of $[0, 5]$. The higher the score, the better the comment quality. For each model, we assign 25 generated comments from different time stamps to each user and take the average score as the his/her evaluation result. Therefore, each user will evaluate 100 comments from three models and ground-truth for each metric. Finally we take the average score from all users as the human evaluation result. As we can see from Table 6, our proposed MML-CG model achieves better scores on all three aspects. Some examples of generated comments with the proposed MML-CG model are shown in Figure 5, where the generated live comments (in purple) interact the surrounding comments (in black) and the video content. Although the proposed MML-CG model outperforms other state-of-the-art models on all three human evaluation metrics, from the results we can see that there still exists a clear gap with human. MML-CG may generate relevant but not appropriate comments. As shown in the third example in Figure 5, MML-CG grasps the topic 'profession', but generates an improper comment as shown in green.

4.2.2 Multimodal Information Analysis

We also investigate the multimodal information impact on live comments generation as shown in Table 7, where V refers to the visual representation, C refers to the surrounding comments representation, and *V&A* refers to the concatenation of audio and visual representation. We employ transformer based seq2seq model for

Figure 5: Examples of generated comments by MML-CG on VideoIC test set. Purple: good comments. Green: improper comments. Black: human posted comments. (best viewed in color).

Table 7: Comparison of comment generation quality based on different modalities with Seq2Seq Transformer. V:visual information, C:surrounding comments, A:audio

Livebot

Modality	R@1	R@5	R@10	MR	MRR
V	7.63	26.28	32.17	22.42	0.179
V&A	7.13	24.72	35.09	19.61	0.177
C	14.46	39.66	49.66	16.60	0.272

VideoIC

Modality	R@1	R@5	R@10	MR	MRR
V	7.03	31.50	40.25	21.26	0.189
V&A	7.08	32.93	42.28	20.25	0.194
C	25.76	53.07	63.82	12.91	0.391

this set of experiments, which can be considered as the single modality version of the Unified Transformer [21]. From the results we can see that: (1) Both visual information and surrounding comments information are helpful for comment generation. However, audio modality has little influence on the performance. In the experiment, we only utilize acoustic information from the audio and we suspect that the speech content might be more helpful. (2) Surrounding comments information boosts the performance most, indicating that danmaku contains rich viewer interactions.

Table 8: Performance under different generation scenario on VideoIC dataset. Modality: Visual and Surrounding comments(V&C). UT: Unified Transformer.

Model	Rate	R@1	R@5	R@10	MR	MRR
UT	20%	19.49	49.16	58.83	14.62	0.328
	50%	23.84	53.20	62.98	13.20	0.369
	80%	25.83	54.24	63.96	12.83	0.385
	100%	26.34	54.66	64.37	12.66	0.390
MML-CG	20%	20.00	49.65	59.47	14.44	0.333
	50%	24.69	53.70	63.53	13.02	0.376
	80%	26.46	55.24	64.95	12.49	0.392
	100%	27.50	56.12	65.68	12.21	0.402

Comparing results in Table 5 and Table 7, Seq2Seq Transformer only using surrounding comments achieves comparable results with the Unified Transformer and achieves even better performance than the Fusional RNN which utilizes multimodal information. Our proposed MML-CG outperforms the single modality models by a large margin. It proves that the proposed MML-CG model can effectively handle the cross-modality interaction.

4.2.3 Generation Scenario Analysis

In real applications, we want to train models with high density comments and apply it in the low comments density scenario (as we mentioned before, the live comment generation can stimulate the group chatting atmosphere). Therefore, we compare the performance of Unified Transformer and MML-CG model under different test scenarios with different surrounding comments density on our VideoIC dataset. We reduce the surrounding comments density of the test set by randomly selecting comments according to a certain rate. Then we evaluate models trained with full training set on these low comment density test data. The results are shown in Figure 8. As we expected, the performance drops with the decrease of comment density. Our model still outperforms the Unified Transformer under low comment density testing scenario.

4.2.4 Ablations Study of Model Components

We conduct experiments to investigate the performance changes with model structure variants. We first explore the influence of the decoder context. In the full model, we generate the comments based on multimodal information $[\boldsymbol{S^v}, V^G, V^L]$. As shown in the decoder block in Table 9, both the video-aware comments representation $\boldsymbol{S^v}$, the global V^G and local V^L video representation contribute to the performance. The global representation brings slightly more improvements than local representation as V^G contains more visual context information than V^L.

Then we explore the cross modalities attention mechanism to generate video-aware comments representation $\boldsymbol{S^v}$ in the multimodal encoder. The results in the attention block of Table 9 indicates

Table 9: Ablation Study. Decoder: different combinations of multi-modal information in the decoder. Attention: different ways to employ attention mechanism in the multi-modal encoder. Task: different task combination

Module	Variant	Livebot					VideoIC				
		R@1	R@5	R@10	MR	MRR	R@1	R@5	R@10	MR	MRR
Decoder	S^v	15.76	39.44	49.94	6.49	0.280	25.83	54.42	63.9	12.95	0.386
	$S^v + V^L$	16.46	40.93	50.92	16.11	0.290	25.75	55.03	64.68	12.62	0.387
	$S^v + V^G$	16.86	41.40	51.31	15.96	0.293	26.37	55.03	64.52	12.71	0.391
	$S^v + V^G + V^L$	**18.57**	**43.80**	**54.09**	**15.37**	**0.312**	**27.50**	**56.12**	**65.68**	**12.21**	**0.402**
Attention	$V^L \rightarrow S^v$	16.49	41.40	50.98	16.15	0.290	26.49	55.60	65.26	12.37	0.394
	$V^G \rightarrow S^v$	**18.57**	**43.80**	**54.09**	**15.37**	**0.312**	**27.50**	**56.12**	**65.68**	**12.21**	**0.402**
Task	CG only	16.52	41.88	51.95	15.91	0.291	26.01	55.19	65.09	12.46	0.390
	TRP + CG	**18.57**	**43.80**	**54.09**	**15.37**	**0.312**	**27.50**	**56.12**	**65.68**	**12.21**	**0.402**

that global information of video clips can lead to better comment representation than local video information of time stamp t, which may due to the fact that the temporal relation between live comments and video content is loose so the global representation containing information of all surrounding video frames is beneficial.

We also evaluate the improvement bring by the multi-task learning. As shown in the Task block of Table 9, the Temporal Relation Prediction task can help model the complex multimodal interaction required for comments generation.

5 RELATED WORK

Live video comments which provides real-time multimodal interactions of viewers can support various related research. Apart from viewer behavior [18, 31], most previous works focus on extracting user preferences information from danmaku to help recommendation, such as key frame prediction [4], video tagging [1, 33] and recommendation [20, 32]. Among various tasks, live video comment generation [21] aims to generate human-like live comments, which is suitable to explore the unique interaction characteristic of danmaku. Different from video caption generation [3, 14] which describes the video content, and Visual Dialog [6] which creates meaningful conversation conditioned on an image, live video comments generation requires generating interactive comments in a free group chatting scenario based on the surrounding context.

TSC [33] is a large scale dataset on specific virtual animation area to support the exploration of user experience improvement. Although containing 32 million comments, TSC has a low comments density, making it unsuitable to support the research for interaction within live comments. Livebot [21] is a danmaku dataset with videos from broad categories. However, low comments density makes it hard to present the interactivity of live comments. Besides, videos are collected by popular queries of the search engine rather than representative topics of the online video website. Comparing to the previous datasets, our VideoIC dataset has higher comment density, broader and more representative video categories. Different from other multimodal video datasets such as VQA dataset [16, 17] and instructional video dataset [7, 27, 30, 36], VideoIC contains complex multimodal and user interactions.

The Encoder-Decoder framework originally introduced for machine translation [5, 26] is widely used in the generation task [23, 29]. The first study for live video comment generation is [21]. It proposes the state-of-the-art model named Unified Transformer. However, it concatenates surrounding comments into a sequence, ignoring the order of appearance and the temporal relation with aligned video content. Therefore, the Unified Transformer model can not fully capture the complex dependency and temporal relation of multimodal information in the live commenting scenario.

Another related research area is the multi-modality interaction and fusion. Various of attention mechanisms can adaptively select important information to achieve effective information flow. Co-attention [2, 19, 24, 34] and bilinear attention [13, 35] methods learn the inter-modality relations between modalities. [10] utilizes self-attention and cross-modal attention to fuse multimodal information. However, there exists a loose relation between video and live comments instead of strict correspondence. Therefore how to model such loose relation is important for live comments generation.

6 CONCLUSION

As an emerging social media, danmaku has attracted more and more users, which produces large amount of valuable interactive multimodal data involving natural language texts along with videos. In this work, we propose a new danmaku dataset named 'VideoIC', with 557 hours of videos and 5 million live comments. The high comments density and diverse video types make it a challenging corpus for various investigations related to danmaku. We explore the live video comments generation task based on this dataset. We propose a multimodal multitasking learning based model which can grasp the temporal relation and interaction between multiple modalities for comments generation. The extensive experiments on both previous Livebot and the new VideoIC datasets prove the effectiveness of our proposed model. The experiment results also demonstrate the importance of utilizing multimodal data and their interaction information. In the future work, we will explore how to assist more tasks with the information in danmaku.

7 ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 61772535) and Beijing Natural Science Foundation (No. 4192028).

REFERENCES

[1] Q. Bai, Q. V. Hu, L. Ge, and L. He. 2019. Stories That Big Danmaku Data Can Tell as a New Media. *IEEE Access* 7 (2019), 53509–53519.

[2] Victor Zhong Caiming Xiong and Richard Socher. 2017. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*.

[3] Shizhe Chen, Yuqing Song, Yida Zhao, Qin Jin, Zhaoyang Zeng, Bei Liu, Jianlong Fu, and Alexander Hauptmann. 2019. Activitynet 2019 Task 3: Exploring Contexts for Dense Captioning Events in Videos. *arXiv preprint arXiv:1907.05092* (2019).

[4] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized key frame recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 315–324.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2634–2641.

[8] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation* (Baltimore, Maryland, USA). Association for Computational Linguistics, 376–380. https://doi.org/10.3115/v1/W14-3348

[9] Carlos Flick. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Workshop on Text Summarization Branches Out*.

[10] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion With Intra-and Inter-Modality Attention Flow for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6639–6648.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*. 1571–1581.

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

[17] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *arXiv preprint arXiv:1904.11574* (2019).

[18] Chenchen Li, Jialin Wang, Hongwei Wang, Miao Zhao, Wenjie Li, and Xiaotie Deng. 2019. Visual-Texual Emotion Analysis with Deep Coupled Video and Danmu Neural Networks. *IEEE Transactions on Multimedia* 22, 6 (2019), 1634–1646.

[19] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*. 289–297.

[20] Guangyi Lv, Kun Zhang, Le Wu, Enhong Chen, Tong Xu, Qi Liu, and Weidong He. 2019. Understanding the Users and Videos by Mining a Novel Danmu Dataset. *IEEE Transactions on Big Data* (2019).

[21] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. In *AAAI 2019*.

[22] Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, WZY, Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stöter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis "Fjord" Hawthorne, CJ Carr, João Felipe Santos, JackieWu, Erik, and Adrian Holovaty. 2018. *librosa/librosa: 0.6.2*. https://doi.org/10.5281/zenodo.1342708

[23] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined Dense Video Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6588–6597.

[24] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6087–6096.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[26] I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS* (2014).

[27] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[30] Weiying Wang, Yongcheng Wang, Shizhe Chen, and Qin Jin. 2019. YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[31] Xiaodong Wang, Ye Tian, Rongheng Lan, Wen Yang, and Xinming Zhang. 2018. Beyond the watching: Understanding viewer interactions in crowdsourced live video broadcasting services. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 11 (2018), 3454–3468.

[32] Wenmian Yang, Wenyuan Gao, Xiaojie Zhou, Weijia Jia, Shaohua Zhang, and Yutao Luo. 2019. Herding Effect Based Attention for Personalized Time-Sync Video Recommendation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 454–459.

[33] Wenmain Yang, Kun Wang, Na Ruan, Wenyuan Gao, Weijia Jia, Wei Zhao, Nan Liu, and Yunyong Zhang. 2019. Time-Sync Video Tag Extraction Using Semantic Association Graph. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13, 4 (2019), 1–24.

[34] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6281–6290.

[35] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 1821–1830.

[36] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*. 7590–7598. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344