Collaborative Reasoning on Multi-Modal Semantic Graphs for Video-Grounded Dialogue Generation

Xueliang Zhao[1,2]*, **Yuxuan Wang**[1,2]*, **Chongyang Tao**[1],
Chenshuo Wang[1,2] and **Dongyan Zhao**[1,2,3]†
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Center for Data Science, AAIS, Peking University
[3]Beijing Institute for General Artificial Intelligence
{xl.zhao,chongyangtao,zhaody}@pku.edu.cn {wyx,wcs}@stu.pku.edu.cn

Abstract

We study video-grounded dialogue generation, where a response is generated based on the dialogue context and the associated video. The primary challenges of this task lie in (1) the difficulty of integrating video data into pre-trained language models (PLMs) which presents obstacles to exploiting the power of large-scale pre-training; and (2) the necessity of taking into account the complementarity of various modalities throughout the reasoning process. Although having made remarkable progress in video-grounded dialogue generation, existing methods still fall short when it comes to integrating with PLMs in a way that allows information from different modalities to complement each other. To alleviate these issues, we first propose extracting pertinent information from videos and turning it into reasoning paths that are acceptable to PLMs. Additionally, we propose a multi-agent reinforcement learning method to collaboratively perform reasoning on different modalities (i.e., video and dialogue context). Empirical experiment results on two public datasets indicate that the proposed model can significantly outperform state-of-the-art models by large margins on both automatic and human evaluations.

1 Introduction

Conversing with computers has become a crucial step toward general artificial intelligence, and it has attracted increasing attention from AI and NLP researchers. Multi-turn dialogue response generation and multi-modal question answering are two high-profile initiatives made toward this goal. The task of multi-turn dialogue response generation necessitates the agent comprehending the key information in the dialogue context in order to provide a cohesive, fluent and informative response (Zhao et al., 2017; Tao et al., 2018). Multi-modal question answering, on the other hand, necessitates the agent's understanding of both the textual and visual contexts (Antol et al., 2015; Tapaswi et al., 2016; Jang et al., 2017). The video-grounded dialogue (Alamri et al., 2018; Pasunuru and Bansal, 2018) is a generalization of the above two tasks, in which the agent must observe multi-modal contents and engage in a conversation with the human, rather than simply responding to the last utterance or ignoring the visual contents. Compared to multi-turn dialogue response generation and multi-modal question answering, the distinctive challenges posed by video-grounded dialogue generation can be summarized as: (1) Unlike traditional multi-turn dialogue that can directly use large-scale pre-trained language models (PLMs), video-grounded dialogue cannot directly use PLMs due to their incapacity to process video input; (2) In comparison to multi-modal question answering, video-grounded dialogue necessitates reasoning on both video and multi-turn textual context, and there is usually a complementarity between different modalities that should be taken into account.

Although having made notable progress in video-grounded dialogue, existing approaches still fail to recognize the aforementioned challenges. On one hand, existing approaches cannot be effectively combined with PLMs, which presents obstacles to exploiting the power of state-of-the-art pre-training technology. The reasons can be summarized into two categories: (1) Simply appending the video features to the text embeddings presents a challenge for the model to obtain an in-depth understanding of the video (Li et al., 2020; Le and Hoi, 2020; Le et al., 2021). To investigate this problem further, we compare the performance of these models before and after removing the video from the input. As demonstrated in Table 1, most metrics only show a tiny shift, and several even increase once the video is removed; and (2) Overly complex designs for the Transformer that are difficult to transfer to PLMs (Le et al., 2020; Kim et al., 2021; Geng et al.,

*Equal Contribution.
†Corresponding author: Dongyan Zhao.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5988–5998
December 7-11, 2022

Model	BLEU4	METEOR	ROUGE-L	CIDEr
with video				
RLM	0.402	0.254	0.544	1.052
VGD-GPT2	0.388	0.251	0.539	0.998
PDC-GPT	0.385	0.260	0.545	1.010
Ours	0.414	0.265	0.558	1.078
w/o video				
RLM	0.401	0.255	0.545	1.038
VGD-GPT2	0.393	0.251	0.537	1.016
PDC-GPT	0.388	0.261	0.543	1.020
Ours	0.405	0.264	0.554	1.064

Table 1: Pilot study on AVSD@DSTC7.

2021). On the other hand, multi-modal information should be used in conjunction with each other, and reasoning on different modalities should be done **collaboratively** rather than **independently**. Existing approaches fall short when it comes to reasoning jointly on multi-modalities, since they either separate the reasoning of different modalities (Li et al., 2020) or employ a cross-modal attention mechanism which is difficult to train without direct supervision (Le et al., 2020; Kim et al., 2021; Geng et al., 2021).

To address the aforementioned issues, we propose extracting relevant information from videos and converting it into reasoning paths, which are in the form of natural language and can be fed directly into PLMs. Besides, we propose a multi-agent reasoning framework that is based on the multi-agent reinforcement learning (MARL) theory. Specifically, we design a video agent and a context agent which learn to find the chains of reasoning on the multi-modal semantic graphs. We further design a central communicator to make the two agents work in a collaborative manner. Our framework has the following advantages: (1) the multi-modal reasoning paths are compatible with the input of PLMs; (2) the reasoning process can be "supervised" by designing appropriate reward functions; and (3) the communication mechanism allows the information from different modalities better complement each other. We conduct extensive experiments on two benchmark datasets for video-grounded dialogue generation, including AVSD@DSTC7 (Alamri et al., 2018) and Twitch-FIFA (Pasunuru and Bansal, 2018). Experiment results show that, thanks to the multi-agent reasoning framework, our model can significantly outperform state-of-the-art methods in terms of both automatic and human evaluations.

Our contributions in the paper are three-fold: (1) Identifying the issue that current PLMs-based approaches are unable to fully comprehend the video content although showing promising results in automatic evaluation metrics. (2) Proposal of a multi-agent reasoning framework upon PLMs that can let information from different modalities reinforce each other and discover multi-modal reasoning paths. (3) Empirical verification of the effectiveness of the proposed model on two benchmarks of video-grounded dialogue generation.

2 Related Work

The majority of early works on dialogue generation use hand-crafted rules or templates to construct dialogue systems (Weizenbaum, 1966; Wallace, 2009). A number of initiatives have been made to develop end-to-end open-domain dialogue generation models (Ritter et al., 2011; Gehring et al., 2017; Vaswani et al., 2017), which have been inspired by the developments in the field of machine translation. Following that, the vanilla encoder-decoder architecture is frequently utilized to enhance response quality, and numerous modifications to this architecture have been made to enhance response diversity (Zhao et al., 2017; Tao et al., 2018), model the structure of conversation contexts (Zhang et al., 2019), introduce external knowledge (Dinan et al., 2019; Zhao et al., 2020) and control response attributes (Wang et al., 2018; See et al., 2019; Wang et al., 2020).

The research on generating dialogue from video was started by Alamri et al. (2018). After that, Hori et al. (2019a) present an LSTM-based encoder-decoder architecture with multi-modal attention that merely combines textual and visual data via a projection matrix. A multi-modal transformer network is introduced in Le et al. (2019) to encode videos and incorporate data from several modalities. Hori et al. (2019b) uses a joint student-teacher learning approach to make up for a missing video description in which the student network is trained to mimic the teacher's response. VGD-GPT (Le and Hoi, 2020) is based on a pre-trained GPT-2 model and formulates the video-grounded dialogue generation as a sequence-to-sequence task. On a pre-trained GPT-2 model, RLM (Li et al., 2020) provides a multi-task learning strategy. Additionally, BiST (Le et al., 2020) models the dependencies between text and visual in two directions: spatial to temporal and temporal to spatial. With visual attention, PDC-GPT (Le et al.,

2021) learns to anticipate the reasoning process on turn-level semantic graphs. For further reasoning, SCGA (Kim et al., 2021) constructs a structured graph based on a multi-modal coreference technique, while STSGR (Geng et al., 2021) introduce a shuffled transformer reasoning framework on semantic scene graph. In contrast to previous approaches, this paper focuses on how to build a multi-modal reasoning approach that can cooperate with PLMs in a way that facilitates the complementary nature of information from various modalities.

The study of reasoning on various types of graph structures for dialogue generation is related to our work. Moon et al. (2019) create a KG walk path for each entity retrieved in an effort to explain conversational reasoning in a natural way. Jung et al. (2020) develop a dialogue-conditioned path traversal model with attention flows and improve the comprehension of the path reasoning process. Xu et al. (2020) propose to represent dialogue transitions as graphs. Previous approaches typically concentrate on textual graphs, but video-grounded dialogue contains multi-modal contexts, which makes it difficult to conduct reasoning.

3 Approach

3.1 Overview

Suppose that we have a dataset $\mathcal{D} = \{V_i, U_i, R_i\}_{i=1}^{N}$ with N denoting the total number of datapoints. For the i-th datapoint, V_i signifies a brief video clip, $U_i = \{u_{i,1}, u_{i,2}, \cdots, u_{i,n}\}$ serves as the dialogue context with $u_{i,j} = \{w_{i,j}^1, w_{i,j}^2, \cdots, w_{i,j}^m\}$ denoting the j-th utterance. n and m are the number of utterances in a context and the number of words in an utterance respectively. R_i is a response that is factually consistent with the video while also catching up with the dialogue context. Our goal is to learn a generation model $p(R|V, U; \theta)$[1] (θ denotes the parameters of the model) from $\mathcal{D}$, so that given a new dialogue context U associated with a video V, one can generate a response following $p(R|V, U; \theta)$.

To alleviate the heterogeneity of different modalities, we first represent the video as well as the dialogue context as semantic graphs (will be elaborated in Section 3.2). Figure 1 illustrates the architecture of the proposed model. In a nutshell, the model is composed of a multi-modal reasoning module and a generation module. The multi-modal reasoning module is responsible for extracting crucial signals from multi-modal contexts (Section 3.3). Specifically, it consists of a video agent, a text agent and a central communicator. The video agent and the text agent are responsible for extracting reasoning paths from the video semantic graph and the text semantic graph respectively. Taking the latest context utterance as input, they determine the query entities from which they start traversing the graphs to find the answer-providing paths. To search for answer-providing paths more efficiently, we devise a central communicator to transport the entire path histories between video and text agents. The reasoning paths, which form interpretable provenances for the prediction, are integrated by the generation module to synthesize a response (Section 3.4).

3.2 Multi-Modal Graph Construction

The crucial step in building the semantic graph for video reasoning is gathering the collection of facts from the unstructured video data, which take the form of subject-predicate-object triplets. Although there have been some previous attempts to extract such triplets from videos using relation detection (Liu et al., 2020), the models that have been made public struggle to build the proper relations because of the dramatic domain discrepancy between their training corpus and the benchmark dataset for video-grounded dialogue. Therefore, we resort to video action recognition (Zhu et al., 2020) to extract meaningful structural representations from video. Specifically, we first employ the slowfast model (Feichtenhofer et al., 2019), which is pre-trained on the Charades (Sigurdsson et al., 2016) and Kinetics dataset (Kay et al., 2017), to extract all potential action classes and only reserve those with a probability greater than 0.5. Given the extracted facts $\{(e_s^v, r, e_o^v)\}$ with e_s^v, r^v and e_o^v standing for subject, predicate and object respectively, we can construct a video semantic graph $\mathcal{G}^v = (N^v, E^v)$ in which the entities e_s^v and e_o^v are represented as nodes (i.e., $e_s^v, e_o^v \in N^v$) and the relation r^v is represented as a labeled edge connecting them (i.e., $(e_s^v, r, e_o^v) \in E^v$).

The semantic graph for dialogue context, $\mathcal{G}^u = (N^u, E^u)$, is constructed in a similar way, except that we employ open information extraction (OpenIE) technology to extract subject-predicate-object triplets. Specifically, we first apply the co-reference resolution tool (e.g., AllenNLP (Gardner et al.,

[1] We omit the subscript to reduce clutter.

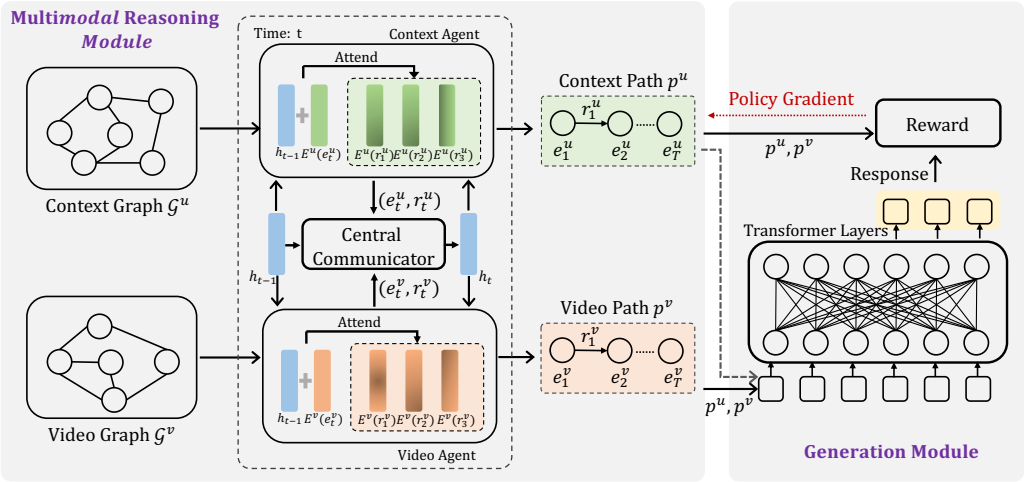

Figure 1: Architecture of the proposed model.

2017) in our experiments) to restore all the pronouns to their original name entities. Then we extract all relation triplets in a dialogue context by combining the outputs of OpenIE 5.1 (Saha and Mausam, 2018) and Stanford OpenIE (Angeli et al., 2015). We further remove unnecessary information after getting all the triplets by combining all entities with high semantic similarity, as determined by the cosine similarity between the word2vec embeddings (Mikolov et al., 2013).

3.3 Multi-Agent Reasoning Process

Inspired by recent advances in graph-grounded generation (Moon et al., 2019; Xu et al., 2020), we decompose the problem of video-grounded generation into two steps: (1) identify answer-providing paths on the graph that might contain crucial signals for catching up with the context; (2) generate a response using the extracted paths as additional information. However, independently extracting the chains of reasoning for each modality will result in a sub-optimal solution, since the video provides crucial guidance for text reasoning and vice versa. To this purpose, we propose a multi-agent reasoning framework, where agents responsible for different modalities can work in a collaborative manner.

We formulate the multi-modal reasoning task as a partially observable multi-agent sequential decision process on semantic graphs $\mathcal{G}^v$ and $\mathcal{G}^u$. Intuitively, we want a state s_t at time t to be a summary of previous experiences: $s_t = (o_1, a_1, \cdots, a_{t-1}, o_t)$, where $o_t = (e_t^v, e_t^u)$ and $a_t = (r_t^v, r_t^u)$ stand for observations/entities of all agents at time t and actions/relations taken by them respectively. $o_1 = (e_1^v, e_1^u)$ is the query entity to start traversing on the graphs and is defined as:

$$
\begin{aligned}
e_1^v &= \arg\max_{e^v \in N^v} E^v(e^v)^\top E(u_n), \\
e_1^u &= \arg\max_{e^u \in N^u} E^u(e^u)^\top E(u_n),
\end{aligned}
\tag{1}
$$

where $E^v(e^v)$, $E^u(e^u)$ and $E(u_n)$ denote the embeddings for e^v, e^u and the last utterance u_n respectively. The state of the environment is ubiquitous and shared by all agents, but in a multi-agent setting, the observation and the action are both private and only accessible by the individual agent. Take the video agent as an example. When receiving a local observation e_t^v, or the current location on $\mathcal{G}^v$, it will select an outgoing edge r_t^v with its own private policy network $p_v(r|s_t)$, and obtain a reward from the environment. We also design a central communicator to encode historical information and promote multiple agents to work in a collaborative way. The details about the central communicator, the private policy network, and the reward will be described as follows:

Central Communicator. To make full use of the information from different modalities and facilitate the reasoning process, we design a central communicator which can get access to the local observations and actions of all agents (Feng et al., 2018). The central communicator works by recursively encoding the historical information (i.e., $(o_1, a_1, \cdots, o_t, a_t)$) into a message h_t and transporting this message between agents. Specifically, we implement the central communicator as a recurrent neural network, with the hidden state h_t

encoding the past observations and actions. At time t, the central communicator takes current observation o_t and action a_t as input, and updates the hidden state as:

$$\begin{aligned} h_t &= \text{RNN}(h_{t-1}, f(a_t, o_t); \phi), \\ f(a_t, o_t) &= W_c[E^v(r_t^v); E^v(e_t^v); E^u(r_t^u); E^u(e_t^u)], \end{aligned} \tag{2}$$

where W_c is a learnable projection matrix, $E^v(r_t^v)$,$E^v(e_t^v)$, $E^u(r_t^u)$ and $E^u(e_t^u)$ are embeddings of r_t^v, e_t^v, r_t^u and e_t^u respectively. Consequently, with the help of message h_{t-1}, the full state can be approximated as $s_t^v \approx (h_{t-1}, e_t^v)$ for the video agent or $s_u^v \approx (h_{t-1}, e_t^u)$ for the context agent.

Private Policy Network. Each agent has its own private policy network that chooses an outgoing edge at the current location. Take the video agent as an example. With the guidance of transported message h_{t-1}, the policy network can be approximated as $p_v(r|s_t) \approx p_v(r|h_{t-1}, e_t^v)$, which is formally defined as:

$$\begin{aligned} p_v(r|h_{t-1}, e_t^v; \psi^v) &= \frac{e^{g(h_{t-1}, e_t^v, r)}}{\sum_{r \in \mathcal{R}(e_t^v)} e^{g(h_{t-1}, e_t^v, r)}}, \\ g(h_{t-1}, e_t^v, r) &= W_a[h_{t-1}; E^v(e_t^v); E^v(r)], \end{aligned} \tag{3}$$

where W_a is a learnable parameter and $\mathcal{R}(e_t^v)$ denotes all outgoing edges of the node e_t^v. We define the private policy network for context agent $p_u(r|h_{t-1}, e_t^u; \psi^u)$ following the same procedure.

Reward. We only have a reward once the complete chain of reasoning is obtained. Given the final state $s_T = (o_1, a_1, \cdots, a_{T-1}, o_T)$ (T is the maximum time constraint), we can obtain the reasoning paths p^v and p^u for video and dialogue context respectively, and define the reward as:

$$\begin{aligned} Re(s_T) &= \text{ROUGE}(p^v, p_{gt}) + \text{ROUGE}(p^u, p_{gt}), \\ p^v &= (e_1^v, r_1^v, \cdots, r_{T-1}^v, e_T^v), \\ p^u &= (e_1^u, r_1^u, \cdots, r_{T-1}^u, e_T^u), \end{aligned} \tag{4}$$

where p_{gt} is the subject-predicate-object triplet extracted from the ground-truth response, and $\text{ROUGE}(\cdot, \cdot)$ is a function that returns the ROUGE-1 score (Lin, 2004) between two sequences.

3.4 Generation Module

We employ the pre-trained GPT-2 (Radford et al., 2019) as the backbone of our generation module, which synthesizes a response conditioning on the reasoning path p^v for video data, the reasoning path p^u for dialogue context and the last utterance in context $u_n = \{w_n^1, \cdots, w_n^m\}$. Formally, the input of the generation module is defined as:

$$\{e_1^v r_1^v \dots e_T^v [\text{SEP}] e_1^u r_1^u \dots e_T^u [\text{SEP}] w_n^1 \dots w_n^m\}, \tag{5}$$

where [SEP] is a special token separating different types of data. The probability of generating the response $R = \{w_r^1, w_r^2, \cdots, w_r^m\}$ is formulated as:

$$\begin{aligned} p(R|V, U; \theta) &\approx p(R|p^v, p^u, u_n; \theta) \\ &= \prod_{i=1}^{m} p(w_r^i | w_r^{<i}, p^v, p^u, u_n). \end{aligned} \tag{6}$$

3.5 Learning Details

To estimate θ (i.e., parameters of the generation module), we directly minimize the negative log-likelihood of response R through MLE loss:

$$\mathcal{L}_{mle} = -\sum_{i=1}^{m} \log p(w_r^i | w_r^{<i}, p^v, p^u, u_n). \tag{7}$$

The parameters of private policy networks (i.e., ψ^v and ψ^u), as well as the parameters of the central communicator (i.e., ϕ), are optimized through policy-gradient method (Sutton et al., 2000). Specifically, we sample reasoning paths $\tilde{p}^v$ and $\tilde{p}^u$ according to the private policy networks and the central communicator, and define the loss as follows:

$$\begin{aligned} \mathcal{L}_{pg}^v &= -Re(\tilde{s}_T) \sum_{t=1}^{T-1} \log p_v(r_t^v | h_{t-1}, e_t^v; \psi^v), \\ \mathcal{L}_{pg}^u &= -Re(\tilde{s}_T) \sum_{t=1}^{T-1} \log p_u(r_t^u | h_{t-1}, e_t^u; \psi^u), \end{aligned} \tag{8}$$

where $Re(\cdot)$ is the reward function defined in Eq. 4 and $\tilde{s}_T$ is the final state when sampling the chains of reasoning. The general training process is conducted by alternately optimizing $\mathcal{L}_{mle}$ and $\mathcal{L}_{pg} = \mathcal{L}_{pg}^v + \mathcal{L}_{pg}^u$.

4 Experiments

4.1 Datasets

We evaluate our model on two benchmark datasets for video-grounded dialogue generation:

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
			with caption				
Naive Fusion	0.628	0.481	0.377	0.298	0.220	0.491	0.748
MTN	0.731	0.597	0.494	0.410	0.274	0.569	1.129
Student-Teacher	0.727	0.593	0.488	0.405	0.273	0.566	1.118
RLM	0.765	0.643	0.543	0.459	0.294	0.606	1.308
VGD-GPT2	0.750	0.621	0.516	0.433	0.283	0.581	1.196
BiST	0.755	0.619	0.510	0.429	0.284	0.581	1.192
PDC-GPT	0.770	**0.653**	0.539	0.449	0.292	0.606	1.295
Ours	**0.776***	0.652	**0.551***	**0.466***	**0.304***	**0.609**	**1.333***
			without caption				
Naive Fusion	0.626	0.485	0.383	0.309	0.251	0.487	0.746
MTN	0.692	0.556	0.459	0.368	0.259	0.537	0.964
Student-Teacher	0.675	0.543	0.446	0.371	0.248	0.527	0.966
RLM	0.694	0.570	0.476	0.402	0.254	0.544	1.052
VGD-GPT2	0.692	0.563	0.464	0.388	0.251	0.539	0.998
BiST	0.715	0.560	0.477	0.390	0.259	0.552	1.030
PDC-GPT	0.713	0.574	0.468	0.385	0.260	0.545	1.010
Ours	**0.717**	**0.590***	**0.491***	**0.414***	**0.265***	**0.558***	**1.078***

Table 2: Automatic evaluation results on the test set of AVSD@DSTC7. Numbers in bold are the best results. Significant improvements over the best baseline results are marked with $\star$ (t-test with p-value < 0.05).

AVSD@DSTC7 This dataset is constructed by Alamri et al. (2018) through crowd-sourcing and contains conversations about Charades videos (Sigurdsson et al., 2016).

Twitch-FIFA. This dataset is collected by crawling live-broadcast soccer game videos and the chats from Twitch.tv (Pasunuru and Bansal, 2018).

To facilitate reproducibility, we adopt the datasets shared by the publishers and conduct preprocessing strictly following the official code. Table 3 reports the statistics of AVSD@DSTC7 and Twitch-FIFA.

	AVSD@DSTC7			Twitch-FIFA		
	Train	Valid	Test	Train	Valid	Test
Number of Dialogues	7,659	1,787	1,710	10,150	2,153	2,780
Number of Utterances	76,590	17,870	6,745	110,602	19,362	31,245
Average Turns per Dialogue	10	10	3.94	10.52	8.99	11.24

Table 3: Statistics of the two datasets.

4.2 Evaluation Metrics

Automatic Evaluation. We choose 4 commonly used reference-based metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007) and CIDEr (Vedantam et al., 2015). We evaluate our models using the official code released by the owner of AVSD@DSTC7 dataset [2].

Human Evaluation. We also conduct a human evaluation to deepen our understanding of the quality of responses produced by different models. We randomly sample 300 examples from the test sets of AVSD@DSTC7, and hire 6 well-educated native speakers to conduct qualitative analysis on the results produced by our model and all competitive baselines, which are randomly mixed to obscure identification. The annotators evaluate the quality of the responses using three criteria: (1) *Language Fluency*: whether the response is fluent and devoid of grammatical errors, (2) *Context Coherence*: whether the response is coherent with the dialogue context, and (3) *Factual Correctness*: whether the response is factually consistent with the events depicted in the video. Each annotator rates each response for each aspect with a score from $\{0, 1, 2\}$ (representing "bad", "fair" and "good" respectively). Each response receives three scores for the aforementioned 3 aspects, and Fleiss' kappa (Fleiss, 1971) is used to gauge the level of agreement between all annotators.

4.3 Baseline Models

The following models are selected as baselines: (1) **Naive Fusion:** A model proposed by Hori et al. (2019a) which combines all modalities with a projection matrix. (2) **MTN:** A model proposed by Le et al. (2019) that is based on transformer architecture and employs query-guided attention to extract query-aware features from videos. (3) **Student-Teacher:** A model proposed by Hori et al. (2019b) that aims to alleviate the reliance on human-generated captions through a student-

[2] https://drive.google.com/open?id=1nz9Pu9YIfuZHzowhASXERajRXqE6DBQx

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
BIDAF	0.092	0.057	0.043	0.035	0.036	0.134	0.099
MTN	0.113	0.06	0.043	0.034	0.039	0.143	0.091
BiST	0.100	0.050	0.031	0.022	0.038	0.159	0.104
RLM	0.102	0.081	0.070	0.060	0.046	0.188	0.130
Ours	**0.128***	**0.101***	**0.083***	**0.069***	**0.052***	**0.193**	**0.176***

Table 4: Automatic evaluation results on the test set of Twitch-FIFA. Numbers in bold are the best results. Significant improvements over the best baseline results are marked with $\star$ (t-test with p-value < 0.05).

teaching learning method. (4) **RLM:** A model proposed by Li et al. (2020) that is trained with multi-task learning objectives to learn joint representations among different modalities. (5) **VGD-GPT2:** A model proposed by Le and Hoi (2020) that leverages the power of pre-trained language models for improving video-grounded dialogue. (6) **BiST:** A model proposed by Le et al. (2020) that exploits both spatial and temporal-level information to promote video understanding. (7) **PDC+GPT2:** A model proposed by Le et al. (2021) that conducts reasoning on the dialogue history to model the information flow at turn level.

All the baselines are taken from their open-source implementations or re-implemented strictly following the details in the original papers.

4.4 Implementation Details

In our experiments, the maximum time constraint T which serves as the stop criteria for reasoning is set as 3. The embedding sizes for relations and entities are all set as 100. The central communicator is implemented as an LSTM network whose size of the hidden state is set as 200. The generation module is implemented on the basis of the pre-trained GPT-2 (small) model which has $117M$ parameters. All models are learned with Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We initialize the learning rates to 0.001 and $6.25e-5$ for the multi-modal reasoning module and the generation module respectively and optimize the model with a linear learning rate decay strategy. The batch size is set as 8 in our experiments. In the test phase, we employ beam search in response decoding and set the beam size, max decode length and the length penalty as 4, 16 and 0.1 respectively. Early stopping on validation is adopted as a regularization strategy. All models are trained on an 8×RTX 3090 Ti machine.

4.5 Evaluation Results

In this section, we will compare the performance of various models on AVSD@DSTC7 and Twitch-FIFA. We conduct two experiment settings for AVSD@DSTC7, including *with caption* and *without caption*, since the video caption is unavailable in most real-world scenarios. Table 2 and Table 4 show the performance of our model on AVSD@DSTC and Twitch-FIFA respectively. From the results, we can observe that:

Model	Language Fluency	Context Coherence	Factual Correctness	Kappa
RLM	1.73	1.67	1.47	0.61
VGD-GPT2	1.76	1.54	1.50	0.65
BiST	1.64	1.58	1.49	0.71
PDC-GPT	1.78	1.68	1.53	0.64
Ours	**1.81**	**1.75**	**1.65**	0.76

Table 5: Human evaluation results on AVSD@DSTC7. Numbers in bold are the best results.

(1) Our model achieves the new state-of-the-art on most metrics in both datasets, illuminating the effectiveness of the proposed multi-agent reasoning framework and the multi-modal semantic graphs. In particular, the proposed model outperforms RLM and PDC-GPT, the two best baselines on AVSD@DSTC7, since they both directly feed the video features for the generation procedure which presents obstacles for the PLMs to conduct multi-modal reasoning. This is also supported by the results in our pilot study (as shown in Table 1).

(2) In the AVSD@DSTC7 dataset, the caption has a significant impact on models since, in the absence of caption, all models' performance significantly degrades. Another intriguing finding is that reasoning-based methods (e.g., BiST, PDC-GPT and ours) rely less on the video caption as compared to methods without explicit reasoning (e.g., RLM and VGD-GPT2). This confirms the need for multi-modal reasoning, and our proposed method of collaborative reasoning is more effective.

(3) Because there is a lot of noise in the live-broadcast data collection, which is closer to the real-world scenario than the manually-labeled dataset, the PLMs-based methods (e.g., RLM and ours) perform better on Twitch-FIFA than others.

Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
Ours	0.717	0.590	0.491	0.414	0.265	0.558	1.078
-$\mathcal{G}^u$	0.706	0.578	0.481	0.403	0.261	0.554	1.067
-$\mathcal{G}^v$	0.704	0.579	0.481	0.405	0.260	0.552	1.063
-$\mathcal{G}^u$ & $\mathcal{G}^v$	0.697	0.573	0.474	0.399	0.257	0.545	1.048
-*Communicator*	0.708	0.578	0.478	0.400	0.259	0.548	1.058

Table 6: Ablation study on AVSD@DSTC7. All experiments are conducted in the *without caption* setting.

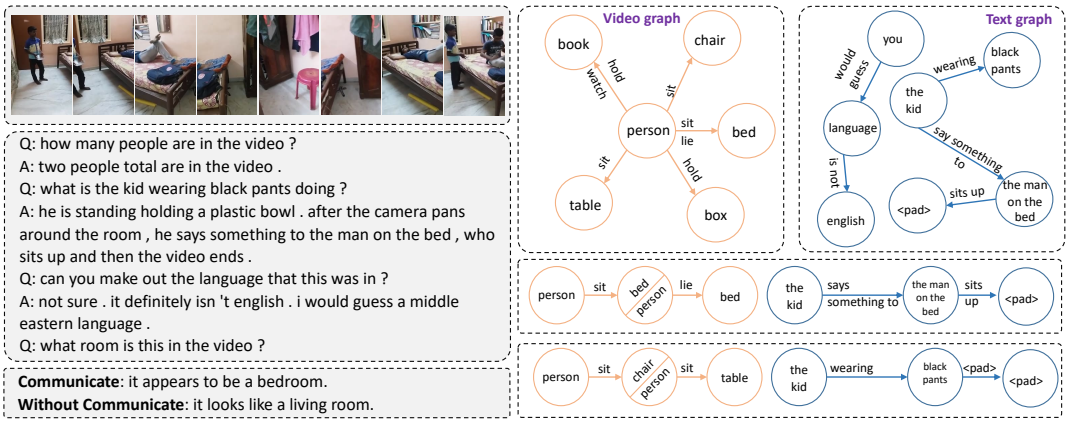

Figure 2: A case from the test set of AVSD@DSTC7.

Model	BLEU1-4	METEOR	ROUGE-L	CIDEr
T=1	0.713/0.584/0.485/0.407	0.262	0.551	1.055
T=2	0.717/0.590/0.491/0.414	0.265	0.558	1.078
T=3	0.714/0.588/0.489/0.412	0.265	0.554	1.082

Table 7: Performance of our model under different maximum time constraints.

This further emphasizes the value of integrating PLMs with multi-modal reasoning, one of the benefits of our proposed method.

Human Evaluation. Table 5 shows the results of human evaluation. Although our model achieves a language fluency score that is comparable to other baselines, it attains a significant improvement in context coherency and factual correctness, which is congruent with the results of our pilot experiments and automatic evaluation. The fact that all kappa values are more than 0.6 shows that the annotators are in agreement.

4.6 Discussions

Ablation Study. In addition to the main experiments, we compare the full model with the following variations to gain a better understanding of how each component affects the general performance: (1) -$\mathcal{G}^u$: the context graph is removed; (2) -$\mathcal{G}^v$: the video graph is removed; (3) -$\mathcal{G}^u$ & $\mathcal{G}^v$: both the context graph and the video graph are removed. Here, the model directly generates the response based on the dialogue context and the video features provided by Alamri et al. (2018); and (4) -*Communicator*: the Central Communicator is removed. In this instance, the context agent and video agent each independently reason on graphs. The experiment results of ablation are shown in Table 6. We can draw the following conclusions: (1) the multi-modal semantic graphs are both significant, as the performance is negatively impacted by deleting one or more of them. Although they are built using off-the-shelf tools with heuristics, they nonetheless contain significant information that enables the agents to locate chains of reasoning that lead to solutions; and (2) the communicator is helpful because it enables crucial signals from different modalities to reinforce each other.

Effect the Maximum Time Constraint T. We continue to look at how sensitive the model is to various selections of the maximum time constraint T. In order to achieve this, we change the value of T in $\{1; 2; 3\}$, and then report the evaluation results in Table 7. As can be shown, our model performs best when $T = 2$ since a larger maximum time constraint will introduce more irrelevant entities and relations into generation, whereas a smaller number (i.e., $T = 1$) limits the reasoning paths to only the entity that is most similar to the last utterance.

Case Study. We further conduct a case study to have a deeper understanding of the multi-modal reasoning process in our model. Figure 2 shows an example from the test set of AVSD@DSTC7. We can see that our model is able to precisely construct the reasoning paths for dialogue context and video respectively, and to produce a response that accurately captures the factual information in the video. For comparison, we also provide the results of a variant in which the central communicator has been eliminated. We can observe that the communication mechanism can effectively assist in retrieving relevant signals from multi-modal data.

5 Conclusion

We propose a multi-modal reasoning framework that can be used in conjunction with PLMs to enable the complementation of information from various modalities. Specifically, we devise a video agent and a context agent to extract reasoning paths on video and dialogue contexts respectively. A central communicator is also designed to transport information between the two agents and enables their cooperative operation. The general framework is optimized through multi-agent reinforcement learning. Evaluation results on two benchmarks indicate that our model can significantly outperform state-of-the-art methods.

Limitations

We also recognize that our model has its certain limitations: (i) Due to multi-modal semantic graphs, our framework needs higher computation overheads to extract triplet relations from video and perform reasoning on dual graphs. Nonetheless, the multi-modal reasoning paths which are compatible with PLMs make our model still practical and scalable. (ii) The performance of our model may be limited to some extent by the quality of the dual graphs created by off-the-shelf tools.

Ethics Statement

This paper studies video-grounded dialogue generation and proposes a multi-modal reasoning framework based on multi-agent reinforcement learning to facilitate the complementarity of information from different modalities. There are no ethical concerns with this study. The datasets we used are widely used by other academics and are typically accessible to the public. No ethical or societal prejudice is introduced by the suggested strategy.

Acknowledgements

We appreciate the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106600).

References

Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batra, and Devi Parikh. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, volume 2.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Jun Feng, Heng Li, Minlie Huang, Shichen Liu, Wenwu Ou, Zhirong Wang, and Xiaoyan Zhu. 2018. Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning. In *Proceedings of the 2018 World Wide Web Conference*, pages 1939–1948.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. 2021. Dynamic graph

representation learning for video dialog via multimodal shuffled transformers. In *AAAI*.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019a. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE.

Chiori Hori, Anoop Cherian, Tim K Marks, and Takaaki Hori. 2019b. Joint student-teacher learning for audio-visual scene-aware dialog. In *INTERSPEECH*, pages 1886–1890.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Jaehun Jung, Bokyung Son, and Sungwon Lyu. 2020. Attnio: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In *EMNLP*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang Dong Yoo. 2021. Structured co-reference graph attention for video-grounded dialogue. *ArXiv*, abs/2103.13361.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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.

Hung Le, Nancy F Chen, and Steven Hoi. 2021. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *International Conference on Learning Representations*.

Hung Le and Steven CH Hoi. 2020. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5842–5848.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven CH Hoi. 2020. Bist: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1859.

Hung Le, Doyen Sahoo, Nancy F. Chen, and Steven C. H. Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *ACL*.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, Cheng Niu, and Jie Zhou. 2020. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *arXiv preprint arXiv:2002.00163*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 2020. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10840–10849.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *COLING*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. 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*, pages 4566–4575.

Richard S. Wallace. 2009. The anatomy of a.l.i.c.e.

Qiansheng Wang, Yuxin Liu, Chengguo Lv, Zhen Wang, and Guohong Fu. 2020. Cue-word driven neural response generation with a shrinking vocabulary. *ArXiv*, abs/2010.04927.

Yansen Wang, Chen-Yu Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. In *ACL*.

Joseph Weizenbaum. 1966. Eliza: A computer program for the study of natural language communication between man and machine. volume 9, pages 36–45. ACM.

Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Conversational graph grounded policy learning for open-domain conversation generation. In *ACL*.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, pages 654–664.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. 2020. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*.