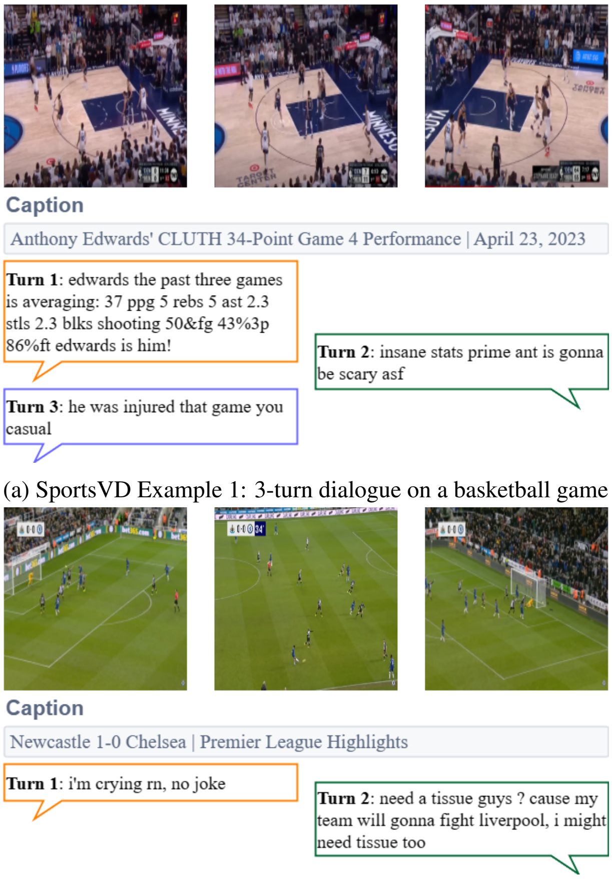
# Event-Content-Oriented Dialogue Generation in Short Video

**Fenghua Cheng, Xue Li, Zi Huang, Jinxiang Wang, Sen Wang**
School of Electrical Engineering and Computer Science,
The University of Queensland, Brisbane, Australia
{fenghua.cheng, xue.li, helen.huang, jixiang.wang, sen.wang}@uq.edu.au

## Abstract

Understanding complex events from different modalities, associating to external knowledge and generating response in a clear point of view are still unexplored in today's multi-modal dialogue research. The great challenges include 1) lack of event-based multi-modal dialogue dataset; 2) understanding of complex events and 3) heterogeneity gap between different modalities. To overcome these challenges, we firstly introduce a novel event-oriented video-dialogue dataset called SportsVD (Sports-domain Video-dialogue Dataset). To our best knowledge, SportsVD is the first dataset that consists of complex events videos and opinion-based conversations with regards to contents in these events. Meanwhile, we present multi-modal dialogue generation method VCD (Video Commentary Dialogue) to generate human-like response according to event contents in the video and related external knowledge. In contrast to previous video-based dialogue generation, we focus on opinion-based response and the understanding of longer and more complex event contents. We evaluate VCD's performance on SportsVD and other baselines under several automatic metrics. Experiments demonstrate VCD can outperform among other state-of-the-art baselines. Our work is available at https://github.com/Cheng-Fenghua/SportsVD.


Caption
Anthony Edwards' CLUTH 34-Point Game 4 Performance | April 23, 2023
**Turn 1**: edwards the past three games is averaging: 37 ppg 5 rebs 5 ast 2.3 stls 2.3 blks shooting 50&fg 43%3p 86%ft edwards is him!
**Turn 2**: insane stats prime ant is gonna be scary asf
**Turn 3**: he was injured that game you casual

(a) SportsVD Example 1: 3-turn dialogue on a basketball game

Caption
Newcastle 1-0 Chelsea | Premier League Highlights
**Turn 1**: i'm crying rn, no joke
**Turn 2**: need a tissue guys ? cause my team will gonna fight liverpool, i might need tissue too

(b) SportsVD Example 2: 2-turn dialogue on a football game

Figure 1: SportsVD Examples

## 1 Introduction

Despite great success achieved in large language models (Adiwardana et al., 2020; Lu et al., 2023; Raffel et al., 2020; Zeng et al., 2022), most of them are still limited to incapacity of reading visual information and multi-modal interactions. To build conversational agents' abilities to interact with human and specific environment by not only reading natural language but also perceiving the physical world using all senses is one of long-term goals of Artificial Intelligence (AI). Communicating in a multi-modal way can greatly enhance the engagement of user into the communication between agents.

Therefore, there has been increasing interest on multi-modal dialogue generation. Recent works have (Shuster et al., 2021) achieved a significant progress in multi-modal dialogue generation. However, there are still research gaps. First, much research interest are focused on image-grounded dialogue. Video-grounded dialogue generation still needs to be explored. State-of-the-art works on video-grounded chatbot are still unsatisfying due to difficulty in extracting information from video

*Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4114–4124
June 16-21, 2024 ©2024 Association for Computational Linguistics

and integrating external knowledge.

Second, existing video-grounded dialogue works mainly focus on entity movement but more complex event content is neglected. However, this differs largely from real-world scenarios. A longer and more complex event is more likely to be the topic of a conversation. To this end, constructing a dialogue generating model that can understand more complex scenarios and events is necessary for AI to interact with the real world. Here we give formal definition of entity movement and event contents to distinguish them easily. Event-contents based dialogue generation is still unexplored.

- **Entity Movement** Entity refers to a concrete object, *e.g.*, a person or a plane. Entity movement refers to some purposeless, simple and short movement of a single entity, *e.g.*, a person raises his arm or a plane takes off.
- **Event Contents** Event refers to a complicated activities and a series of interaction involved by multiple entities. An event which could specify participants, time and place.

Third, much attention is concentrated on question-answer pairs in multi-modal dialogue generation. Containing only question and answer in a dialogue is also inconsistent with reality as simple Q&A is largely based on facts whereas opinions are much more involved in real dialogue. Hence, opinion-based response generation still needs to be explored. Here we give a formal definition of fact-based dialogues and opinion-based dialogues:

- **Fact-based Dialogue:** should be about the existence, reality, or truth. Normally fact-based dialogue is in QA form.
- **Opinion-based Dialogue:** should state the speakers' feelings, beliefs, or views.

To this end, we propose a new dataset SportsVD which contains dialogue-video pairs on sports-domain events together with VCD, a novel multi-modal dialogue generation method based on understanding of event contents and external knowledge. SportsVD contains 5114 sports-events videos and 39K corresponding dialogues. We collect videos and multi-turn comments as dialogues from Youtube. Comments from real users based on their own knowledge help form high-quality opinion-based conversations. Figure 1 demonstrates some examples in SportsVD. We choose sports-domain events to form our dataset because we believe people are highly likely to express diverse opinions when watching a sports event. VCD utilizes video encoding from pre-trained video-language model InternVideo (Wang et al., 2022) and unified transformer structure and input representation to reduce heterogeneity gap. Additionally, we utilize prompt to integrate external knowledge into dialogue generation. We evaluate VCD and other baselines on SportsVD under both automatic metrics and human evaluation. Experiments demonstrates the efficiency and effective of our work. Our model can outperform other baselines.

In summary, the contributions of this work include:

1. We introduce a new opinion-based video-dialogue dataset SportsVD in which all video-dialogue pairs are related to sports-domain events on Youtube. To our best knowledge, SportsVD is the first video-dialogue dataset based on complex events.
2. We present a novel dialogue generation method VCD which generate response based on event contents shown in a short video and external knowledge.
3. We provide data analysis of the new dataset SportsVD and compares it with other multi-modal dataset.
4. We evaluate VCD on SportsVD and compare it to other state-of-the-art baselines.

## 2 Related Works

We conclude recent advances achieved in two related areas which are related to this work.

### 2.1 Multi-modal Dialogue System

Due to the lack of visual information perception, text-only dialogue systems can hardly produce human-like conversations in which multi-modal interaction happens. Therefore, recent works raised increasing interest in developing a multi-modal conversational agent with visual ability. According to different tasks, multi-modal dialogue systems can be roughly categorized into:

(1) Image-grounded question-answering dialogue system (Das et al., 2017; Murahari et al., 2020; Nguyen et al., 2020) aims to generate answers to given questions based on a given image. (2) Instead of conversing based on facts in

Q&A form, image-grounded chit-chat dialogue system (Shuster et al., 2020) generates opinion-based responses. (3) Visual-evidence-embedded dialogue system (Liang et al., 2021; Shen et al., 2021) takes text-only context as input and generate text-only response as well. However, non-paired image, acting as implicit evidence, is introduced in response generation. (4) Image-Response Dialogue system (Zang et al., 2021; Sun et al., 2022) needs to generate not only textual but also visual response to the given dialogue context. (5) Video-grounded question-answering dialogue system (Le et al., 2019, 2022) generates answer to given question based on video and audio. (6) Video-grounded real-time commenting task (Ma et al., 2019) aims to generate reasonable live comments as opinions of a specific video clip near a timestamp.

Different from these works, we focuses on generating opinion-based response based on the understanding of more complex events in short video.

### 2.2 Video Understanding

Unlike understanding images, understanding videos raises more challenges. A video can be regarded as continuous set of images and therefore understanding videos needs more computation. Moreover, due to dense frames in video and slow information variance, there is much redundant information in video frames.

InternVideo (Wang et al., 2022) presents a novel idea of understanding video by fusing two types of representations from two ways of self-supervised learning. Video Swin Transformer (Liu et al., 2022) adopts swin transformer (Liu et al., 2021) into video understanding inspired from its success in image understanding. VideoMAE (Tong et al., 2022) adopts masked token learning (Devlin et al., 2019) which achieved great success in NLP to pre-train a vanilla vision transformer based autoencoder.

## 3 SportsVD Dataset

To create a dataset specially for the event-content dialogue generation is an inevitable task in this work due to lack of training corpus. To this end, we present SportsVD which contains 5114 sports-event based videos and 39K opinion-based dialogues regarding to them. We desire these videos to contain at least one event and dialogues to be not question-answering but opinion-based and highly related to events in the video. In this section, we firstly introduce the data collection method and then detail the analysis of the new dataset.

### 3.1 Data Collection

We collect videos on Youtube, together with titles and captions. Each video describes a specific sports event, *e.g.*, a game highlight. We also collect comments under the video as dialogues related to the event contents.

To guarantee that all videos are event-based, we focus on videos about sports game highlights. We collect both basketball and football game highlights in world-wide famous sports league or association. Considering the uneven quality of videos posted on Youtube, we select videos carefully. We assume that a video with high number of likes and times of watching is less likely to have poor quality. Therefore, we filter retrieved videos by : 1) the length of video should be less than 4 minutes; 2) the number of comments under the video should be more than 100; 3) the number of view counts of the video should be greater than 10000.

On the other hand, we use comments and replying relationship to construct multi-turns dialogues related to the video. Similar as filtering videos, comments with high number of likes proves to be meaningful and informational replies. To filter comments which are meaningless and less relevant to the video, we select comments under such conditions: 1) comments should be in English and 2) the number of votes for the comment should be greater than 5. For selected comments, we reconstruct them to multi-turn dialogues according to the replying relationship. Besides, we clean some meaningless text like "@someone".

### 3.2 Data Analysis

We present basic statistics of SportsVD in table 1. There are in total 5114 sports-event videos and 39097 conversations regarding to them. We split the dataset into training and testing set in 8:2 ratio.

| Split | #V | #D | #S | Avg-T | Avg-L |
|---|---|---|---|---|---|
| Train | 4,065 | 30,708 | 153,161 | 2.23 | 162.7s |
| Valid | 1,049 | 8,389 | 42,299 | 2.22 | 162.1s |
| Total | 5,114 | 39,097 | 195,460 | 2.23 | 162.5s |

Table 1: Basic statistics of SportsVD. "#V", "#D", "#S" denotes the number of videos, dialogues and sentences. "Avg-T" and "Avg-L" denotes the average turns of a dialogue and the average duration of videos.

We also construct analysis on the quality of our dataset SportsVD. We aim to measure how the dialogue is related to the event contents in the video and how it is related to external knowledge. We sample 100 dialogues from SportsVD and ask volunteers to rate from 0 to 2 in these questions:

- Whether the response is related to one or more events;
- Whether the response is based on opinions;
- Whether the response is reasonable and understandable without video;
- Whether the response is reasonable and understandable without external knowledge;

We present comparison between some main multi-modal dialogue datasets and SportsVD in table 2. Further more, to compare with other datasets, we also sample 100 dialogues from OpenViDial 2.0 (Wang et al., 2021) and AVSD (Alamri et al., 2019) and conduct the same analysis. OpenViDial is a open-domain video-frame-dialogue dataset from movies and TV series. AVSD is a dialogue dataset which consists of questioning-answering pair based on entity-movement video. Figure 2 shows the comparison between three datasets.

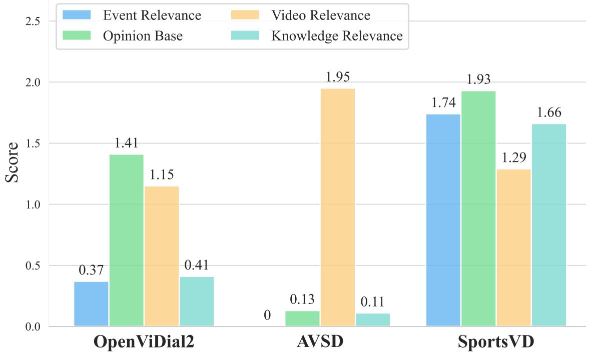


Figure 2: Human comparison between OpenViDial2, AVSD and SportsVD in four aspects.

SportsVD has the highest score for event relevance, opinion base and external knowledge relevance. The results indicate that SportsVD has the richest multi-modal information among three datasets. This is because SportsVD collects only-event videos and external knowledge is frequently associated when event-oriented dialogue happens. On the contrary, AVSD demonstrates the highest score in video relevance and the lowest score for opinion base because all dialogues in AVSD are based on facts in video in simple Q&A form.

## 4 Video Commentary Dialogue

We propose VCD (Video Commentary Dialogue), a new multi-modal dialogue generation method that understand event contents from multiple modalities in a video and generate opinion-based response. To understand longer and more complex event contents in a video, we samples several frames from the video and utilizes VideoIntern (Wang et al., 2022) to encode them. Meanwhile, in order to reduce modality gaps, unified transformer structure and input representation are applied. Besides, we integrate external knowledge from ATOMIC (Sap et al., 2019) and google knowledge graph[1] by natural language prompt.

### 4.1 Problem Formulation

We formulate video-dialogue generation task as follows. Suppose we have a Video-Dialogue set $D = \{(V_i, C_i, T_i, R_i)\}_{i=1}^{n}$ where $V_i$ is the video, $C_i$ is the caption of the video, $T_i$ is the dialogue context and $R_i$ is the response to $T_i$ based on understanding of $V_i$. The goal of VCD is to learn a generative model $P(R|V, C, T; \theta)$ from $D$ where $\theta$ means the parameters of the model. Hence, we can generate response according to Equation 1.

$$R = argmax_R P(R|V, C, T; \theta) \tag{1}$$

### 4.2 Model Architecture

In order to fill semantic gaps between different modalities, we propose to use unified transformer to integrate information from different modalities. Figure 3 shows the high-level architecture of VCD. VCD consists of: (1) a multi-layer transformer encoder and (2) a video encoder. In our experiments, we simply use pre-trained BERT (Devlin et al., 2019) as the multi-layer transformer encoder. The multi-layer transformer encodes video, caption, context and knowledge part bidirectionally and encodes response part unidirectionally. The input representation will be elaborated later.

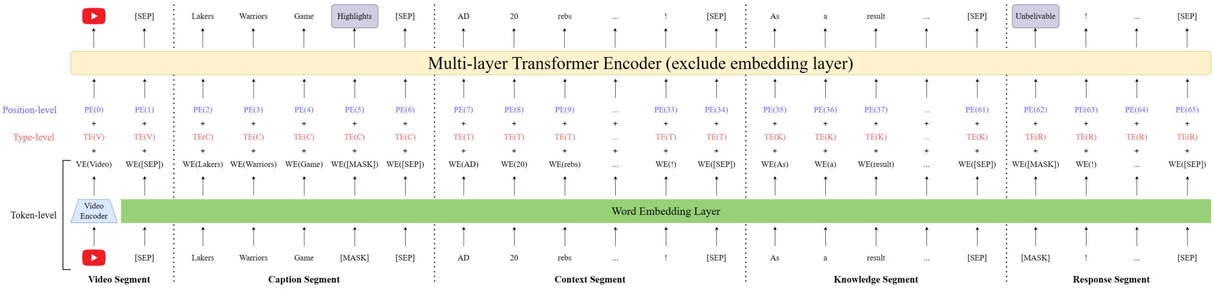
### 4.3 Video Encoder

To understand complex events in a video, we utilize frozen InternVideo (Wang et al., 2022) to produce the video embedding. The video representation is a 768-dimensional vector before fusing with language representation. We use a projection layer and layer normalization to reduce modality gap between visual and textual input.

[1] https://cloud.google.com/enterprise-knowledge-graph/docs/search-api

| Name | Type | Modalities | #Dialogues | #Images | #Videos | Language |
|---|---|---|---|---|---|---|
| VisDial (Das et al., 2017) | Fact & Entity | I, T | 1.2M | 120K | - | English |
| IGC (Mostafazadeh et al., 2017) | Opinion & Entity | I, T | 250K | 250K | - | English |
| MMChat (Zheng et al., 2022) | Opinion & Entity | I, T | 120K | 204K | - | Chinese |
| MMDialog (Feng et al., 2023) | Opinion & Entity | I, T | 1.0M | 1.5M | - | English |
| Image-Chat (Shuster et al., 2020) | Opinion & Entity | I, T | 202K | 202K | - | English |
| OpenViDial2.0 (Wang et al., 2021) | Opinion & Entity | I, T | 5.6M | 5.6M | - | English |
| PhotoChat (Zang et al., 2021) | Opinion & Entity | I, T | 12K | 12K | - | English |
| AVSD (Alamri et al., 2019) | Fact & Entity | V, T | 18K | - | 18K | English |
| VideoIC (Wang et al., 2020) | Opinion & Entity | V, T | 5.3M (comments) | - | 4.9K | Chinese |
| LiveBot (Ma et al., 2019) | Opinion & Entity | V, T | 896K (comments) | - | 2.3K | Chinese |
| TikTalk (Lin et al., 2023) | Opinion & Entity | V, T | 367K | - | 38K | Chinese |
| **SportsVD (Ours)** | Opinion & Event | V, T | 39.1K | - | 5.1K | English |

Table 2: Comparison between SportsVD and other multi-modal dataset. "I", "V" denotes "Image" and "Video".



[SEP]
Lakers
Warriors
Game
Highlights
[SEP]
AD
20
rebs
...
!
[SEP]
As
a
result
...
[SEP]
Unbelivable
!
...
[SEP]
Multi-layer Transformer Encoder (exclude embedding layer)
Position-level
PE(0) PE(1) PE(2) PE(3) PE(4) PE(5) PE(6) PE(7) PE(8) PE(9) ... PE(33) PE(34) PE(35) PE(36) PE(37) ... PE(61) PE(62) PE(63) PE(64) PE(65)
Type-level
TE(V) TE(V) TE(C) TE(C) TE(C) TE(C) TE(C) TE(T) TE(T) TE(T) ... TE(T) TE(T) TE(K) TE(K) TE(K) ... TE(K) TE(R) TE(R) TE(R) TE(R)
VE(Video) WE([SEP]) WE(Lakers) WE(Warriors) WE(Game) WE([MASK]) WE([SEP]) WE(AD) WE(20) WE(rebs) ... WE(!) WE([SEP]) WE(As) WE(a) WE(result) ... WE([SEP]) WE([MASK]) WE(!) ... WE([SEP])
Token-level
Video Encoder
Word Embedding Layer
[SEP] Lakers Warriors Game [MASK] [SEP] AD 20 rebs ... ! [SEP] As a result ... [SEP] [MASK] ! ... [SEP]
Video Segment
Caption Segment
Context Segment
Knowledge Segment
Response Segment

Figure 3: VCD Architecture and Input Representation. "+" denotes element-wise summation.

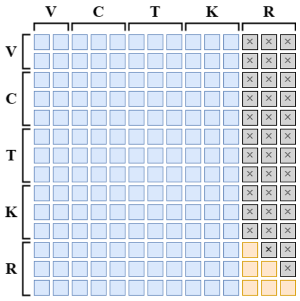


Figure 4: Self-attention mask used in VCD. Areas with cross are masked out. Unidirectional encoding for response part and bidirectional encoding for other parts.

## 4.4 Input Representation

We integrate information from video, context and external knowledge by introducing a unified input representation. There are different types and levels of input representation. To avoid confusing between types and levels of input, we specify them in Figure 3. Type indicates which segment the token is in, *e.g.*, video, context or knowledge and level indicates which level the sequence represents, *e.g.*, token level or positional level. The final input representation to the transformer network is the element-wise summation from all levels of embeddings. We explain them for details later.

### 4.4.1 Types of Input

There are five types of inputs in a sequence including video, caption, context, knowledge and response. For video input, there is only one token embedding, video embedding obtained by the video encoder. For other parts, token embeddings are all embedding of natural language word.

### 4.4.2 Levels of Input

The final input representation to the transformer is the element-wise summation of three levels of embeddings: token level, type level and position level. Each level's embeddings is a sequence of basic embeddings in dimension $[L, D]$ where $L$ indicates the length of the sequence and $D$ indicates the dimension of basic embedding. In our experiments, $L = 240$ and $D = 1024$. Hence, the final input representation is also in shape $[L, D]$.

Concatenating all five types token embeddings together with special token [SEP], we get token-level embeddings. Video embeddings are through a projection layer to keep in align with word embeddings in dimension. Token-level embeddings fuse information from different segment and modalities. Formally, the token-level embeddings can be formulated as Equation 2.

$$\begin{aligned}&TextSequence = \{w_{c1}, \dots, w_{cm}, [SEP], w_{t1}, \dots, w_{tn}, [SEP],\\ &w_{k1}, \dots, w_{kj}, [SEP], w_{r1}, \dots, w_{rl}\},\\ &Token = Cat(VE(v_1, \dots, v_j), WE(TextSequence))\end{aligned} \tag{2}$$

where $w_{xi}$ means the $i$th word in corresponding segment, $Cat$ means the concatenate operation, $VE$ means video encoder and $WE$ denotes the word embedding layer. $v_i$ is the $i$th frames sampled from the video.

Type level embeddings indicate which segment the token is in, *i.e.*, video, caption, context, knowledge or response. We simply use a embedding layer. Type level embeddings help the model to specify different segments in token-level embeddings explicitly.

Positional level embeddings indicate the position that tokens are in one sequence. We simply use the same positional embedding as original transformer.

### 4.5 Knowledge Extraction

Since SportsVD contains many external-knowledge-related dialogues, VCD tries to integrate both commonsense and domain-specific knowledge about the events and context into dialogue generation. We extracts commonsense knowledge by COMET-ATOMIC (Hwang et al., 2021). Taking description of an event as input, It predicts relationship tuple from commonsense knowledge graph ATOMIC (Sap et al., 2019), *e.g.*, <PersonX, xReact, Happy> indicates "As a result, PersonX feels happy". VCD inputs event captions and dialogue context to COMET-ATOMIC to extract commonsense knowledge.

In order to utilize domain-specific knowledge related to the event contents and context, we use a pipeline workflow. Specifically, we firstly use google ner[2] to recognize all entities, including person and location, in video captions and dialogue context. Then we extract sports knowledge about these entities by google knowledge graph and filtering "sports" keyword.

The external knowledge is represented by natural language and concatenated together with caption and context as Equation 2 shows.

### 4.6 Training Policy

For model training, we use three training objectives, MCP (masked caption prediction), MKP (masked knowledge prediction) and MRP (masked response prediction) inspired from success of (Devlin et al., 2019) respectively. In the training phase, 70% tokens in response segment are randomly masked. Among these tokens, 80% tokens are masked by a special token [MASK], 10% tokens are replaced by a random token sampled from the vocabulary and 10% tokens are unchanged. The model is required to predict these masked tokens. Similarly, to reduce semantic gap between video contents and video itself, we adopt masked caption prediction and masked knowledge prediction as another training objective. We randomly mask 15% tokens in MCP and MKP. By masking caption and knowledge tokens, the model have a better understanding of video and external knowledge.

We simply use CrossEntropy loss of the masked tokens as loss for MCP, MKP and MRP. Specifically, Equation 3 denotes the loss of these two training tasks, where $\hat{R}$, $\hat{K}$ and $\hat{C}$ denotes the set of masked tokens in response segment, knowledge segment and caption segment respectively.

$$\begin{aligned}&L_{MRP}(V,C,T,R) = -\sum_{w_{ri}\in\hat{R}} logp_i(w_{ri}|V,C,T,R),\\ &L_{MKP}(V,C,T,R) = -\sum_{w_{ki}\in\hat{K}} logp_i(w_{ki}|V,C,T,R),\\ &L_{MCP}(V,C,T,R) = -\sum_{w_{ci}\in\hat{C}} logp_i(w_{ci}|V,C,T,R),\\ &L = L_{MRP} + L_{MCP} + L_{MKP}\end{aligned} \tag{3}$$

For inference phase, the model encodes the video, caption, context, knowledge and special [BOS] token as input and generates the first token of response by predicting a [MASK] token over the vocabulary. Then the [MASK] token is replaced by the generated token and repeat predicting a new [MASK] token appending to the input sequence until ending condition satisfies.

## 5 Experiments

We conduct experiments among VCD and other state-of-the-art baselines on SportsVD under several automatic metrics. In this section, we elaborate the experiments and results.

### 5.1 Implementation Details

VCD uses InternVideo-MM-L-14[3] as video encoder and a bert-large-uncased[4] as base transformer

[2] https://cloud.google.com/natural-language/docs/analyzing-entities?hl=en

[3] https://github.com/OpenGVLab/InternVideo

[4] https://huggingface.co/bert-large-uncased

network. We only tune parameters of the base transformer network. The embedding dimension $D$ used in VCD is 1024. We set the max length $L = 240$ in which the max length of response, caption, context and knowledge is 40, 40, 80 and 80 respectively. For training details, we optimize via Adam optimizer (Kingma and Ba, 2014) and use MultiStepLR to adjust the learning rate. All training work was completed on NVIDIA Tesla V100 SMX2 GPU.

## 5.2 Evaluation Metrics

For each video-dialogue pair, we have a reference response. We call response generated from VCD and other baseline candidate. We evaluate the quality of generated response candidates by both automatic metric and human evaluation.

### 5.2.1 Automatic Metric

We employ two aspects of automatic metrics to evaluate VCD and other baselines:

- **Relevance:** we use (1) Rogue-L (Lin, 2004), (2) BLEU (Papineni et al., 2002), (3) CIDEr (Vedantam et al., 2015) and (4) Meteor (Lavie and Agarwal, 2007) to evaluate relevance between candidates and reference.
- **Diversity:** we simply use (1) Dist-1 and (2) Dist-2 (Li et al., 2016) to measure the diversity of candidates themselves. We use NLG evaluation code[5] to calculate these metrics.

### 5.2.2 Human Evaluation

In order to comprehensively evaluate the performance of VCD and other baselines, we adopt human evaluate on sampled generated candidates. Specifically, we randomly select 100 video-dialogue pairs from test set. The volunteer watches the video, reads dialogue context and rates corresponding candidates generated by VCD and other baselines. The volunteer is asked to rate the generated response from (0, 1, 2) in:

- **Fluency:** whether the response is fluent, reasonable and logical.
- **Context-relevance:** whether the response is relevant to the given dialogue context.
- **Event-relevance:** whether the response is relevant to the event contents shown in the video.
- **Opinion-clarity:** whether the response has a clear and not self-contradictory opinion.
- **Knowledge-relevance:** whether the response involves a range of external knowledge or common sense.
- **Overall quality:** overall quality of generated response considering event contents, dialogue context and external knowledge.

## 5.3 Baselines

We compare VCD with other state-of-the-art baselines. We classify them into different types according to what kind of information they accept while generating response.

- **Context only:** only takes dialogue context as input. DialoGPT (Zhang et al., 2020) is a dialogue generation model based on GPT-2 (Radford et al., 2019) architecture and trained on 147M multi-turn text-only dialogue corpus. We finetune a DialoGPT model on SportsVD.
- **Context + Event:** takes dialogue context and event contents as evidence. (1) Blip-2 (Li et al., 2023) is a state-of-the-art vision-language model. It utilizes a trainable Q-former to bridge the frozen vision encoder and large language model OPT (Zhang et al., 2022). Since Blip-2 only accepts image as input, we randomly sampled one frame from the video as the visual information. (2) *VCD* is our method, reading both visual and textual information and generating response.
- **Context + Event + Knowledge:** takes dialogue context, event contents and external knowledge into consideration while predicting. (1) ChatGPT[6] has been the best performing benchmark in dialogue generation since released. Since the implementation details is not open source, we don't finetune it on SportsVD. Due to incapability of reading video information by ChatGPT api, we add event captions together with dialogue context and ask ChatGPT to generate response. Since ChatGPT has its own knowledge base, we regard it as knowledge-enabled baseline. (2) *VCD+Common* is our method, based on *VCD* but integrating commonsense knowledge into response generation. (3) *VCD+Domain* is based on *VCD*, integrating domain-specific knowledge.

## 5.4 Main Results

We summarize main experiments results in Table 3 and Table 4. For relevance metrics, *VCD+Domain* achieves almost the best performance among all baselines. Meanwhile, *VCD+Common* achieves the second best performance regarding to relevance evaluation. Without external knowledge, relevance metric decreased slightly for *VCD* and

[5] `https://github.com/Maluuba/nlg-eval`

[6] `https://openai.com/blog/introducing-chatgpt-and-whisper-apis`

| Type | Methods | BLEU_3 | BLEU_4 | METEOR | ROUGE_L | CIDEr | distinct_1 | distinct_2 |
|---|---|---|---|---|---|---|---|---|
| C | DialoGPT | 0.08 | 0.03 | 1.38 | 3.15 | 1.26 | 11.57 | 47.50 |
| C+E | *VCD (Ours)* | 0.29 | 0.11 | 2.29 | 5.25 | **2.95** | 5.65 | 34.78 |
|  | Blip-2 | 0.25 | 0.11 | 1.82 | 3.68 | 2.02 | **14.02** | **54.94** |
| C+E+K | ChatGPT | 0.34 | 0.08 | **2.65** | 5.40 | 1.74 | 6.23 | 31.70 |
|  | *VCD+Common (Ours)* | 0.44 | 0.17 | 2.48 | 5.49 | 2.73 | 5.18 | 31.98 |
|  | *VCD+Domain (Ours)* | **0.50** | **0.20** | 2.55 | **5.59** | 2.80 | 5.16 | 31.85 |

Table 3: Evaluation results on automatic metrics. Bold number indicates the best performance among baselines and number with underline denotes the second best performance.

Blip-2. On the contrary, due to incapability of reading information of the event and external knowledge, DialoGPT can hardly produce comparable response to other methods. Results validates the significance of events understanding and external knowledge association in event-contents-oriented dialogue generation. This is also in align with our conclusion that SportsVD involves large amount of external knowledge. Human evaluation results also validate the high relevance to event and external knowledge of candidates generated by our method. *VCD+Domain* and *VCD+Common* has highest overall quality score. Blip-2 and ChatGPT performs bad in human evaluation due to too many meaningless words in generated candidates.

| Method | F | Ctx-R | Evnt-R | Opn-C | Knwl-R | Oa-Q |
|---|---|---|---|---|---|---|
| DialoGPT | 1.01 | 1.05 | 0.77 | 1.02 | 0.67 | 1.0 |
| *VCD* | 1.38 | **1.63** | 1.08 | 1.38 | 0.8 | 1.44 |
| Blip-2 | 0.56 | 0.75 | 0.38 | 0.23 | 0.1 | 0.26 |
| ChatGPT | **1.63** | 0.95 | 1.08 | 1.11 | 0.18 | 1.02 |
| *VCD+Common* | 1.55 | 1.61 | 1.11 | 1.45 | 0.93 | 1.49 |
| *VCD+Domain* | 1.41 | 1.62 | **1.31** | **1.47** | **1.22** | **1.55** |

Table 4: Human evaluation results on Fluency (F), Context-relevance (Ctx-R), Event-relevance (Evnt-R), Opinion-clarity (Opn-C), Knowledge-relevance (Knwl-R) and Overall quality (Oa-Q). Bold number indicates the best performance among baselines and number with underline denotes the second best performance.

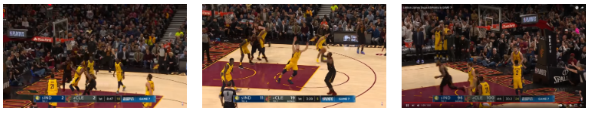
However, for diversity evaluation, Blip-2 and DialoGPT outperform other methods. We assume that external knowledge limits the ability of generating more different words with less possibility. Besides, the pre-training data distribution between different methods and scale of parameters also causes low diversity in response generated by *VCD*.

## 5.5 Ablation Study

We provide a brief ablation study to investigate the importance of integrating external knowledge into dialogue generation. From Table 3 and Table 4, *VCD+Common* and *VCD+Domain* outperform *VCD* in both automatic evaluation and human evaluation. On the other hand, *VCD+Domain* achieves higher score in almost all metrics which indicates domain knowledge provides better evidence in both video understanding and response generation. Hence, the introduction of external knowledge is effective and efficiency.

## 5.6 Case Study

We provide a case of generated responses in Figure 5 to further investigate the performance of our method and other baselines. In Figure 5, *VCD+Domain* can infer the results of the event (cavs wins the series), the speaker's position (perhaps a mavs fan because mavs fans hate lebron) and counterattack. We see the impressive ability of integrating external knowledge of VCD+Domain and clear point of view. ChatGPT acts more like echoing speaker's opinion with weak understanding of the event contents. Without seeing event contents, DialoGPT can hardly generate any informative response. Due to taking only one frame as vision information, Blip-2 also cannot generate informative response. *VCD* can understand the event contents but the generated response is not quite related to the context in this case. While other baselines can generate reasonable response, they can hardly integrate external knowledge and event contents at the same time except *VCD+Domain*.



**Caption:** LeBron James Drops 45 Points in Game 7!
**Context:** refs couldn't let lebron lose in the first 1st round lol
**Ground Truth:** the guy above is a warriors bandwagon
**DialoGPT:** the refs wanted the cavs to lose
**Blip2:** lebron is overrated
**VCD:** as a hater we have to admit lebron is the most clutch player in nba history
**ChatGPT:** Wow, LeBron really showed up in game 7 regardless of any controversies
**VCD+Domain:** mavs only made the first round when lebron went to the second round lol

Figure 5: Case study: an example comparison between VCD and other baselines

## 6 Conclusion

In this paper, we introduce a new multi-modal dialogue dataset SportsVD. To our best knowledge, it is the first events-content-based video-dialogue dataset. We collect videos and dialogues in sports domains on Youtube. Real users' comments ensure the quality of dialogue and clarity of views towards the event. Meanwhile, we present VCD, a novel video-dialogue generation method which can understand event contents, integrate external knowledge and generate reasonable response. Experiments between VCD and other baselines on SportsVD demonstrate the outstanding performance of our method.

## Acknowledge and Ethical Statement

We would like to thank all reviewers for their valuable comments and volunteers for human evaluation. This research is partially supported by Australian Research Council (ARC) Discovery Project Grant. The Funding ID is: ARC DP230101753.

At last, we discuss the potential ethical concern. Our dataset SportsVD is collected from YouTube following fair use on YouTube legally. We have anonymised the authors of the videos and comments. All identifiable information in the dataset was removed. Considering all videos are from sports highlights which is not sensitive and raw data is already public, there is no concern about privacy leak and offensive content.

## References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Hung Le, Nancy Chen, and Steven Hoi. 2022. VGNMN: Video-grounded neural module networks for video-grounded dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3377–3393, Seattle, United States. Association for Computational Linguistics.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. 2021. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5596–5611, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, and Zhiwu Lu. 2023. Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 1303–1313, New York, NY, USA. Association for Computing Machinery.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, Toronto, Canada. Association for Computational Linguistics.

Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. Livebot: Generating live video comments based on visual and textual contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6810–6817.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 223–240. Springer.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. 2021. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4287–4296, New York, NY, USA. Association for Computing Machinery.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

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, page 2599–2607, New York, NY, USA. Association for Computing Machinery.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. PhotoChat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, Online. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France. European Language Resources Association.