Describing Unseen Videos via Multi-Modal Cooperative Dialog Agents

Ye Zhu[1], Yu Wu[2,3], Yi Yang[2], and Yan Yan[1][*]

[1] Texas State University, San Marcos, USA
{ye.zhu, tom_yan}@txstate.edu
[2] ReLER, University of Technology Sydney, Australia
yu.wu-3@student.uts.edu.au,yi.yang@uts.edu.au
[3] Baidu Research, China

Abstract. With the arising concerns for the AI systems provided with direct access to abundant sensitive information, researchers seek to develop more reliable AI with implicit information sources. To this end, in this paper, we introduce a new task called video description via two multi-modal cooperative dialog agents, whose ultimate goal is for one conversational agent to describe an unseen video based on the dialog and two static frames. Specifically, one of the intelligent agents - *Q-BOT* - is given two static frames from the beginning and the end of the video, as well as a finite number of opportunities to ask relevant natural language questions before describing the unseen video. *A-BOT*, the other agent who has already seen the entire video, assists *Q-BOT* to accomplish the goal by providing answers to those questions. We propose a QA-Cooperative Network with a dynamic dialog history update learning mechanism to transfer knowledge from *A-BOT* to *Q-BOT*, thus helping *Q-BOT* to better describe the video. Extensive experiments demonstrate that *Q-BOT* can effectively learn to describe an unseen video by the proposed model and the cooperative learning method, achieving the promising performance where *Q-BOT* is given the full ground truth history dialog. Codes and models are available at `https://github.com/L-YeZhu/Video-Description-via-Dialog-Agents-ECCV2020`.

Keywords: Video Description, Dialog Agents, Multi-Modal.

1 Introduction

It is becoming a trend to exploit the possibilities to develop artificial intelligence (AI) systems with subtle and advanced reasoning abilities as humans, usually by providing AI with direct access to abundant information sources. However, what comes with this tendency is the arising concerns over the security and privacy issues behind such AI systems. Although direct access to rich information assists AI becoming more intelligent and competent, the general public starts
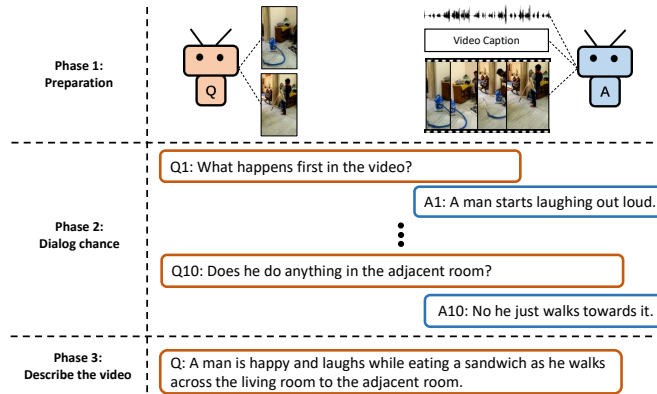
---

[*] Corresponding author.

Fig. 1. Task setup: Describing an unseen video via two multi-modal cooperative dialog agents. The entire process can be described in three phases. The ultimate goal is for *Q-BOT* to describe what happens between the beginning and the end of an unseen video based on the dialog history with *A-BOT*.

to question whether their sensitive personal information, such as the identifiable face images and voices, is in safe hands.

Despite the general concerns, nowadays research in AI and computer vision (CV) fields is experiencing a rapid transition from traditional 'low-level' tasks within single modality data such as image classification [21, 16, 15], object detection [32, 41, 24], machine translation [6], to more challenging tasks that involve multiple modalities of data and subtle reasoning [13, 51, 19, 40, 34], such as visual question answering [3, 4, 50] and visual dialog [10, 20, 46]. A meaningful and informative conversation, either between human-computer or computer-computer, is an appropriate task to demonstrate such a reasoning process due to the complex information exchange mechanism during the dialog. With the emergence of large-scale datasets such as VQA [4], GuessWhat [12] and AVSD [1, 18], much effort is devoted to study the techniques for machines to maintain natural conversation interactions in a sophisticated way [20, 46, 22, 28]. While most existing works still focus on the dialog itself and only involve a single agent, generally by providing the agent with direct access to sensitive information (*e.g.,* the complete video clips with identifiable human faces and voices), we wish to take a step forward to more secure and reliable AI systems with implicit information sources. To this end, in this paper we introduce a novel natural and challenging task with implicit information sources: describe an unseen video mainly based on the dialog between two cooperative agents.

The setup for our task is illustrated in Fig. 1, which involves two dialog agents, *Q-BOT* and *A-BOT*. Imagine the *Q-BOT* to be the actual AI system, and the *A-BOT* to be humans. The entire process can be described in three phases: In the first preparation phase, two agents are provided with different information. *A-BOT* is able to see the complete video with the audio signals

and captions, while Q-BOT is only given two static frames from the beginning and the end of the video. It is worth noting that the static frames given to Q-BOT have no specific requirements, they could be static images without visible humans or just the back of the person as shown in Fig. 1 and Fig. 2, which largely reduces the risks for the AI systems to recognize the actual person. In the second phase, Q-BOT has 10 opportunities to ask A-BOT relevant questions about the video, such as the event happened. After 10 rounds of question-answer interactions, Q-BOT is asked to describe the unseen video based on the initial two frames and the dialog history with A-BOT. Under this task setup, the AI system Q-BOT accomplishes a multi-modal task without direct access to the original information, but learns to filter and extract useful information from a less sensitive information source, $i.e.$, the dialog. It is highly impossible for AI systems to identify a person based on the natural language descriptions. Therefore, such task settings and reasoning ability based on implicit information sources have great potential to be applied in a wide practical context, such as the smart home systems, improving the current AI systems that rely on direct access to sensitive information to accomplish certain tasks. Notably, instead of directly asking for the final video descriptions from human users, our task formulation that requires AI systems to gradually ask questions helps to reduce the bias and noises usually contained in the description directly given by human individuals.

It is a challenging and natural task compared to previous works in the field of the visual dialog. The difficulty mainly comes from the implicit information source of multiple modalities and the more complex reasoning process required for both agents. Specifically, this task also differs from the traditional video captioning task due to the fact that Q-BOT has never seen the entire video. Intuitively, it can be considered as establishing an additional information barrier($i.e.$, the dialog) between the direct visual data input and the natural language video caption output. Fig. 2 shows some examples from the AVSD dataset [1, 18], whose data collection process resembles to our task setup. We observe that the ground truth video captions shown to A-BOT and the expected final descriptions given by Q-BOT are quite different, thus revealing the actual gap existing in human reasoning. This fact again emphasizes the difficulties of our task.

The key aspect to consider in this work is the effective knowledge transfer from A-BOT to Q-BOT. A-BOT, who plays the role of humans, has full access to all the information, while Q-BOT only has an ambiguous understanding of the surrounding environment from two static video frames after the first phase. In order to describe the video with details that are not included in the initial input, Q-BOT needs to extract useful information from the dialog with A-BOT. Therefore, we propose a QA-Cooperative network that involves two agents with the ability to process multiple modalities of data. We further introduce a cooperative learning method that enables us to jointly train the network with a dynamic dialog history update mechanism. The knowledge gap and transfer process are both experimentally demonstrated.

The main contributions of this paper are: 1) We propose a novel and challenging video description task via two multi-modal dialog agents, whose ultimate
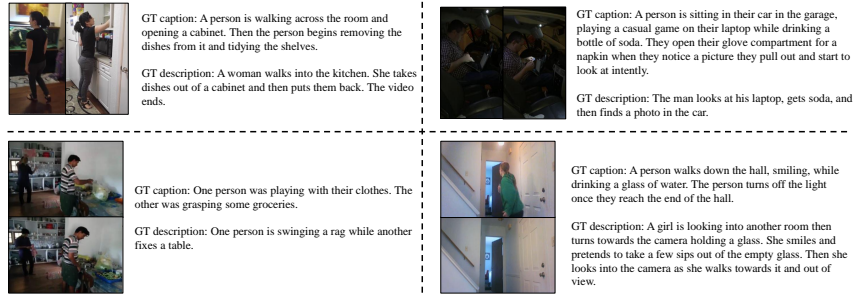
GT caption: A person is walking across the room and opening a cabinet. Then the person begins removing the dishes from it and tidying the shelves.

GT description: A woman walks into the kitchen. She takes dishes out of a cabinet and then puts them back. The video ends.

GT caption: A person is sitting in their car in the garage, playing a casual game on their laptop while drinking a bottle of soda. They open their glove compartment for a napkin when they notice a picture they pull out and start to look at intently.

GT description: The man looks at his laptop, gets soda, and then finds a photo in the car.

GT caption: One person was playing with their clothes. The other was grasping some groceries.

GT description: One person is swinging a rag while another fixes a table.

GT caption: A person walks down the hall, smiling, while drinking a glass of water. The person turns off the light once they reach the end of the hall.

GT description: A girl is looking into another room then turns towards the camera holding a glass. She smiles and pretends to take a few sips out of the empty glass. Then she looks into the camera as she walks towards it and out of view.

Fig. 2. Examples from the AVSD dataset [1, 18]. Two static images are the beginning and the end frame of the video clips. *GT caption* is the caption shown to the *A-BOT* under our task setup. *GT description* is actual the summary given by the human questioner at the end of the dialog during the data collection without directly watching the video, which corresponds to the video description required in our task. Intuitively, our task can be considered as establishing an additional information barrier(*i.e.*, the dialog) between the direct visual data input and the natural language caption output.

goal is for one agent to describe an unseen video mainly based on the interactive dialog history. This task establishes a more secure and reliable setting by providing implicit information sources to AI systems. 2) We propose a QA-Cooperative network and a cooperative learning method with a dynamic dialog history update mechanism, which helps to effectively transfer knowledge between the two agents. 3) We experimentally demonstrate the knowledge gap as well as the transfer process between two agents on the AVSD dataset [18]. With the proposed method, our *Q-BOT* achieves very promising performance comparable to the strong baseline situation where full ground truth dialog is provided.

2   Related Work

Image and Video Captioning. Image or video captioning refers to the task that aims to obtain textual descriptions for the given image or video. It is one of the first well-exploited tasks that combines both computer vision and natural language processing. You *et al.* [53] adopt Recurrent neural networks (RNNs) with selective attention to semantic concept proposals to generate image captions. Adaptive attention and spatial attention have also been studied in [25, 8]. Anderson *et al.* [3] propose to use bottom-up and top-down attention for object levels and other salient image regions to address this task. Yao *et al.* [52] use the attributes to further improve the performance. Rennie *et al.* [33] propose to train the image captioning system with reinforcement learning. Wang *et al.* [45] adopt conditional variational auto-encoders with an additive Gaussian encoding space to generate image descriptions. Wu *et al.* [47] propose a disentangled framework to generalize image captioning models to describe unseen objects. RecNet [44]

is introduced for video captioning. Mahasseni*et al.* [27] proposes to summarize the video with adversarial lstm networks in an unsupervised manner.

Visual Question Answering. Another similar research field beyond describing an image or a video is visual question answering (VQA). The objective of VQA is to answer a natural language question about the given image [4]. Different attention mechanisms including hierarchical attention [26], question-guided spatial attention [50], stacked attention [51], bottom-up and top-down attention [3] have been exploited. Das *et al.* [9] look into the question whether the existing attention mechanisms attend to the same regions as humans do. Shih *et al.* [38] map textual queries and visual features into a shared space to answer the question. Dynamic memory networks is also used for VQA [49].

Visual Dialog. Visual dialog is another succession of vision-language problem after visual captioning and VQA that is closely related to our work. Different from the VQA that only involves a single round of natural language interaction, visual dialog requires machines to maintain multiple rounds of conversation. Several datasets have been collected for this task [10, 12]. Jain *et al.* [20] proposes a symmetric baseline to demonstrate how visual dialog can be generated from discriminative question generation and answering. Additionally, attention mechanisms [29, 37] and reinforcement learning [46, 11] have also been exploited in the context of visual dialog.

Audio Data. As another important source of information, the audio modality has recently gained popularity in the research field of computer vision. There have been emerging studies on combining audio and visual information for applications such as sound source separation [14, 30], sound source localization [54, 5] and audio-visual event localization [42, 48].

Audio-Visual Scene-Aware Dialog. Audio-visual scene-aware dialog is a recently proposed multi-modal task that combines the previously mentioned research fields. The AVSD dataset introduced in [18] and [1] contains videos with audio streams and a corresponding sequence of question-answer pairs. Hori *et al.* [18] enhance the quality of generated dialog about videos using multi-modal features. Schwartz *et al.* [35] adopt the multi-modal attention mechanism to extract useful features for audio-visual scene-aware dialog task.

Cooperative Agents and Reasoning. Das *et al.* [11] are the first to propose goal-driven training for dialog agents, during which two interactive dialog agents are also involved to select an unseen image. Wu *et al.* [46] propose to generate reasoned dialog via adversarial learning. An information theoretic algorithm for goal-oriented dialog is then introduced in [22] to help the question generation. Unlike the previous work that has separate agents for questioner and answerer, Massiceti *et al.* [28] propose to use a single generative model for both roles. The reasoning process is essential in developing such intelligent agents. The idea of multi-step reasoning in the field of VQA has been exploited in [46, 13, 51, 19, 40].

Our work is related to the works mentioned above, yet differentiates from them in multiple aspects, including the task setup and formulation with implicit information sources, the QA-Cooperative network design, and the cooperative training method.

Table 1. Notations for the video description task.

s - Video description	V_s - start static frame of the video
$\mathcal{S}$ - Vocabulary	V_e - end static frame of the video
$i(i \leq 10)$ - Question-Answer round	$x_{A,i}$ - input for A-BOT at round i
A - Audio data	$x_{Q,i}$ - input for Q-BOT at round i
V_A - Video data for A-BOT	r_m - original data embedding for modal m
C - Video caption	a_m - attended data embedding for modal m
H_{i-1} - Existing dialog history at round i	d_m - dimension of the embedding for modal m
p_i - i-th pair of question-answer	$n_{\{C,H,q,a\}}$ - length of textual sequence
q_i - i-th question	m - modality notation, specified in context
a_i - i-th answer	$m \in \{A, V, C, H, q, a\}$

3 Video Description via Cooperative Agents

In this section, we respectively introduce the QA-Cooperative network for the two dialog agents, each component of the proposed network and the cooperative learning method with a dynamic dialog history update mechanism.

Notations used in our task formulation is presented in Table 1. In this video description task, we expect Q-BOT to describe an unseen video with a sentence $s = (s_1, s_2, ..., s_n)$ in n words after 10 rounds of question-answer interactions, each word s_k arises from a vocabulary $\mathcal{S}$. At i-th round of question-answer interaction, A-BOT takes the audio signals, video data, video caption and the existing dialog history as input, $x_{A,i} = (A, V_A, C, H_{i-1})$, with $H_{i-1} = \{p_1, ..., p_{i-1}\}$ and $p_{i-1} = (q_{i-1}, a_{i-1})$. For Q-BOT at the same round i, $x_{Q,i} = (V_s, V_e, H_{i-1})$. The final description task for Q-BOT is formulated as the inference in a recurrent model with the joint probability given by:

$$p(s|x_Q) = \prod_{k=1}^{n} p(s_k|s_{<k}, x_Q),$$
(1)

where we maximize the product of conditionals for each word in description s, given the input at 10-th round x_Q.

3.1 QA-Cooperative Network

The overall architecture of QA-Cooperative network is presented in Fig. 3. Q-BOT consists of a visual module, a history encoder, a visual LSTM-net, a multimodal attention module, a question decoder and the final description generator. A-BOT has an audio module, a visual module, a caption encoder, a history encoder, two attention modules and an answer decoder. In general, Q-BOT generates i-th question q_i based on the input $x_{Q,i}$, A-BOT responds to the question by generating the corresponding answer a_i. The new question-answer pair $p_i = (q_i, a_i)$ is used to update the existing dialog history. Note that we observe from the experiments that separate history decoders for two agents do not help improve the performance. Therefore, we choose the shared history encoder design to reduce the network redundancy.
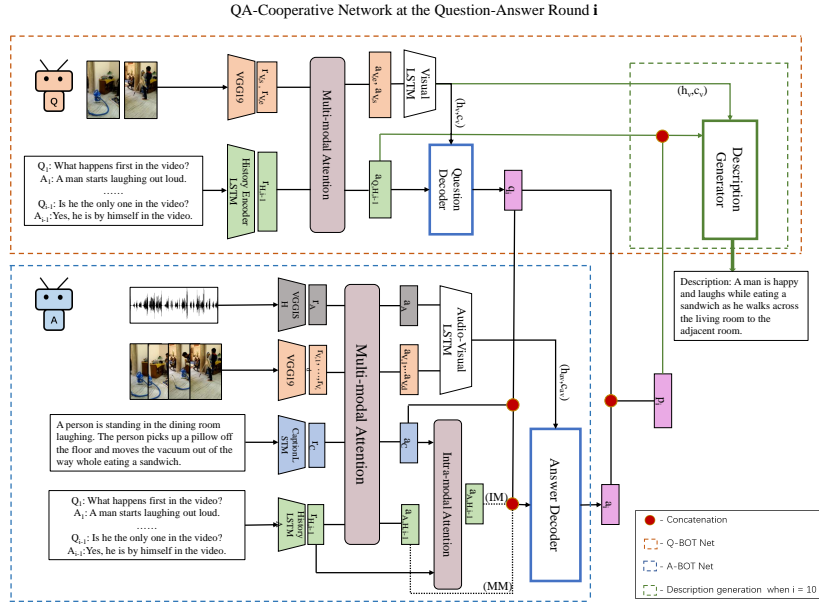
Fig. 3. QA-Cooperative Network at the question-answer interaction round i. Details about the network architecture and learning method are presented in Sec. 3.1, Sec. 3.2 and Sec. 3.3.

3.2   Model Components

There are multiple components in our QA-Cooperative network, which will be explained in detail in this section. We consider the situation at the i-th round of question-answer interaction.

Caption Encoder. The caption encoder contains a linear layer and a single layer LSTM-net. We firstly represent each word in the captions with one-hot vectors. Next, we find the longest caption sentence in each batch and zero-pad other shorter ones. The final caption embedding $r_C \in \mathbb{R}^{n_C \times d_C}$ is obtained from the last hidden state of the LSTM-net. This component is designed for $A\text{-}BOT$ to encode video captions during the preparation phase.

History Encoder. The history encoder contains a linear layer and an LSTM-net. Similarly to the processing steps for the video captions, we start with a list of one-hot word representations for a pair of question-answer. The longest question-answer pair of length is selected, the other pairs are zero-padded to fit the maximum length. The LSTM-net is used to obtain the pair-level embedding $r_{H,i-1} \in \mathbb{R}^{n_T \times n_H \times d_H}$. n_T here denotes the number of question-answer pairs in the dialog history (*i.e.*, $i - 1$). This encoder is a common component for both $Q\text{-}BOT$ and $A\text{-}BOT$, since the dialog history is visible to both agents.

Audio-visual LSTM. It is an LSTM-net with $d + 1$ units, where d is the number of visual frames visible to $A\text{-}BOT$. It takes the attended audio embedding

$a_A \in \mathbb{R}^{d_A}$ and $a_{V,j} \in \mathbb{R}^{d_V}$ with $j = \{1,...,d\}$ as input, the context vector (h_{av}, c_{av}) generated from this LSTM-net is used as the initial states input to the answer decoder. This component is used by $A\text{-}BOT$ to process the audio and visual information in addition to the cross-modal attention.

Visual LSTM. Similar to the audio-visual LSTM, this component takes the attended visual embedding $a_{V,s} \in \mathbb{R}^{d_V}$ and $a_{V,e} \in \mathbb{R}^{d_V}$ as input, the context vector (h_v, c_v) from this LSTM-net is used as the initial states for the question decoder and the final description generator. It is used by $Q\text{-}BOT$ to summarize the visual information from the initial two static frames.

Question Decoder. The question decoder is formed by an LSTM-net. It takes the attended history embedding $a_{Q,H,i-1} \in \mathbb{R}^{d_H}$ as input, with initial state $(h_0, c_0) = (h_v, c_v)$. The question generator generates the new question q_i that imitates the i-th question in the ground truth dialog.

Answer Decoder. Similar to the question decoder, we use the answer decoder to generate the answer embedding close to the i-th answer in the ground truth dialog. The answer LSTM decoder takes the concatenation of the attended history embedding $a_{A,H,i-1} \in \mathbb{R}^{d_H}$, the attended caption embedding $a_c \in \mathbb{R}^{d_C}$ and the newly generated question embedding $r_{q,i}$ as input, with initial state $(h_0, c_0) = (h_{av}, c_{av})$. The output is the answer a_i for the given question. The newly generated question-answer pair at i-th round is obtained by combining the i-th question and answer.

Description Generator. This LSTM generator generates the final description s for the unseen video based on 10 rounds of question-answer interactions history and the two static frames given in the first phase. When $i = 10$, the generator computes the following conditional probabilities based on the input, which is the attended history embedding $a_{A,H,10} \in \mathbb{R}^{d_H}$ including 10 rounds of question-answer interactions:

$$p(s_k|s_{k-1}, h_{k-1}, x_Q) = g(s_k, s_{k-1}, h_{k-1}, x_Q), \qquad (2)$$

where h_{k-1} is the hidden states obtained from the previous $k-1$ step. Note that h here is the hidden states of LSTM-net, differnt from the history notation H. The initial state is the same as the question decoder, thus we have $(h_0, c_0) = (h_v, c_v)$. the LSTM-net g predicts the probability distribution $p(s_k|s_{k-1}, h_{k-1}, x_Q)$ over words $s_k \in \mathcal{S}_k$, conditioned on the previous words s_{k-1}. The final probability distribution for natural language description is obtained by transforming the output of the LSTM-net by a FC-layer and a Softmax.

Attention module. Since the dialog is a key information source in our task, we propose two different attention mechanisms for processing the information contained in the dialog history: The multi-modal (MM) attention among visual, audio and textual modalities, and the intra-modal (IM) attention between dialog history and another textual sequence.

For the MM attention, we use the factor graph attention mechanism proposed in [36]. For $A\text{-}BOT$, this multi-modal attention module takes the audio embedding r_A, visual embedding $r_{V,j}$ with $j = \{1,...,d\}$, caption embedding r_C and the history embedding $r_{H,i-1}$ as input. Each visual frame is treated as an

individual modality as in [36]. The output of this multi-modal attention module are the attended audio embedding a_A, the attended visual embedding $a_{V,j}$ with $j = \{1, ..., d\}$, and the attended history embedding $a_{Q,H,i-1}$. Similarly for Q-BOT, we have the attended output $a_{V,s}$, $a_{V,e}$ and $a_{A,H,i-1}$ after taking $r_{V,s}$, $r_{V,e}$ and $r_{H,i-1}$ as input. Note that the history embedding $r_{H,i-1}$ before the multi-modal attention module is the same for Q-BOT and A-BOT because a shared history encoder is used, but the attended history embedding becomes different due to different inputs for two agents.

For the IM attention module in Fig. 3, it is a simple softmax attention between the dialog history $r_{H,i-1}$ and the concatenation of a_C and $r_{q,i}$.

3.3   Cooperative Learning

We propose to learn the QA-Cooperative network with a dynamic dialog history update mechanism in a goal-driven manner considering the following two aspects.

Firstly, the ultimate objective of our task is for Q-BOT to describe what happens between the beginning and the end of the video in a concrete way. Considering the fact that only two static frames are given to Q-BOT during the first preparation phase, the principal information source for Q-BOT is the dialog with A-BOT. Therefore, dialog history is the key to effectively learning our QA-Cooperative network. To this end, we propose a dynamic dialog history update mechanism to help with the knowledge transfer from A-BOT to Q-BOT. Fig. 4 illustrates the dialog history update operation by fusion. Notably, to emphasize the information from the newly generated question-answer pair, we set the dimension of the history embedding equal to the dimension of $r_{p,i}$ before the attention module.

Fig. 4. Dialog history update. Specifically, we maintain the history embedding dimension equal to the dimension of $r_{p,i}$ to emphasize the information from the newly generated question-answer pair.

Secondly, for the internal question and answer generation process, we encourage the two agents to imitate the questions and answers from the corresponding rounds of the ground truth dialog. Intuitively, this generation process can also be realized by pre-trained question and answer decoders using the existing methods [35, 18, 20, 46] trained in traditional VQA tasks. However, because the pre-training process is directly optimized to generate questions or answers based on the ground truth dialog, it sets the barrier for the final performance in our video description task. In other words, Q-BOT is unlikely to surpass the performance obtained in the situation where full ground truth dialog is provided. As for comparisons, the goal-driven training [11] is optimized in the final description phase. Although we also use the ground truth dialog as the internal imitation
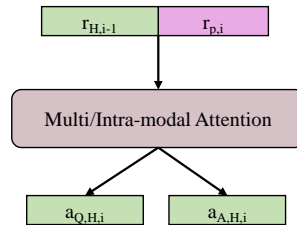
reference for two agents, it leaves more space for flexible question and answer generations (*i.e.*, to ask and answer the questions that directly help Q-BOT to better describe the video in the last phase). Our experiments support this assumption in Sec. 4.3.

Fig. 3 schematically illustrates the general learning process at the i-th question-answer round. Q-BOT takes the two static frames and the existing dialog history that consists of $i-1$ pairs question-answer as input, $x_{Q,i} = (V_s, V_e, H_{i-1})$. The visual and history embedding $r_{V,s}$, $r_{V,e}$ and $r_{H,r-1}$ are processed in the MM attention module [36] to acquire the attended embedding $a_{V,s}$, $a_{V,e}$ and $a_{Q,H,r-1}$. Given (h_v, c_v) from the visual LSTM as the initial state, the question decoder outputs the question q_i with its embedding $r_{q,i}$ after taking $a_{Q,H,r-1}$ as input. In the meanwhile, A-BOT takes the audio signals, video frames, video caption and the existing dialog history as input, $x_{A,i} = (A, V_A, C, H_{i-1})$. The corresponding attended embedding a_A, $a_{V,j}$ with $j = 1, ..., d$, a_C and $a_{A,H,i-1}$ is obtained after the MM attention. Another way to obtain the $a_{A,H,i-1}$ is through the IM attention module. The input for answer decoder is the concatenation of the attended embedding $a_{A,H,i-1}$, a_C and $r_{q,i}$, with initial state $(h_0, c_0) = (h_{av}, c_{av})$ generated from the audio-visual LSTM. The generated question and answer embedding are combined to form the new i-th question-answer pair $r_{p,i}$, and used to update the existing history embedding $r_{H,i-1}$ by fusion.

4 Experiments

4.1 Dataset

We use the recent AVSD v0.1 dataset [18] for experiments. The AVSD dataset consists of annotated dialog about 9848 short videos taken from CHARADES [39]. The dialog collection process resembles our task setup, during which two Amazon Mechanical Turk (AMT) workers play the roles of Questioner and Answerer. The Questioner was shown only the first, middle and last static frames of the video, while the Answerer had already watched the entire video, including the audio stream and the original video caption. After having a conversation about the events that happened between the frames through 10 rounds of question-answer interactions, the Questioner is asked to summarize the entire video.

The current AVSD v0.1 is split into 7659 training, 1787 validation and 1710 testing dialog, respectively. Hori *et al.* [18] also propose a 'prototype validation-set and test-set', which are sub-splits of the original validation set since the original test set does not include ground truth dialog at the moment. Our experiments are conducted on the original training and 'prototype' validation-test splits of the AVSD dataset.

4.2 Implementation Details

Data Representations. Our cooperative dialog agents have data input of visual, audio and textual modalities. For the visual modal, we take the video

representations extracted from the last conv layer of a VGG19 as input. We sample 4 equally spaced frames from the beginning of the original video, and each frame representation is of dimension $7 \times 7 \times 512$. The spatial and visual embedding dimensions are 49 and 512, respectively. $Q\text{-}BOT$ is shown the first and the last frames, while the $A\text{-}BOT$ is able to see all the frames. For the audio modal, we obtain the 256-dim audio feature via VGGish [17]. For the textual representations, we extract the language embedding from the last hidden state of their corresponding LSTM-nets. The dimensions are $d_C = 256$, $d_q = 128$, $d_a = 128$ and $d_H = 256$.

Network Training. The proposed QA-Cooperative network is trained using a cross-entropy loss on the probabilities $p(s_k|s_{<k}, x_Q)$ on the final video descriptions. All the components are jointly learned. The total amount of trainable parameters is approximately 19M. We use the Adam optimizer with a learning rate of 0.001 and a batch size of 64 for training. During the training, the perplexity metric is used for evaluating the performance on the validation set.

4.3   Cooperative Video Description

We split our experiments into two major groups to provide a more comprehensive and objective analysis of our work as shown in Table 2, which presents the quantitative results using the BLEU1-4 [31], METEOR [7], SPICE [2], ROUGE_L [23], and CIDEr [43] as the evaluation metrics.

Standard Test Setting. During our test, the performance of $Q\text{-}BOT$ is evaluated at each question-answer round-level. In other words, for a given video in the test split, the start question-answer round number i ranges from 1 to 10. For example, if the start round number $i = 1$, then no existing dialog history is given to $Q\text{-}BOT$ and $A\text{-}BOT$, they will generate all the ten questions and answers by themselves. However, if the start round number $i = 6$, then 5 rounds of question-answer pairs are given to two agents as the existing history, in which case, $Q\text{-}BOT$ still has 5 opportunities to freely ask questions. For a given video, tests with different start round numbers are independent, representing 10 different test cases. Therefore, for the 733 videos from the 'prototype test set' of the AVSD dataset [18], we have in total 7330 different test cases. We refer this testing process as the standard test setting, it is consistent with the learning process explained above in the Sec. 3.3. The only exception is the strong baseline situations for $Q\text{-}BOT$, during which the full ground truth dialog history is given before the testing. For the strong baseline situation, since the evaluation is only conducted at the end of full dialog history, there will be 733 testing cases.

Description Ability for $A\text{-}BOT$. The first group of experiments focuses on the video description ability of $A\text{-}BOT$, which is used as a comparison and performance reference for $Q\text{-}BOT$. Under our task setup, $A\text{-}BOT$ has full access to all the information while $Q\text{-}BOT$ only sees two static images. Intuitively, $A\text{-}BOT$ should have better performance than $Q\text{-}BOT$. We also compare the performance of our $A\text{-}BOT$ with the other two recent $A\text{-}BOT$ baselines from [18, 35]. The models proposed in [18, 35] are initially designed for question answering, but they also take all modalities of data as input, therefore, we modify the generators to

Table 2. Quantitative experimental results of video description tasks by different agents using multiple methods. *HIS Att* stands for *History attention*. The experiments are split into two groups, one group for *A-BOT*, and another group for *Q-BOT*. The comparisons between *A-BOT* and basic *Q-BOT* baselines show the actual knowledge gap between the two agents. We obverse that both *A-BOT* and *Q-BOT* from the proposed QA-Cooperative network achieve the best performance. Notably, our *Q-BOT* with cooperative learning is able to achieve comparable performance with the strong baseline, during which the full ground truth dialog history is given as input. It demonstrates the effective knowledge transfer process.

Agent	Method	HIS Att	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	SPICE	ROUGE_L	CIDEr
A-BOT (with info from all modalities)	Hori *et al.* [18]	-	34.2	17.1	8.4	4.8	11.5	11.4	24.9	20.7
	S. *et al.* [35]	-	32.1	16.2	8.7	5.1	12.1	11.6	27.6	21.6
	S. *et al.* [35]	IM	33.8	16.9	9.1	5.3	12.7	11.8	27.7	22.7
	S. *et al.* [35]	MM	33.8	17.6	9.9	5.9	12.9	13.5	28.5	25.6
	Ours	IM	**37.9**	**21.6**	12.5	7.6	**15.2**	**18.5**	31.1	38.1
	Ours	MM	37.5	21.5	**12.9**	**8.2**	**15.2**	17.9	**31.2**	**39.3**
Q-BOT Basic baselines	Ours	IM	32.0	15.7	8.1	4.7	11.6	11.1	26.4	18.3
	Ours	MM	33.2	16.4	8.6	5.0	12.5	11.5	27.1	20.2
Q-BOT Strong baselines	Ours (full GT HIS)	IM	33.3	17.0	9.1	5.4	12.6	11.7	27.3	21.3
	Ours (full GT HIS)	MM	32.7	17.2	**9.7**	**6.0**	12.6	**13.6**	**27.9**	**26.3**
Q-BOT Cooperative	Ours (pre-trained)	MM	33.3	17.0	9.2	5.4	**12.9**	11.4	27.4	21.5
	Ours QA-C	IM	33.3	16.9	9.1	5.3	12.7	11.6	27.7	22.7
	Ours QA-C	MM	**33.3**	**17.3**	9.5	5.5	12.8	12.4	**27.9**	23.1

generate video descriptions after 10 rounds of question-answer interactions. We observe that our *A-BOT* largely outperforms the baseline models, mainly with the help of separate caption encoder and the attention operation on the dialog history, while the previous works directly incorporate the video captions into the dialog history. Different from [35] where the usage of attention for dialog history does not yield improvements for the classic question answering task, we find it is helpful in improving the performance for our video description task either by MM or IM attention method.

Description Ability for *Q-BOT*. The second group of experiments focuses on the video description ability of *Q-BOT*. The basic baseline for *Q-BOT* is obtained under the standard test setting without cooperative learning, in other words, *Q-BOT* does not involve question-answer interactions with *A-BOT*. The comparisons between the basic baselines and *A-BOT* reveal the actual knowledge gap between the two agents. The strong *Q-BOT* baselines are obtained by providing *Q-BOT* with the full ground truth dialog history that contains 10 rounds of question-answer pairs. We also compare our cooperative learning method with the pre-trained methods mentioned in Sec. 3.3. The experimental results are consistent with our expectations. Although outperforming the basic baseline, the pre-trained learning method can hardly surpass two strong baselines. As for comparisons, our *Q-BOT* from the proposed QA-Cooperative network with the cooperative learning method is able to achieve more promising performance. It outperforms the strong baseline with IM attention setting, and
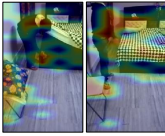
| Input of Q-BOT | Video descriptions | Ground Truth Dialog | Generated Dialog |

Q1: whats going on in this video?
A1: the guy hits his bed frame and then throws a pillow from a chair and then the video ends.
Q2: is this a male or female?
A2: the guy in the video is male.
Q3:is there anybody else with him?
A3: there is nobody with him.

Human(GT): a young man in his early 20s hits his bed frame with his hand. picks up some pillow from a nearby chair and throws in on the bed.

Q-Basic: a man walks into a room with a broom. he then picks up a book and walks out of the room.

Q-Strong: a man walks into a room and sits on a chair. he looks at his phone. then he picks up a book and leaves the room.

A-BOT: a man is sitting on a bed. he wakes up and throws a pillow on the bed. he then walks out of the room.

Ours: a man walks into a room and sits down on a couch , he picks up a pillow and throws it on the floor.

Q4: is there sound to the video?
A4: yes there is sound. nothing important.
Q5: so what does he hit his bed frame with?
A5: he hits the bedframe with his hand.
Q6: is he making his bed?
A6: the bed is already made.
Q7: does he pick anything up besides the pillow?
A7: he throws a pillow onto the bed from the chair.
Q8: do u see the expression on his face?
A8: there are no expression from the males face.
Q9: how old would u say the man is?
A9: the man is probably in his 20s.
Q10: is the room well lit?
A10: the room is well lit with some incandescent lightbulbs.

Q4: does the person chat ?
A4: no, nothing.
Q5: do you see anything ?
A5: socks and chatting .
Q6: do you see sports ?
A6: no, nothing.
Q7: do you see sitting ?
A7: yes.
Q8: do you see sitting ?
A8: yes.
Q9: do you see anything prepared ?
A9: no, i don't think so.
Q10: do you see anything hoping ?
A10: yes he does.

Fig. 5. Example of qualitative results. Our *Q-BOT* is able to describe more details about the unseen video. More qualitative results in the supplementary material.

achieves comparable performance with the strong baseline under MM attention setting. The experimental results demonstrate that cooperative learning helps *Q-BOT* to extract even more useful information to better describe the unseen video. Moreover, we find the MM attention for the dialog history is very helpful in improving the performance of the strong baseline settings.

Examples of qualitative results are shown in Fig. 5. We observe that our *Q-BOT* with cooperative learning is able to describe the unseen videos with more concrete details, achieving comparable performance with the strong baseline. More qualitative results and analysis can found in the supplementary material.

4.4   Ablation Studies

More experimental results on ablation studies are presented in Table 3. All the performance reported is achieved by *Q-BOT* under the standard test setting.

Full Update vs. Partial Update. In the proposed cooperative learning, the newly generated pair $r_{p,i}$ is fused with the existing dialog history $r_{H,i-1}$ before the attention module as shown in Fig. 4. With the updated history embedding, other outputs from the attention modules also become different compared to the previous dialog round. We compare the performance between the case of full update and partial update. In the full update setting, all the outputs from the attention modules are updated, while the partial update setting only updates the attended dialog history. The full update setting is more competitive than the partial one.

Initial States. In our proposed QA-Cooperative network, the initial states of the question decoder and the answer decoder are obtained from the visual LSTM and audio-visual LSTM, respectively. These two LSTM-nets summarize the audio and visual information for two agents. We observe from Table 3 that the initial states help achieve better performance under the full update setting, but have less impact on the partial update setting. It is reasonable because the full update setting mainly changes the attended audio and visual information, which is later used as the initial states.

Table 3. Quantitative experimental results on ablation studies. All the results are obtained by *Q-BOT* under the standard task setup as explained in Fig. 1. *init.* means the initial states provided by the visual and audio-visual LSTM for the question and answer decoder. *w/o his for A* means we remove the attended history embedding from the input for the answer decoder.

Settings	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	SPICE	ROUGE_L	CIDEr
parital	33.2	16.9	9.2	5.2	12.8	12.1	27.6	21.6
full w/o init.	33.1	16.3	8.6	4.9	12.2	11.6	27.1	20.3
partial w/o init.	32.7	16.2	8.8	5.1	12.1	11.8	27.4	20.3
w/o caption	31.5	15.3	7.9	4.6	12.7	11.1	26.3	19.5
w/o audio	33.2	17.2	9.4	5.4	12.8	12.2	27.8	22.3
w/o his for A	32.5	16.9	9.3	5.4	12.1	12.2	27.1	23.0
shuffled QA order	32.0	15.7	8.3	4.8	11.7	11.1	26.3	18.9
proposed QA-C	**33.3**	**17.3**	**9.5**	**5.5**	**12.8**	**12.4**	**27.9**	**23.1**

Modality Input. Our two agents, especially *A-BOT* take multiple modalities of information as input. We test the settings when the caption and the audio information are removed from the input of *A-BOT*. Experimental results show that the caption is of vital importance in achieving good performance for video description. Audio information also contributes to the better final performance. *w/o his for A* in Table 2 stands for the setting when the history embedding $a_{H,i-1}$ is removed from the input of answer decoder for *A-BOT*. It explains the fact that *A-BOT* does not heavily rely on the dialog history to provide answers to *Q-BOT* since it has already watched the entire video.

Order of Question-Answer Pairs. We also investigate the influence of the order of the question-answer pairs in the input. Similar to [1], the order of question-answer pairs is a significant factor for better final performance.

5   Conclusions

In summary, in this paper we propose a novel video description task via two multi-modal dialog agents, *Q-BOT* and *A-BOT*. We establish a new task setting for AI systems to accomplish a multi-modal task without direct access to the original visual or audio information. We further propose a QA-Cooperative network and a cooperative learning method with a dynamic dialog history update mechanism. Extensive experiments prove that *Q-BOT* is able to achieve very promising performance via our proposed network and learning method.

Acknowledgements

References

1. Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al.: Audio visual scene-aware dialog. In: CVPR (2019)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
5. Arandjelovic, R., Zisserman, A.: Objects that sound. In: ECCV (2018)
6. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
7. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
8. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR (2017)
9. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017)
10. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017)
11. Das, A., Kottur, S., Moura, J.M., Lee, S., Batra, D.: Learning cooperative visual dialog agents with deep reinforcement learning. In: ICCV (2017)
12. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: CVPR (2017)
13. Gan, Z., Cheng, Y., Kholy, A., Li, L., Liu, J., Gao, J.: Multi-step reasoning via recurrent dual attention for visual dialog. In: ACL (2019)
14. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: CVPR (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP. IEEE (2017)
18. Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A., et al.: End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In: ICASSP. IEEE (2019)
19. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. In: ICLR (2018)
20. Jain, U., Lazebnik, S., Schwing, A.G.: Two can play this game: visual dialog with discriminative question generation and answering. In: CVPR (2018)

21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
22. Lee, S.W., Heo, Y.J., Zhang, B.T.: Answerer in questioner's mind: information theoretic approach to goal-oriented visual dialog. In: NeurIPS (2018)
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
25. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR (2017)
26. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NeurIPS (2016)
27. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: CVPR (2017)
28. Massiceti, D., Siddharth, N., Dokania, P.K., Torr, P.H.: Flipdial: A generative model for two-way visual dialogue. In: CVPR (2018)
29. Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., Wen, J.R.: Recursive visual attention in visual dialog. In: CVPR (2019)
30. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
32. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
33. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)
34. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: NeurIPS (2017)
35. Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: CVPR (2019)
36. Schwartz, I., Yu, S., Hazan, T., Schwing, A.G.: Factor graph attention. In: CVPR (2019)
37. Seo, P.H., Lehrmann, A., Han, B., Sigal, L.: Visual reference resolution using attention memory for visual dialog. In: NeurIPS (2017)
38. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: CVPR (2016)
39. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV. Springer (2016)
40. Song, X., Shi, Y., Chen, X., Han, Y.: Explore multi-step reasoning in video question answering. In: ACM Multimedia (2018)
41. Szegedy, C., Toshev, A., Erhan, D.: Deep neural networks for object detection. In: Advances in neural information processing systems. pp. 2553–2561 (2013)
42. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
43. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
44. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: CVPR (2018)

45. Wang, L., Schwing, A., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In: NeurIPS (2017)
46. Wu, Q., Wang, P., Shen, C., Reid, I., Van Den Hengel, A.: Are you talking to me? reasoned visual dialog generation through adversarial learning. In: CVPR (2018)
47. Wu, Y., Zhu, L., Jiang, L., Yang, Y.: Decoupled novel object captioner. In: ACM Multimedia (2018)
48. Wu, Y., Zhu, L., Yan, Y., Yang, Y.: Dual attention matching for audio-visual event localization. In: ICCV (2019)
49. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: ICML (2016)
50. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV. Springer (2016)
51. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)
52. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: ICCV (2017)
53. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016)
54. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: ECCV (2018)