

# Saying the Unseen: Video Descriptions via Dialog Agents

Ye Zhu, Yu Wu, Yi Yang, and Yan Yan\*

**Abstract**—Current vision and language tasks usually take complete visual data (e.g., raw images or videos) as input, however, practical scenarios may often consist the situations where part of the visual information becomes inaccessible due to various reasons e.g., restricted view with fixed camera or intentional vision block for security concerns. As a step towards the more practical application scenarios, we introduce a novel task that aims to describe a video using the natural language dialog between two agents as a supplementary information source given incomplete visual data. Different from most existing vision-language tasks where AI systems have full access to images or video clips, which may reveal sensitive information such as recognizable human faces or voices, we intentionally limit the visual input for AI systems and seek a more secure and transparent information medium, i.e., the natural language dialog, to supplement the missing visual information. Specifically, one of the intelligent agents - *Q-BOT* - is given two semantic segmented 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 access to the entire video, assists *Q-BOT* to accomplish the goal by answering the asked questions. We introduce two different experimental settings with either a generative (i.e., agents generate questions and answers freely) or a discriminative (i.e., agents select the questions and answers from candidates) internal dialog generation process. With the proposed unified QA-Cooperative networks, we experimentally demonstrate the knowledge transfer process between the two dialog agents and the effectiveness of using the natural language dialog as a supplement for incomplete implicit visions.

**Index Terms**—Video Description, Dialog Agents, Multi-modal Learning.

## 1 INTRODUCTION

CLASSIC vision-language tasks such as video captioning and visual question answering (VQA) have been well explored in previous work [3], [6], [14], [16], [55], [60], [61], [80] and have achieved promising performance. Most existing research studies on these tasks provide AI systems with full access to images or videos. However, these images or videos may reveal sensitive personal biometric information (e.g., recognizable human faces or voices), thus aggravating the arising concerns on the privacy and security issues of AI from the general public in recent years. Although directly taking the original visual data as input usually helps with the performance improvement, we also observe that it is not always necessary to fulfill the final task in practical scenarios (e.g., we do not need to directly look at the human faces to tell their actions or gestures). In addition, in more practical application scenarios, we may encounter the situations where part of the visual information is inaccessible due to reasons such as restricted view of fixed cameras. Based on the above observations, we make efforts to introduce a new multi-agent task that aims to describe a video based on implicit visions in this work. The concept of implicit vision refers to the idea that the given visual information is intentionally made incomplete to protect user privacy. We then propose to supplement the missing visual information via a less sensitive information medium, i.e., the natural language

dialog. Unlike video clips, AI systems, or even humans, can hardly identify the biometric information of a person based on the natural language descriptions from the dialog. In addition, natural language dialog is more transparent for humans in the sense that humans can understand and interpret the sentences compared to traditional obscure feature embedding in matrix forms. Overall, we have two objectives to fulfill in this work: to propose a novel video description task setup that addresses the privacy concerns by providing implicit visual data, and to demonstrate that the natural language dialog can be a more secure yet effective source to supplement the missing visual information.

Our task involves two agents, the questioning robot *Q-BOT* and the answering robot *A-BOT*. In practical scenarios such as smart homes, *Q-BOT* could be the actual AI system, while *A-BOT* plays the role of human users. Human users can naturally perceive all the information sources and answer questions related to the surrounding environment. In contrast, *Q-BOT* (AI system) only has a sketchy perception of the general environment such that it will not see the entire home setting. The proposed task shares some similarities with the classic video captioning [39], the visual dialog task [14]. Video captioning aims to generate a natural language summary of the video based on direct visual input, and visual dialog aims to answer a series of questions related to the visual content in the form of a dialog. Our task includes both components but differentiates them from multiple aspects. Firstly, the task inputs and motivations are different. While the previous tasks take the original complete visual data as input and seek to achieve better performance, our work intentionally provides the AI model with implicit visual input to exploit a trade-off between the

- Ye Zhu and Yan Yan are with the Department of Computer Science, Illinois Institute of Technology, USA. (Corresponding Author: Yan Yan) E-mail: yzhu96@hawk.iit.edu, yyan34@iit.edu
- Yu Wu and Yi Yang are with University of Technology Sydney, Australia. E-mail: yu.wu-3@student.uts.edu.au, yi.yang@uts.edu.au
- Codes will be updated at: <https://github.com/L-YeZhu/Video-Description-via-Dialog-Agents-ECCV2020>

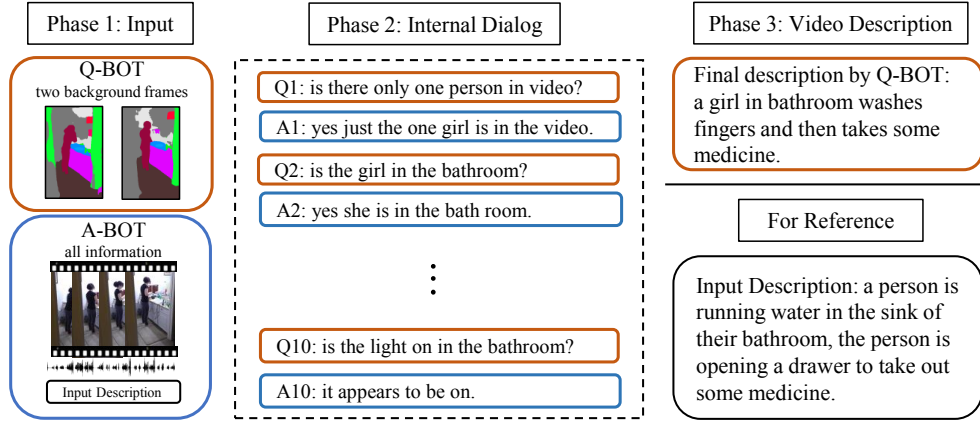


Fig. 1. Unseen video description task via interpretable knowledge transfer between dialog agents. The task setup includes three phases, and the ultimate goal is for *Q-BOT* to describe the unseen video mainly based on the dialog. The input description is also presented for reference. The difference between the input descriptions for *A-BOT* and the final descriptions given by *Q-BOT* reveals the actual knowledge gap due to the lack of direct access to the original video data.

performance and the visions. Secondly, the task goals are different. The original visual dialog task focuses on learning the AI systems to answer natural language questions. In contrast, our task emphasises to enable the AI models (*i.e.*, *Q-BOT*) to achieve a concrete vision-related goal (*i.e.*, video description) using the natural language dialog as a supplementary information source.

The concrete setup is illustrated in Figure 1, which resembles the data collection process of the AVSD dataset [3]. Initially, *Q-BOT* takes as input two semantic segmented frames (*i.e.*, semantic segmentation results of static video frames, thus no visible human faces) from the beginning and the end of the video. *A-BOT* has full access to the information of all modalities, including the entire video, audio stream, and the original video description sentences. Afterwards, *Q-BOT* has 10 chances to ask questions to collect necessary information of the video, and *A-BOT* collaboratively provides answers to the questions. After 10 rounds of dialog, *Q-BOT* is asked to summarize the unseen video based on the segmented visual input and the dialog history. Under our task setup, *Q-BOT* learns to accomplish the video description task without directly seeing the video.

There are two principal considerations behind our task formulation that using the dialog as the supplementary information medium, instead of directly asking for the final descriptions from human users (*i.e.*, *A-BOT*). *Firstly*, the overall descriptions directly given by humans are usually noisy and biased without given hints or templates in the sense that different humans may pay attentions to completely different parts given the same visual content [30], [58]. In contrast, the answers given by human users for specific questions are less biased. For a question like “*How many persons are there in the video?*”, we can expect the answer to be a specific number in most cases. *Secondly*, from a higher-level perspective, AI systems have different objectives in practical scenarios, the question-guided dialog interactions help AI systems to better extract the necessary information required for accomplishing specific downstream tasks. For example, human users may want to create a better sound experience in their living rooms via the smart home system,

which usually requires acoustic engineers to perform professional acoustic compensation based on the relative spatial relations among loudspeakers. For an acoustic expert, the process to acquire the spatial structure among loudspeakers is rather systematic via a succession of structured questions (*e.g.*, *How many loudspeakers do you have in the room? Are they placed in the corner close to the wall?*), while users may find it more challenging to directly provide the necessary spatial information. In this case, the AI system is expected to ask guided questions and to extract the necessary information from the answers provided. The above motivations inspire us to explore the possibilities of using the dialog as our primary choice for supplementing the insufficient visions.

One of the unique challenges in our task is the effective knowledge transfer process from *A-BOT* to *Q-BOT*. To better illustrate and clarify the knowledge transfer process, we introduce the concepts of *Input video descriptions* for *A-BOT* and *Final output video descriptions* by *Q-BOT* (referred as *Input Descriptions* and *Final Descriptions* in the remaining of the paper). The main difference between the two types of video descriptions lies within the fact whether the person/agent has seen the entire video before giving the description. The input descriptions example in Figure 1 contain more concrete details compared to the final descriptions. This fact demonstrates the knowledge and reasoning gap caused by the lack of direct access to the original video data, which also implies that although the natural language dialog could be an effective supplementary information source, it is rather challenging to completely alternate the incomplete visual information. One significant step that leads to the successful accomplishment of our task is to reduce this gap by an effective knowledge transfer process between the agents.

To accomplish the challenging task, we propose two different experimental settings with their corresponding unified QA-Cooperative networks. For the first experimental setting, *Q-BOT* and *A-BOT* generate questions and answers freely during the dialog interactions. We introduce a cooperative learning method with a dynamic dialog history update mechanism, which helps to transfer knowledge between the two agents effectively. Under this generative setting, we

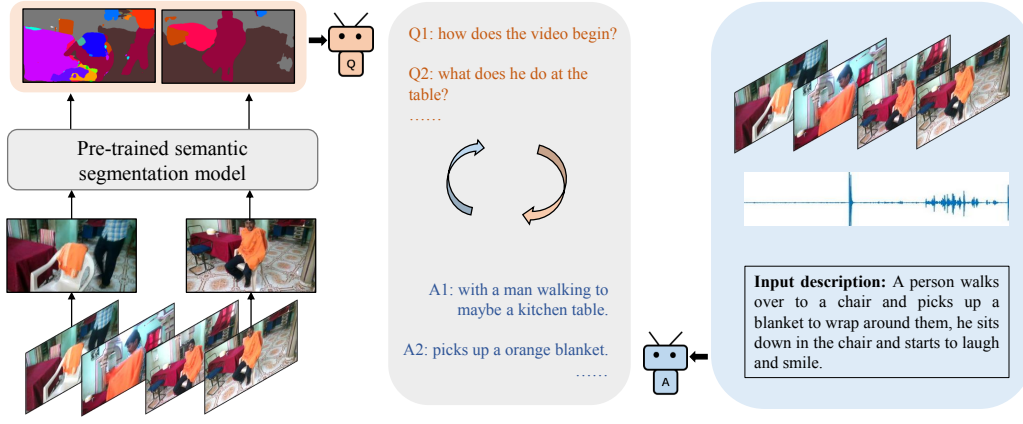


Fig. 2. Different inputs for Q-BOT and A-BOT in shaded orange and blue boxes, respectively. For Q-BOT, we extract the first and the last frames from the video clip and process the extracted frames using a pre-trained semantic segmentation model. We then input semantic segmented frames to Q-BOT, which only provides a sketchy perception of the general environment and does not reveal any sensitive information. In contrast, A-BOT has access to all the information, including entire videos, audio signals, and descriptions.

achieve promising performance and successfully transfer knowledge from A-BOT to Q-BOT. However, we also observe from the qualitative results that the generated internal dialog sometimes lacks clear logic and tends to be repetitive, which is a common issue in similar tasks [15]. We believe a meaningful and informative internal dialog is in line with our final objective to obtain a precise final description. Therefore, to further enhance the internal reasoning abilities and the interpretability of the dialog agents, we propose an improved experimental setting where agents proceed internal dialog in a discriminative way, meaning that Q-BOT and A-BOT pick questions and answers from the given candidates. We then introduce an improved version of the QA-Cooperative network, and propose a different learning method with an internal selective mechanism to enhance the interpretability and quality of the internal dialog. With the improved setting and model design, we make significant improvements in the effectiveness of the knowledge transfer, leading to better final descriptions. Through extensive experiments on the AVSD dataset [3], [27], we demonstrate: (a) An effective knowledge transfer process between two agents via the proposed QA-Cooperative networks and learning methods. (b) A meaningful and informative internal dialog indeed helps with our final objective to achieve better descriptions for unseen videos. (c) Multiple data modalities and proposed model components contribute to the final performance.

This paper is an extended work following [82]. Compared to the vanilla version [82], we incorporate a considerable amount of extension work from three aspects: the task setup, the methodology, and the experiments. *For the task setup:* (a) we modify the initial task setup from [82] to further enhance the security aspect of the task system. Specifically, the previous task setup of [82] allows Q-BOT to take two original RGB frames from the video as visual input. Although this task setup largely reduces the risk of exposing sensitive face images to AI systems, it can not ensure that the observed frames contain absolutely no biometric information. In this paper, we incorporate the semantic segmentation as pre-processing for Q-BOT as shown in Figure 2,

and thus resolving the previous concern. (b) we add an improved discriminative setting for the internal dialog generation process, which brings us more interpretability for the internal reasoning process, as well as the improvement for the final descriptions. *For the methodology:* (a) we propose an enhanced network architecture with modified question and answer decoders for the discriminative setting, which aims to enhance the quality and the interpretability of the generated dialog. (b) we incorporate an adapted internal selective mechanism from [43] for improved cooperative training that leads to better knowledge transfer and final performance. *For the experiments:* (a) we perform additional extensive quantitative and qualitative analysis for both the final descriptions and the dialog, which better interpret the internal process. (b) we conduct an simulated human test to evaluate the ability of A-BOT. (c) we achieve significant performance improvements for the final descriptions, raising the primary metric CIDEr [59] from 22.9 to 27.1.

Our overall contributions for this work can be summarized as follows:

- We propose a novel and challenging task that aims to describe an unseen video via two multi-modal dialog agents. The proposed task uses the natural language dialog as the supplementary information source for the incomplete visual input to address the potential privacy concerns.
- We introduce two QA-Cooperative framework designs for the internal dialog generation process. The proposed frameworks allow the two agents to fulfill the objective of unseen video description via a generative or discriminative internal dialog.
- We conduct extensive experiments and analysis to show the effectiveness of the proposed methods for our novel task, achieving very competitive performance that beats multiple baselines. We also experimentally demonstrate the knowledge transfer process between two agents and the feasibility of using the natural language dialog as a supplementary for incomplete visual input.

## 2 RELATED WORK

### 2.1 Image and Video Captioning

Image and video captioning is a classic vision-language task that aims to textually describe the given image or video input. Previous work on image and video captioning usually provides the network models with direct access to original visual data [5], [11], [33], [60], [71], [74]. Rennie *et al.* [45] formulate the image captioning problem with reinforcement learning and optimize the problem using the self-critical sequence training. A disentangled framework is proposed by Wu *et al.* [67] to generalize image captioning models to describe unseen objects for the zero-shot captioning task. Transformer-based [12] or attention-based [23], [40] methods have also been adopted to tackle the problem.

Although the output of image and video captioning tasks is also the textual descriptions, our task has a different formulation with its focus on the internal knowledge transfer process between two agents via natural language dialog.

### 2.2 VQA

Visual Question Answering (VQA) is another popular vision-language task that aims to answer natural language questions relevant to the given visual data [5], [6], [13], [34], [52], [55], [65], [69], [70], [72]. More recent research works in VQA starts to bring the causality theory into the field. Chen *et al.* [10] propose a model-agnostic Counterfactual Samples Synthesizing (CSS) training scheme for robust question answering. Agarwal *et al.* [2] propose to reveal and reduce the spurious correlations for VQA models to achieve more robustness. Efforts are also made to achieve better performance and more diversity using techniques such as variational auto-encoders [9], [62], attributes learning [73], reinforcement learning [15], [64] and pre-training [79].

Most existing models for VQA aim to answer the given questions about the visual content as the task objective, while our work has a concrete objective (*i.e.*, describing the unseen video) and uses the QA interactions as the medium for information transfer.

### 2.3 Visual Dialog

Unlike VQA that seeks to answer a single question, research works on visual dialog [13], [14], [16], [19], [29], [51], [64] extend the QA interactions into multiple rounds that form a complete and meaningful dialog with more internal logical relations. Several datasets have been collected [14], [16]. Most existing works in Visual Dialog emphasize the ability of AI to answer the questions, however, few researches have been done to exploit the other side of the problem, which is the question asking. Learning to ask meaningful and informative questions about the visual content is also an insightful research topic worth exploiting. Jain *et al.* [29] also look in to the problem of question asking. Qi *et al.* [43] exploit the causality effect for the visual dialog task and propose two causal principles for improving existing models. Guo *et al.* [22] introduce a Context-Aware Graph (CAG) neural network for the visual dialog task. Different attention mechanisms, such as the hierarchical attention [34], question-guided spatial attention [70], stacked

attention [72], multi-step reasoning [55], bottom-up and top-down attention [5] have also been exploited and proven to be effective. Agarwal *et al.* [1] recently study the role of history for visual dialog and reveal its potential shortcomings. Works that leverage the advantages of pre-trained language models such as BERT [17] and then fine-tuned for visual dialog have also been exploited in [36], [63].

Despite some similarities in the task setup, our work takes the incomplete visual data as input and uses the dialog interactions to supplement the missing information. In addition, we shift the model focus from answering the questions to question asking.

### 2.4 Audio Modality

Audio modality is another important source of information that has gained research popularity in recent years. There have been emerging studies on combining audio and visual information for various downstream tasks such as the sound source separation [7], [20], [38], [76], sound source localization [20], [37], [50] and audio-visual event localization [18], [46], [57], [66], [68], [81]. Hu *et al.* [28] introduce Deep Multimodal Clustering for capturing the audio-visual correspondence. Gao *et al.* [21] propose to recognize actions in untrimmed video using audio as a preview mechanism to eliminate visual redundancies.

Audio-visual scene-aware dialog (AVSD) [3], [27], [48] is another recently proposed task that resembles the visual dialog, it additionally incorporates audio signal compared to previous tasks and datasets [14], [16]. Hori *et al.* [27] firstly propose an end-to-end model using multimodal attention-based video features to tackle the task. Alamri *et al.* [3] further propose a benchmark for the AVSD task. While the AVSD task still focuses on answering questions, our work seeks to describe the entire video, which requires the model to further extract useful information from the dialog. In reference to the original AVSD dataset [3], the input descriptions proposed in our work correspond to the video captions given by a human annotator (the role of A-BOT in our task) after watching the entire video. In contrast, the final descriptions correspond to the video summaries given by human annotators (the role of Q-BOT in our task) without directly seeing the video.

Our work uses the AVSD dataset for experiments and shows that the audio data is also an important information source that contributes to better performance for our task.

### 2.5 Cooperative Agents

The research studies on dialog agents mainly have two categories. They either focus on maintaining a meaningful conversation [35], [64] or designing in a goal-driving manner to accomplish certain final objectives [15], [24], [31], [44], [53], [56], [75] (*e.g.*, retrieve the image that the dialog is about from candidates). Our work falls into the categories of the goal-oriented dialog systems. Early works about goal-oriented dialog agents usually focus on the single modality of natural language and formulate the problem using machine techniques such as Markov chain process [54] and probabilistic learning [47]. More recent works seek to incorporate data of other modalities into the framework. Das *et al.* [15] train the dialog agents with reinforcement learning

TABLE 1  
Notations for the unseen video description task.

$s$ - Final descriptions
$\mathcal{S}$ - Vocabulary
$i(i \leq 10)$ - Question-Answer round
$A$ - Audio data
$V_A$ - Video data for A-BOT
$C$ - Input video descriptions
$H_{i-1}$ - Existing dialog history at round $i$
$p_i$ - $i$ -th pair of question-answer
$q_i$ - $i$ -th question
$a_i$ - $i$ -th answer
$N_{\{q,a\}}$ - number of candidates
$V_s$ - start semantic segmented frame of the video
$V_e$ - end semantic segmented frame of the video
$x_{A,i}$ - input for A-BOT at round $i$
$x_{Q,i}$ - input for Q-BOT at round $i$
$r_m$ - original data embedding for modal $m$
$a_m$ - attended data embedding for modal $m$
$d_m$ - dimension of the embedding for modal $m$
$n_{\{C,H,q,a,s\}}$ - length of textual sequence
$m$ - modality notation, specified in context $m \in \{A, V, C, H, q, a\}$
$e_{\{q,a\}}$ - embedding vector of candidates
$h_{\{v,av\}}$ - hidden states of LSTM
$c_{\{v,av\}}$ - cell states of LSTM

to select dialog-related images. Guo *et al.* [24] also propose to optimize the interactive dialog for retrieve images using deep reinforcement learning with a user simulator. An information theoretic algorithm for goal-oriented dialog is introduced in [31] to assist the question generation.

The goal of our dialog system is to describe the unseen video. One of our challenges compared to the above works is the complexity of natural language descriptions, especially with incomplete visual input. Unlike the image retrieval task that aims to find the target image, the video descriptions are more various and difficult to evaluate.

### 3 VIDEO DESCRIPTION VIA DIALOG AGENTS

We firstly present task formulations in Section 3.1. The proposed QA-Cooperative networks for two settings are explained in Section 3.2. We then introduce their respective learning methods in Section 3.3. Notations used in our formulations are listed in Table 1.

#### 3.1 Task Formulation

##### 3.1.1 Video Description

For the proposed unseen video description task, our primary goal is for Q-BOT to describe an unseen video with a sentence  $s = (s_1, s_2, \dots, s_{n_s})$  in  $n_s$  words after 10 rounds of QA interactions. Each word  $s_k$  arises from a vocabulary  $\mathcal{S}$ . At  $i$ -th round of QA interaction, A-BOT takes the video data, audio signals, input description and the existing dialog history as input. Denote the input data to be  $x_{A,i} = (A, V_A, C, H_{i-1})$ , where  $A$  is the audio data,  $V_A$  is the complete video data,  $C$  is the input video descriptions, and the dialog history  $H_{i-1} = \{p_1, \dots, p_{i-1}\}$  with  $p_i$  to be the QA pairs  $p_i = (q_i, a_i)$ . For Q-BOT at the same round  $i$ , we extract the first and last frames from the video, and then perform semantic segmentation on these two frames using pre-trained models to obtain  $V_s$  and  $V_e$ . The semantic segmented frames  $V_s$  and  $V_e$  eliminate the possibilities to

reveal recognizable human faces to Q-BOT, as shown in Figure 2. The input data for Q-BOT is  $x_{Q,i} = (V_s, V_e, H_{i-1})$ . The final description task for Q-BOT is formulated as the inference in a recurrent model with the joint probability given by:

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

where we maximize the product of conditionals for each word in description  $s$ , given the input at 10-th round  $x_Q$ . From Eqn. (1), the core is how to generate a better dialog history  $H$  in  $x_Q$ . Next, we illustrate how to generate the internal dialog in two ways.

##### 3.1.2 Generative Dialog

One intuitive and straightforward way to formulate the internal dialog process is for both agents to directly generate the questions and answers. In this case, Q-BOT and A-BOT have the flexibility to freely ask questions and to provide answers. The generated questions and answers are formulated in a similar way as the final description. At  $i$ -th round of QA interactions, the  $i$ -th question  $q_i$  is given by:

$$p(q_i|x_{Q,i}) = \prod_{k=1}^{n_q} p(q_{i,k}|q_{i,<k}, x_{Q,i}), \quad (2)$$

where  $n_q$  is the number of words for the  $i$ -th question. Similarly for A-BOT, the  $i$ -th answer is generated following the same equation by replacing  $Q$  and  $q$  with  $A$  and  $a$ , respectively. Under this setting, the information is cooperatively exchanged through the dialog  $H$ . However, it is more challenging to guarantee the quality of the generated dialog due to the lack of supervision for the generation process.

##### 3.1.3 Discriminative Dialog

Another way to obtain the internal dialog is to provide possible candidates for Q-BOT and A-BOT to choose from. More specifically,  $q_i$  and  $a_i$  are picked from potential candidates  $\{q_i^1, q_i^2, \dots, q_i^{N_q}\}$  and  $\{a_i^1, a_i^2, \dots, a_i^{N_a}\}$  by Q-BOT and A-BOT, respectively. Those candidates are selected from the dataset. During inference, all the questions and answers from testing dialog are provided as candidates for Q-BOT and A-BOT. During training, we provide 100 questions and 100 answer candidates for each case. All the ground truth questions/answers, except those provided as input, are included in 100 candidates. Other candidates are randomly selected from the training set. Additionally, all the candidates are provided in pairs. In other words, for each question from the question candidates for Q-BOT, we also include its corresponding ground truth answer as an answer candidate for A-BOT. However, it should be mentioned that if a new question other than the ground truth ones is picked by Q-BOT, the picked question may not be valid (*i.e.*, the question may be irrelevant to the given video). In this case, there may not exist valid answers.

Considering the fact that comparing all the candidates at each QA round is very time-consuming during inference, we deploy a two-phase selection mechanism: cluster selection and question/answer selection. Q-BOT and A-BOT firstly select the pre-clustered question or answer type, and then pick the specific candidates from the previously chosen

cluster. For the question and answer types, we represent each question and answer from the testing set with the Glove embedding [42] and use the K-Means algorithm to cluster sentences into 10 classes. Specifically, we use the pre-trained Glove model to convert the each word from sentences into feature vectors and perform the clustering on the obtained sentence embeddings. During inference, the agent first picks a sentence cluster, and then further choose a concrete sentence within the cluster.

A discriminative setting for the internal dialog helps with alleviating the bias commonly existing in vision-language models and provides more interpretability for the internal dialog process. Overall, we experimentally demonstrate that both internal dialog settings are viable for our unseen video description task in Section 4. The generative setting may be more flexible for general deployment, while the discriminative setting leads to the stronger final performance and better reveals the internal reasoning process.

### 3.2 QA-Cooperative Networks

#### 3.2.1 Model Components

Our QA-Cooperative networks include multiple model components, which are presented in detail in this section. We focus on the situation at  $i$ -th round of dialog.

**Components of Q-BOT.** The *Q-BOT* contains a visual LSTM [26] that processes the input frames, a history encoder that gathers information from the dialog history, a question decoder for generating questions, and a description generator that finally describes the unseen video.

*Visual LSTM.* It is an LSTM with 2 units, 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 hidden states and cell states  $(h_v, c_v)$  from this LSTM is used as the initial states for the question decoder and the final description generator. *Q-BOT* uses this component to process the visual information from the two semantic segmented frames.

*History Encoder.* It consists of a linear layer and a single layer LSTM. We start with a list of one-hot word representations for a QA pair. The longest QA pair of length is selected, the other pairs are zero-padded to fit the maximum length. The LSTM is used to obtain the pair-level embedding  $r_{H,i-1} \in \mathbb{R}^{n_T \times n_H \times d_H}$ , where  $n_T$  denotes the number of QA pairs in the current dialog history (i.e.,  $i - 1$ ).

*Question Decoder.* The question decoder has different functions for generative and discriminative dialog settings. For the generative dialog setting, the question decoder is formed by an LSTM. It takes the attended history embedding  $a_{Q,H,i-1} \in \mathbb{R}^{d_H}$  as input, with initial states  $(h_0, c_0) = (h_v, c_v)$ . The question generator generates the new question  $q_i$  based on the  $i$ -th question in the ground truth dialog for the generative dialog setting. For the discriminative dialog setting, the question decoder is used to calculate the score based on the similarity between the question candidate embedding vector and the input for the question decoder  $Score_i^n = \langle e_{q_i^n}, a_{Q,H,i-1} \rangle$ , where the notation  $\langle \cdot \rangle$  denotes the inner product. The score computing is applied to each possible candidate. The scores are then processed by a Softmax operation to obtain the probabilities of all the candidates. The candidate with the highest probability is selected as the final  $q_i$ .

*Description Generator.* This LSTM generator generates the final description  $s$  for the unseen video based on 10 rounds of QA interaction history and the two semantic segmented 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 QA interactions:

$$p(s_k | s_{k-1}, h_{k-1}, x_Q) = g(s_k, s_{k-1}, h_{k-1}, x_Q), \quad (3)$$

where  $h_{k-1}$  is the hidden states from the previous  $k - 1$  step. The LSTM  $g$  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 the description is obtained by transforming the output of the LSTM by a FC-layer and a Softmax operation.

**Components of A-BOT.** The *A-BOT* contains an audio-visual LSTM that processes the audio and visual input, the same history encoder as *Q-BOT* that gathers information from the dialog history, an input description encoder that processes the input video descriptions, and an answer decoder used to generate answers.

*Audio-visual LSTM.* It is an LSTM with  $d + 1$  steps, where  $d$  is the number of visual frames visible to *A-BOT*. The extra one step is for processing the audio input. It takes the attended audio embedding  $a_A \in \mathbb{R}^{d_A}$  and the attended visual embedding  $a_{V,j} \in \mathbb{R}^{d_v}$  with  $j = \{1, \dots, d\}$  as input. The hidden states and cell states  $(h_{av}, c_{av})$  generated from this LSTM are used as the initial states input to the answer decoder. This component is for *A-BOT* to process the audio and visual information in addition to the cross-modal attention.

*Input Description Encoder.* It consists of the same structure as the history encoder with a linear layer and an LSTM. The input description embedding  $r_C \in \mathbb{R}^{n_C \times d_C}$  is obtained from the last hidden state of the LSTM. This component is designed for *A-BOT* to encode the input descriptions.

*History Encoder.* It is the same encoder as the one for *Q-BOT* since the history is a common input visible for both agents.

*Answer Decoder.* Similar to the question decoder for *Q-BOT*, this component is used to get the answer  $a_i$  for question  $q_i$ . The only difference is that this answer decoder takes the concatenation of the attended history embedding  $a_{A,H,i-1} \in \mathbb{R}^{d_H}$ , the attended input description 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 QA pair at  $i$ -th round is obtained by combining the  $i$ -th question and answer.

**Attention modules** Since the dialog is a key information source in our task to supplement the missing visual input, we propose two different attention mechanisms for processing the information contained in the dialog history: (1) the multi-modal (MM) attention among visual, audio, and textual modalities, and (2) the intra-modal (IM) attention between dialog history and another textual sequence.

*MM Attention.* We use the factor graph attention mechanism proposed in [49] for MM attention module. For *A-BOT*, this MM attention module takes the audio embedding  $r_A$ , visual embedding  $r_{V,j}$  with  $j = \{1, \dots, d\}$ , input description embedding  $r_C$  and the history embedding  $r_{H,i-1}$  as input. Each visual frame is treated as an individual modality as in [49]. The output of this multi-modal attention module



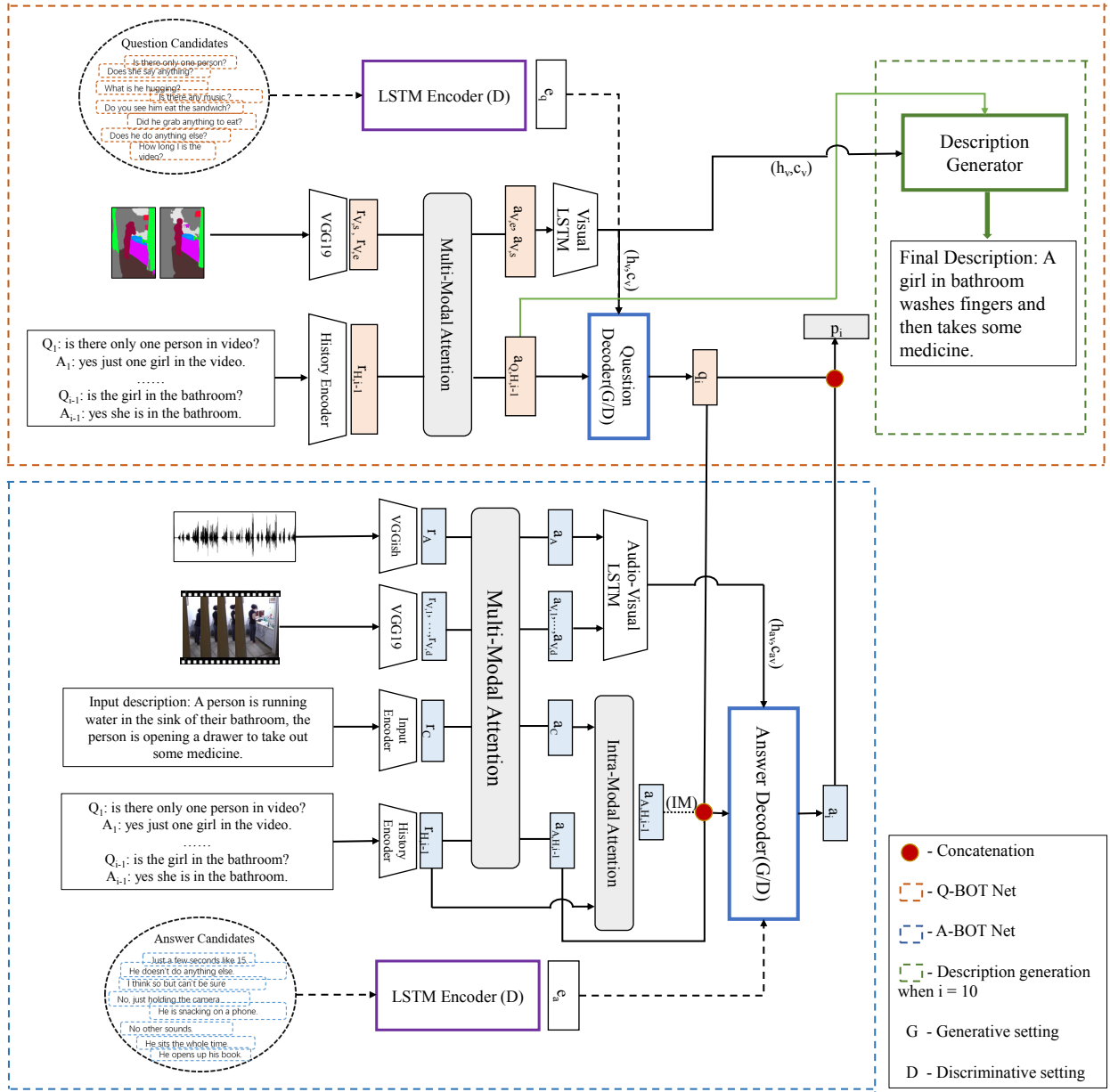


Fig. 3. QA-Cooperative network at QA round  $i$ . History dialog  $H$  is a common input for both agents. The model components for  $Q$ -BOT are in orange color boxes, while those for  $A$ -BOT are in blue. The dashed lines represent the processing of the question and answer candidates uniquely for the discriminative setting, while the solid lines are operations for both generative and discriminative settings.

are the attended audio embedding  $a_A$ , the attended visual embedding  $a_{V,j}$  with  $j = \{1, \dots, d\}$ , and the attended history embedding  $a_{Q,H,i-1}$ . Similarly for  $Q$ -BOT, we have the attended visual embedding  $a_{V,s}$ ,  $a_{V,e}$  and the attended history embedding  $a_{A,H,i-1}$  as output, after taking their original embedding  $r_{V,s}$ ,  $r_{V,e}$  and  $r_{H,i-1}$  as input. Note that the history embedding  $r_{H,i-1}$  before the MM attention module is the same for  $Q$ -BOT and  $A$ -BOT because of the shared history encoder, but the attended history embedding becomes different due to different inputs for two agents.

**IM Attention.** We adopt a softmax attention consisting of the concatenation and dot product operations between the dialog history embedding  $r_{H,i-1}$  and the question embedding  $r_{q,i}$  as the IM attention.

### 3.2.2 QA-Cooperative Framework

The architecture for our proposed QA-Cooperative networks is presented in Figure 3. The main difference in network architecture for two internal dialog settings is the design of question/answer encoder/decoder as explained in model components.

In general, the dialog history consisting  $i - 1$  QA pairs is a common input for both agents since it is visible to both agents in practical scenarios.  $Q$ -BOT processes the visual input (two semantic segmented frames) and the dialog history input via VGG19 and the history encoder to obtain the visual and history embedding  $r_{V,s}$ ,  $r_{V,e}$ , and  $r_{H,i-1}$  respectively. They are later processed by the MM attention module to obtain the attended embedding  $a_{V,s}$ ,  $a_{V,e}$  and  $a_{H,i-1}$ . The

attended visual embedding  $a_{V,s}$  and  $a_{V,e}$  are then fed into the Visual LSTM to get the states output  $(h_v, c_v)$ . The question LSTM decoder takes the attended history embedding  $a_{H,i-1}$  as input and outputs the  $i$ -th question  $q_i$ . Similarly for *A-BOT*, it takes the video frames, audio signal and original input descriptions as input. These modalities of input data are processed by VGG19, VGGish [25] and the input description encoder to obtain their respective embedding  $(r_{V,1}, \dots, r_{V,d}, r_A$  and  $r_C$ . The MM attention module takes those embedding and the history embedding  $r_{H,i-1}$  as input and outputs the attended embedding vectors. While the attended audio and visual embedding vectors are processed by the audio-visual LSTM to obtain the states  $(h_{av}, c_{av})$ , the attended history embedding  $a_{H,i-1}$  is fused with  $i$ -th question embedding  $r_{q_i}$  to form the input for answer decoder. After having obtained  $i$ -th answer  $a_i$ , the  $q_i$  and  $a_i$  are used to form the  $i$ -th QA pair and to update the existing dialog history. When the dialog history includes 10 QA pairs, *Q-BOT* generate the final descriptions.

### 3.3 Cooperative Learning

We propose to learn the proposed QA-Cooperative networks with corresponding cooperative learning methods, which have different emphasis for the generative and discriminative dialog settings.

#### 3.3.1 Dynamic History Update Mechanism

The dialog history is an important supplementary information source for *Q-BOT* to describe the unseen video in our task. Under the generative internal dialog setting, we therefore propose to update the dialog history in a dynamic way [82]. To be more concrete, we maintain the embedding dimension of the newly generated QA pair equal to the dimension of the existing dialog history to emphasize the new information at each QA round. We deploy a linear layer to reduce the dimension of the existing dialog history  $d_{H,i-1}$  to the size of the current QA pair  $d_{p_i}$ , and then concatenate the dialog history embedding with the embedding obtained for the  $i$ -th QA round.

#### 3.3.2 Internal Selective Mechanism

We introduce a mechanism to improve the quality of the internal dialog under the discriminative setting where the dialog agents are expected to select appropriate questions and answers from given candidates.

It is usually more difficult to guarantee the quality of internal dialog under the generative internal dialog setting, since no internal evaluations are implemented on the dialog round level. However, we believe that the quality of the internal dialog is a significant factor that contributes to a better final description, which is also experimentally demonstrated in our experiments in Section 4. In order to improve the quality of the dialog for the discriminative setting, we propose an internal selective mechanism via the sparse annotations similar to [43]. Specifically, it can be considered as a pre-training stage during which the agents learn to reason. We compute the internal loss using the sparse annotations during the internal selection process as:

$$\mathcal{L}_{internal} = \sum_i y_i \log \text{softmax}(\text{logit}_i), y_i \in \{0, 1\}.$$

$y_i$  is 1 if the selected question/answer is not the ground truth one at each round, 0 otherwise. The sparse annotation refers to the fact that we only consider the binary selections of the ground truth questions and answers while computing the internal loss, which contrasts to the idea of considering their dense relevance scores [43].

Overall, we have two loss terms during the entire training process, i.e.,  $\mathcal{L}_{internal}$  and  $\mathcal{L}_{CE}$ .  $\mathcal{L}_{internal}$  is the loss computed using sparse annotations in order to enhance the reasoning ability of the agents.  $\mathcal{L}_{CE}$  is the cross-entropy loss on the probabilities of the final description. Thus we combine the internal loss term to enhance the reasoning ability of two agents and improve the quality of the internal dialog:  $\mathcal{L} = \lambda \mathcal{L}_{internal} + (1 - \lambda) \mathcal{L}_{CE}$ .  $\lambda$  is the weight for the internal loss term. We empirically set  $\lambda$  to be 0.1 in our experiments. In our training process, we first optimize our model using the above loss function. To stabilize the optimization, we train our model using only the cross-entropy loss  $\mathcal{L}_{CE}$  in the last several training epochs.

## 4 EXPERIMENTS

### 4.1 Dataset

We evaluate the proposed method on the AVSD dataset [3], [27]. The data collection process reassembles our task setup where two Amazon Mechanical Turks (AMT) play the role 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 input descriptions. After having a conversation about the events that happened between the frames through 10 rounds of QA interactions, the Questioner is asked to summarize the entire video. The AVSD V0.1 is split into 7659 training, 734 prototype validation and 733 prototype testing dialog, each dialog consists of 10 rounds of question and answer pairs, accompanying the corresponding video clip, audio signals and input descriptions. Our experiments are performed on the provided training, validation and testing split.

### 4.2 Implementation

#### 4.2.1 Evaluation Metrics.

The BLEU1-4 [41], METEOR [8], SPICE [4], ROUGE\_L [32], and CIDEr [59] are used as the quantitative evaluation metrics for our generated final descriptions. CIDEr, which measures the similarity of a sentence to the consensus, is our primary metric for evaluations. For the internal dialog interactions between two agents under the discriminative setting, we further compute and analyze the ground truth question and answer selection ratios during training as additional quantitative evaluations.

#### 4.2.2 Data Representations

Our cooperative dialog agents have multiple modalities of data input including visual, audio, and textual data. For the visual data of *A-BOT*, we take the video representations extracted from the last conv layer of a VGG19 as input. We sample four equally spaced frames from the beginning of the original video, and each frame representation is of



dimension  $49 \times 512$ , where spatial and visual embedding dimensions are 49 and 512, respectively. For the visual input of *Q-BOT*, which only serves the purpose of sketchy perception of the general environment, we begin with getting the first and last frames of the video, and then perform the semantic segmentation using the pre-trained PSPNet with ResNet-50 on the ADE20K dataset [77], [78]. The segmentation result images are used to extract the representations following the same procedure as for *A-BOT*, the final representation is of dimension  $28 \times 512$ . For the audio modal, we obtain the 256-dim audio feature via VGGish [25]. For the textual representations, we extract the language embedding from the last hidden state of their corresponding LSTM. The dimensions are  $d_C = 256$ ,  $d_q = 128$ ,  $d_a = 128$  and  $d_H = 256$ .

#### 4.2.3 Test Settings

During our test, the performance of *Q-BOT* is evaluated at each QA round-level. In other words, each dialog is split into ten independent evaluation cases with the starting round number ranges from 1 to 10. For example, if the start round number  $i$  is 1, then no existing dialog history is given to *Q-BOT* and *A-BOT*, they will generate all the 10 questions and answers by themselves. However, if the start round number  $i$  equals 6, then five rounds of QA pairs are given to two agents as the existing history, in which case, *Q-BOT* still has another five changes to freely ask questions. For a given video, testing with different start round numbers is independent, resulting in 10 different test cases. Therefore, for the 733 videos from the test set of the AVSD dataset [27], we have in total 7330 different test cases. We refer to this testing process as the standard test setting. We also conduct a "strong baseline" experiment with the full ground truth dialog provided as input. For the strong baseline situations, there are only 733 test cases due to the fact that the entire dialog history is provided.

#### 4.2.4 Implementation Details

The description generator of our proposed QA-Cooperative networks is trained using a cross-entropy loss on the probabilities  $p(s_k | s_{<k}, x_Q)$  on the final descriptions. All the components are jointly trained in an end-to-end manner. The total amount of trainable parameters is approximately 19M for the generative dialog setting and 12M for the discriminative dialog setting. We use the Adam optimizer with a learning rate of 0.001 and a batch size of 64 for training. During training, we evaluate the performance on the validation set with a perplexity metric. The training stops after two consecutive epochs with no improvement in the perplexity metric.

### 4.3 Compared methods

We categorize the experiments into multiple groups to provide a more comprehensive and objective analysis for the unseen video description task and the proposed methods.

**A-BOT for Description.** To better understand and illustrate the knowledge gap as well as the transfer process between two dialog agents, we include the experiments for *A-BOT* to accomplish the same video description task. In this case, our video description task can be also considered as a classic video captioning task. *A-BOT* has access to the video data

(i.e., the visual frames and audio signals), and is asked to describe the video. The dialog history and the original input descriptions are removed from the input for *A-BOT* to reduce bias for this group of experiments.

**Basic Baselines.** The basic baselines are obtained without the cooperative internal dialog process, which means that the *Q-BOT* is asked to directly describe the video based on the existing dialog history input without additional chances to ask questions. In this case, the number of testing cases remains to be 7330.

**Strong Baselines.** The strong baselines are established by providing the full 10 rounds *ground truth dialog* to *Q-BOT*. *Q-BOT* can thus directly generate the final descriptions without asking questions. It can be regarded as an "upper bound case" for the generative dialog setting to some extent, since the internal dialog is trained to imitate the ground truth QA interactions.

**Other Baselines.** We also investigate the performance using the previously proposed methods [27], [48]. However, the previous methods are initially designed for question answering tasks, therefore, we modify the generators to generate video descriptions after 10 rounds of QA interactions and fine-tune the models for our task. All the data input for these baselines remains the same as for the proposed QA-Cooperative networks.

**Simulated Human Evaluation.** Considering the intended practical scenario for our proposed task involves the interactions between AI systems (i.e., the role of *Q-BOT* in this work) and the real human users (i.e., the role of *A-BOT* in this work), we also perform a simulated human evaluation test for the discriminative dialog setting. During the simulated test, the answers given by *A-BOT* are replaced by the ground truth answers correspond to the picked question by *Q-BOT* in inference. The training process remains the same as previously described in Section 3.3.

### 4.4 Video Description Results

The quantitative experimental results for the final description are shown in Table 2. We observe that *A-BOT* performs better than *Q-BOT* as expected. However, with the proposed QA-Cooperative networks and cooperative learning methods, our *Q-BOT* can achieve very promising performance under both generative and discriminative internal dialog settings, especially with the discriminative setting. In the meanwhile, the above observations also show that although the dialog has been proven to be an effective information source to supplement the incomplete visual data compared to the basic baseline setting without dialog, it is rather difficult to completely alternate the missing visual information. The results are also consistent with the knowledge gap observed from two types of descriptions from the dataset as described in Section 1 due to the lack of visual data.

We also present the experimental results obtained under the initial task setup from [82], where the visual input for *Q-BOT* is two original static RGB frames from the video without semantic segmentation. The performance shows no evident gap between the two task setups, demonstrating that our models extract the useful information from the dialog history, instead of benefiting from the possible bias introduced in the visual input. It is worth noting that this

TABLE 2

Quantitative experimental results of the unseen video description task. *HIS Att* stands for *History attention*. *G* and *D* denote the Generative or Discriminative dialog settings. The experiments are split into multiple groups, the group for *A-BOT* helps to understand the knowledge gap between two dialog agents. We observe that both *A-BOT* and *Q-BOT* from the proposed QA-Cooperative networks successfully transfer the knowledge by achieving very competitive performance that beats multiple baselines. For better visualization, we mark the scores in descending order with the green-blue-black color gradient.

Group	Method	HIS Att	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	SPICE	ROUGE_L	CIDEr
A-BOT	Hori <i>et al.</i> [27]	-	34.2	17.1	8.4	4.8	11.5	11.4	24.9	20.7
	S. <i>et al.</i> [48]	-	32.1	16.2	8.7	5.1	12.1	11.6	27.6	21.6
	S. <i>et al.</i> [48]	IM	33.8	16.9	9.1	5.3	12.7	11.8	27.7	22.7
	S. <i>et al.</i> [48]	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 w/o dialog	-	28.1	12.4	6.5	3.5	11.0	8.2	25.0	14.2
	Ours	IM	31.8	15.6	8.1	4.5	11.6	11.0	25.8	18.0
	Ours	MM	33.1	16.0	8.3	5.1	12.5	11.2	27.8	22.1
Q-BOT Strong baselines	Ours (full GT HIS)	IM	33.5	17.0	8.9	5.4	12.7	11.5	27.0	21.2
	Ours (full GT HIS)	MM	34.7	18.4	10.2	6.1	13.6	14.2	28.7	25.9
Q-BOT Cooperative	Ours-G (pre-trained)	MM	31.4	17.1	9.2	5.4	12.7	11.4	27.1	21.3
	Our QA-C [82]	IM	33.3	17.0	9.1	5.4	12.6	11.7	27.3	21.3
	Our QA-C [82]	MM	33.3	17.3	9.5	5.5	12.8	12.4	27.9	23.1
	Our QA-C(G)	IM	31.8	16.2	9.1	5.3	12.7	11.6	27.0	21.2
	Our QA-C(G)	MM	32.4	16.3	9.7	5.4	12.9	11.3	27.7	22.9
	Our QA-C(D)	IM	33.8	17.7	9.7	5.9	12.8	13.2	28.2	26.1
	Our QA-C(D)	MM	34.7	18.0	10.2	6.1	13.2	13.6	28.6	27.1
	QA-C(D) w/ simulated A	MM	34.3	18.4	10.3	6.3	13.4	14.1	28.6	27.6

observation is not contradictory to the previous statement about the significance of the visual data. The previous one emphasizes its importance from the existence and temporal aspect, while the raw/segmented visual information addresses more on the details within the same frame.

The improvement for the final description performance compared to basic baselines shows the effectiveness of the knowledge transfer process between two agents with unbalanced input data. For the generative internal dialog setting, our *Q-BOT* with the QA-Cooperative network can achieve comparable performance close to the strong baselines where the full ground truth dialog is provided. In contrast, for the discriminative setting, our *Q-BOT* is able to outperform the strong baselines for most of the evaluation metrics, the primary metric CIDEr score achieves 27.1. The simulated human evaluation yields better performance compared to the case with both dialog agents. In addition, we also notice that the MM attention mechanism helps with performance improvement compared to the IM attention mechanism.

Figure 4 shows examples of qualitative results. Due to the limited space, more qualitative examples can be found in Appendix. The qualitative examples reveal the consistent results with our quantitative evaluations, the video descriptions generated by *Q-BOT* with our proposed QA-Cooperative networks contain more detailed information compared to the basic baselines and are more close to the strong baseline cases where the full ground truth dialog is provided as input. The examples in Figure 4 is challenging test cases due to the fact that only a few rounds of QA pairs are included in the input, however, the final descriptions given by our *Q-BOT* contains the concrete information such as the room types (*e.g.*, the kitchen) that are not included in the input. It demonstrates that our *Q-BOT* does benefit from the effective knowledge transfer process via the natural language dialog to describe the unseen videos. We also notice

from the qualitative results that the internal dialog obtained under the generative internal dialog setting tends to contain repetitive information, which is also observed from previous work on the dialog agents [15]. As for comparisons, the questions and answers selected under the discriminative internal dialog setting are more diverse and informative, which explains the reason for the better final descriptions.

## 4.5 Ablation Studies

We continue to conduct extensive ablation studies on model components, data modalities, QA pairs, and beam width in this section to better analyze the proposed methods. Note that all the experiments in this section adopt the MM attention module as the attention mechanism since the MM attention module is proved to be more effective than IM attention module in previous experiments in Table 2.

### 4.5.1 Model Components

**Attention Modules.** We propose two attention mechanisms in our QA-Cooperative network architectures, *i.e.*, the MM (Multi-Modal) attention module and the IM (Intra-Module) attention module. Interestingly, we observe that both attention mechanisms help to improve the final performance of the video description, which is different from the results in [48]. In [48], the authors find that the attention on the dialog history does not yield performance improvements for the AVSD task. One possible reason for this difference could probably be explained from the perspective of causal inference as in [43], where the dialog history is found to be a spurious and biased factor and should be removed for classic question answering tasks. This again emphasizes the difference of our task from the classic VQA and visual dialog tasks from a novel angle of causality. The dialog history for *Q-BOT* in our unseen video description task is

TABLE 3

Quantitative results for ablation studies on model components, data modalities, QA pairs, beam width and cluster numbers. *G* denotes the Generative dialog setting, while *D* denotes the Discriminative one. For each group, we use the green-blue-black color gradient to mark the scores in descending order for better visualizations.

Group	Setting	Ablation	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	SPICE	ROUGE_L	CIDEr
Model Components	G	w/o Att.	31.5	16.3	8.8	4.9	12.3	11.2	26.8	20.4
		w/o AV-LSTM	32.1	16.2	8.8	5.1	12.1	11.3	27.1	20.3
	D	w/o Att.	33.8	17.1	9.0	5.8	12.7	12.3	27.6	25.3
		w/o AV-LSTM	32.6	16.8	8.8	5.3	12.5	12.0	27.5	23.7
		w/o Reasoning	34.3	18.3	10.0	6.3	12.8	13.1	27.9	26.3
Data Modalities	G	w/o visual frames	33.1	16.1	7.3	4.4	11.6	10.8	26.1	20.0
		full segmented frames	34.0	17.6	9.8	6.1	12.6	11.9	27.7	23.0
		w/o Audio	32.2	16.2	9.4	5.4	12.8	11.2	27.2	22.3
		w/o Input description	31.5	15.3	7.9	4.6	12.7	11.1	26.3	20.1
		w/o HIS for <i>A-BOT</i>	32.5	16.3	9.3	5.4	12.1	11.2	27.1	23.0
	D	w/o segmented frames	33.0	16.8	9.1	5.9	12.7	12.6	27.8	25.3
		full segmented frames	34.6	18.8	10.6	6.5	14.0	13.9	28.7	27.7
		w/o Audio	34.6	18.2	10.2	6.2	13.3	13.1	28.6	26.6
		w/o Input description	32.8	17.3	8.3	6.1	13.1	12.8	27.7	25.8
		w/o HIS for <i>A-BOT</i>	34.0	18.1	10.5	6.3	12.9	13.0	28.3	26.3
QA Pairs	G	Shuffled order	31.4	15.5	8.4	4.9	11.7	11.1	26.3	20.0
		Round#2	27.9	13.3	7.0	4.0	10.8	9.7	24.9	16.7
		Round#5	32.7	16.7	9.4	5.6	12.2	12.1	27.6	22.9
		Round#8	34.1	17.6	9.8	5.9	12.9	12.6	28.5	25.5
	D	Shuffled order	34.0	18.0	10.2	6.1	12.3	13.0	28.1	25.4
		Round#2	32.2	17.4	9.0	5.2	12.3	11.9	27.6	23.0
		Round#5	34.3	18.1	9.6	5.8	13.1	13.0	28.1	26.4
		Round#8	35.1	18.5	10.4	6.5	13.4	13.2	28.8	28.4
Beam Width	G	Beam width=1	28.5	14.8	7.9	4.2	10.9	10.4	26.0	17.4
		Beam width=3	32.4	16.3	9.7	5.4	12.9	11.3	27.7	22.9
		Beam width=5	33.4	16.3	9.8	5.5	12.9	11.0	27.5	22.8
	D	Beam width=1	33.5	17.5	9.8	5.7	12.8	13.1	28.8	24.7
		Beam width=3	34.7	18.0	10.2	6.1	13.2	13.6	28.6	27.1
		Beam width=5	34.9	18.4	10.4	6.3	13.3	13.7	28.8	27.1
Clusters	D	k = 5	34.7	17.7	9.5	5.8	13.1	13.4	28.3	26.8
		k = 10	34.7	18.0	10.2	6.1	13.2	13.6	28.6	27.1
		k = 15	34.8	18.2	10.2	6.2	13.1	13.7	28.8	27.0

not a spurious factor but a significant information source. The attention modules help to raise the CIDEr scores from 20.4 to 22.9 for the generative setting, and from 25.3 to 27.1 for the discriminative setting.

**Audio-Visual LSTM.** The Audio-Visual LSTM component is designed for *A-BOT* to process the audio and visual input data in addition to the MM module. The output of this model component is used as the initial state input for *A-BOT*. We test the ablation experiments by removing this component. Experiments show that the audio-visual LSTM contributes to the final description performance. This module accounts for a raise of 2.6 and 3.4 on CIDEr scores for the generative and discriminative settings, respectively.

**Internal Selection.** Under the discriminative setting, we address the internal selective abilities of two agents by adopting the sparse annotations as explained in Section 3.3.2. This ablation experiments prove the contributions of this module, leading to an increase of 0.8 for the CIDEr score.

#### 4.5.2 Data Modalities

Our unseen video description task incorporates multiple modalities of data, we therefore perform the ablation studies to analyze the impact of different data modalities.

**Visual Data.** In our experimental settings, *Q-BOT* takes two segmented visual frames as the implicit visual input. We exploit the impact of different types of visual input on the final video description performance. Specifically, we conduct ablation studies with the full segmented frames

and without any visual frames under both dialog settings. The experimental results demonstrate the significance of the visual data for our proposed task. It is worth noting that even the visual input are processed with segmentation operation, we observe performance improve given more video frames. In contrast, the complete removal of visual data from input causes the relatively poor performance for video descriptions.

**Audio Data.** Audio data forms part of the input for *A-BOT*, since the audio perception is another important information source in addition to the vision for humans. We remove the audio data from the input for *A-BOT* to investigate its influence on the final description performance. We observe from Table 3 that audio data contributes to the better final description performance under both internal dialog settings.

**Input Description Data.** The input descriptions obtained from the human annotators are also provided to *A-BOT* under our task setup. We study the performance of final descriptions after removing the input description data from the input for *A-BOT*. Table 3 shows that the input descriptions also contribute to the final description performance.

**Dialog History.** The dialog history is a common input for both *Q-BOT* and *A-BOT*. It is the major information source for *Q-BOT* to describe the unseen video, which is already demonstrated in the first basic baseline situation in Table 2. Therefore, for the ablation studies, we conduct experiments under the situation where the dialog history is invisible to *A-BOT*. Interestingly, the performance is not much impaired

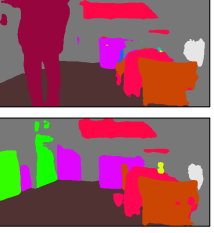
Input of Q-BOT	Video descriptions	Ground Truth Dialog	Generative Dialog	Discriminative Dialog
 <p>Q1: is there a person in the video? A1: yes, there is <b>one person</b> in the video. Q2: is it a male or female? A2: the person is <b>male</b>.</p>	<p>Human(GT): a man is vacuuming the kitchen floor. when he finishes he picks up some garments and leaves the kitchen.</p> <p>Q-Basic: a man walks into a room and takes off his shirt. he then walks away.</p> <p>Q-Strong: a man walks into a room and takes off his jacket. he then takes a bite of a sandwich and then walks away.</p> <p>A-BOT: a man is vacuuming the floor. he then goes to the living room and grabs a vacuum cleaner.</p> <p>Ours (G): a man is in the kitchen. he is holding a cup of coffee. he takes a drink from a cup and walks out of the room.</p> <p>Ours (D): a man walks into a room and puts on a jacket, he then picks up a broom and leaves the room.</p>	<p>Q3: does more than one person appear in the video? A3: there is only one person. Q4: is he in the kitchen? A4: yes, did you see the machine he uses. Q5: is he holding something? A5: yes, he is holding a handheld vacuum. Q6: is he cleaning? A6: yes he is, he is vacuuming the floor. Q7: does he pick up anything else? A7: he does very good, he picked up a garment. Q8: does he walk out of the kitchen? A8: yes he walks out after picking up the garment. Q9: where does he pick up the clothes from? A9: towards the back of the kitchen behind the table. Q10: does he come back into the kitchen after taking the clothes? A10: no he does not.</p>	<p>Q3: is there one person? A3: yes, there is. Q4: is there any person in pics? A4: yes, there is one person. Q5: is there anything in the kitchen? A5: yes. Q6: is there anything prepared? A6: yes, i think so. Q7: is there anything crossed? A7: no, there is one person. Q8: is there any person in pics? A8: yes, i think so. Q9: is there any person in pics? A9: yes, i think so. Q10: is there any person in pics? A10: yes.</p>	<p>Q3: does he pick up the bag? A3: not true, maybe his twinties. Q4: what room does this take place? A4: by having him go through a box of dishes. Q5: can you hear sound or any speaking in the video? A5: yes he does look around some for a short time. Q6: so a man in a room, picks up a plate of food and starts to eat. What is next? A6: he does some kind of jogging dance. Q7: is his phone on as he looks at it? A7: yes, i think so. Q8: where does he go in the video? A8: yes, he stays in the room. Q9: how many people are in the whole video? A9: yes it is just him sitting. Q10: does anyone else appear in the video? A10: yes, they seem to be alone.</p>

Fig. 4. Example of qualitative results. We present the input of Q-BOT, different video descriptions, and the internal dialog. The descriptions given by our Q-BOT include more details compared to multiple baselines. The color intensities in the figure represent attention weights. More examples can be found in Appendix.

in this case compared to other data modalities. Intuitively, it is also reasonable due to the fact that A-BOT does not rely on the dialog history to provide answers to the questions raised by Q-BOT, since A-BOT has already watched the entire video. This finding is also consistent with the previous findings for classic question answering tasks in [48] and [43]. In [48], the authors find that the attention on the dialog history does not yield performance improvement for answering questions. In the work of Qi *et al.* [43], the dialog history is proven to be a spurious factor that ultimately impairs the performance for question answering tasks.

#### 4.5.3 QA Pairs

**Order of QA Pairs.** We test the impact of the order of the QA pairs by randomly shuffle the orders in the dialog history. Similar to the observations from [3], [82], the QA in the dialog history order is an important factor that influences the performance of the final descriptions. With the shuffled dialog history, the performance is impacted under both generative and discriminative dialog settings. We observe that the primary CIDEr scores drop 2.9 and 1.7 for the generative and discriminative settings, respectively.

**Number of Input QA Pairs.** We also take a closer look at the experimental results with different numbers of QA pairs included in the input dialog history. In other words, we modify the number of starting round for the testing cases. Unsurprisingly, the more ground-truth QA pairs in the dialog history usually lead to the better final performance for describing the unseen videos.

#### 4.5.4 Other Hyper-parameters

**Beam Search.** We use the beam search when generating the final descriptions. We experiment with different numbers of beam width. The experimental results from Table 3 shows that with wider beam width, the final performance for the unseen video description tasks improves. However, a beam width of 3 is generally adequate for achieving good results. For the main experimental results reported in Table 2, we adopt the beam width of 3.

**Number of Clusters.** In the discriminative internal dialog setting, Q-BOT selects the questions following a two-phase

selection mechanism. The question candidates are firstly processed using unsupervised k-means clusters algorithm. We test different numbers of clusters to study its impact. The experimental results show that a larger number of clusters leads to slightly better performance. In our main experiments, we use 10 clusters in the first selection phase.

## 5 CONCLUSION AND DISCUSSION

In this work, we propose a novel multi-modal task that aims to describe an unseen video based on the incomplete visual input and the natural language dialog. There are two primary motivations behind this work: to introduce a more reliable task setup by providing AI with implicit visual input, and to demonstrate the effectiveness of using the natural language dialog as the additional source to supplement the missing visual information. We propose two different experimental settings with their corresponding cooperative network models that effectively help with the knowledge transfer process between two agents. Extensive experiments demonstrate the promising and competitive performance of the proposed methods over multiple baselines.

There are research directions that worth further exploiting in the future: (a) One possible direction could be encouraging more efficient dialog interactions between two agents. Specifically, we observe from the experiments that the dialog agent Q-BOT does not always need 10 question chances to achieve good performance for the video descriptions. It may already have enough information to summarize the video at the end of eight or nine rounds of QA interactions. It would be therefore interesting to further encourage more efficient information exchange, and to exploit the possible early stop mechanism for the dialog interactions. (b) The simulated human evaluation results indicate that there is still room to enhance A-BOT's ability to achieve better performance for the proposed video description task. (c) It would also be interesting to further refine the task formulation and to design more specific ultimate objectives other than general video descriptions, *e.g.*, we could ask Q-BOT to generate a scene graph mainly based on the natural language dialog. (d) A more sophisticated mechanism that enables Q-BOT to

ask guided and structured questions could also be useful when applied in real-life scenarios.

## 6 ACKNOWLEDGEMENTS

This research was partially supported by NSF NeTS-2109982 and the gift donation from Cisco. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

## REFERENCES

- [1] Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. History for visual dialog: Do we really need it? In *ACL*, 2020.
- [2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020.
- [3] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *CVPR*, 2019.
- [4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.
- [5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
- [7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.
- [8] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.
- [9] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. In *NeurIPS*, 2019.
- [10] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020.
- [11] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017.
- [12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.
- [13] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 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.
- [14] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.
- [15] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, 2017.
- [16] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [18] Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV*, 2021.
- [19] Zhe Gan, Yu Cheng, Ahmed Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *ACL*, 2019.
- [20] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.
- [21] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.
- [22] Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. Iterative context-aware graph inference for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [23] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *CVPR*, 2020.
- [24] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesaro, and Rogerio Schmidt Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018.
- [25] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*. IEEE, 2017.
- [26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.
- [27] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP*. IEEE, 2019.
- [28] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019.
- [29] Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *CVPR*, 2018.
- [30] Sabine Kastner, Mark A Pisk, Peter De Weerd, Robert Desimone, and Leslie G Ungerleider. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 1999.
- [31] Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. Answerer in questioner’s mind: information theoretic approach to goal-oriented visual dialog. In *NeurIPS*, 2018.
- [32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [33] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
- [34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.
- [35] Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. Flippdial: A generative model for two-way visual dialogue. In *CVPR*, 2018.
- [36] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*. Springer, 2020.
- [37] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.
- [38] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision*, 126(10):1120–1137, 2018.
- [39] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.
- [40] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.
- [41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- [42] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
- [43] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020.
- [44] Janarthanan Rajendran, Jatin Ganhotra, Satinder Singh, and Lazaros Polymenakos. Learning end-to-end goal-oriented dialog with multiple answers. In *EMNLP*, 2018.
- [45] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.
- [46] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*. IEEE, 2019.
- [47] Nicholas Roy, Joelle Pineau, and Sebastian Thrun. Spoken dialogue management using probabilistic reasoning. In *ACL*, 2000.
- [48] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019.
- [49] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G



- Schwing. Factor graph attention. In *CVPR*, 2019.
- [50] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.
- [51] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. In *NeurIPS*, 2017.
- [52] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.
- [53] Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulka-rni, Matthew Turk, and William Yang Wang. What should i ask? using conversationally informative rewards for goal-oriented visual dialog. In *ACL*, 2019.
- [54] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. Reinforcement learning for spoken dialogue systems. In *Nips*, 1999.
- [55] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *ACM Multimedia*, 2018.
- [56] Florian Strub, Harm De Vries, Jérémie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*, 2017.
- [57] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.
- [58] Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 2000.
- [59] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.
- [60] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018.
- [61] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018.
- [62] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*, 2017.
- [63] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*, 2020.
- [64] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *CVPR*, 2018.
- [65] Yu Wu, Lu Jiang, and Yi Yang. Revisiting embodiedqa: A simple baseline and beyond. *IEEE Transactions on Image Processing*, 29:3984–3992, 2020.
- [66] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021.
- [67] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *ACM Multimedia*, 2018.
- [68] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019.
- [69] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.
- [70] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*. Springer, 2016.
- [71] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [72] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
- [73] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.
- [74] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.
- [75] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton Van Den Hengel. Goal-oriented visual question generation via intermediate rewards. In *ECCV*, 2018.
- [76] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.
- [77] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.
- [78] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2018.
- [79] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.
- [80] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018.
- [81] Ye Zhu, Yu Wu, Hugo Latapie, Yi Yang, and Yan Yan. Learning audio-visual correlations from variational cross-modal generation. In *ICASP*, 2021.
- [82] Ye Zhu, Yu Wu, Yi Yang, and Yan Yan. Describing unseen videos via multi-modal cooperative dialog agents. In *ECCV*, 2020.



## APPENDIX A

### KNOWLEDGE GAP AND TWO TYPES OF DESCRIPTIONS FROM THE AVSD DATASETS

Our proposed task involves two different types of video descriptions as mentioned in the Introduction. In this Appendix, we provide further clarifications on the differences between the input and final descriptions, which helps to better understand the concept of knowledge gap caused by the implicit visions.

Table 4 shows the language scores computed between the two types of descriptions from the original AVSD dataset [3], [27] using the ground truth captions obtained after watching the entire video (*i.e.*, the input descriptions) as references. The average word lengths of the input and final descriptions are 23.8 and 23.0 words, respectively. Although the word lengths are rather similar, the relatively low scores between two types of descriptions reveal their evident disparity.

TABLE 4  
Language scores between the input and final descriptions from the AVSD dataset

Metric	BLEU4	METEOR	SPICE	ROUGE_L	CIDEr
GT annotations	5.12	15.0	16.9	27.0	26.5

Figure 6 shows more qualitative examples from the AVSD dataset that reveals the difference between the input and the final descriptions as well as the knowledge gap. We also provide the raw frames and the segmented frames for comparisons. We observe from the figure that the final descriptions given by human annotators lack certain details compared to the input descriptions, which are the video captions obtained after watching the entire video.

## APPENDIX B

### DETAILS ABOUT THE MODEL

We list the details about the model components used in our experiments in this section in Table 5. In addition, for the LSTM layers, we use the Xavier weight initialization [?].

## APPENDIX C

### MORE EVALUATION ABOUT THE INTERNAL SELECTION MECHANISM

In addition to the final description performance in *w/o Reasoning* setting, we also calculate the ground truth question and answer selection ratio as qualitative evaluation as shown in Figure 7. The ground truth selection ratios increase after deploying the internal reasoning mechanism. Additionally, we also observe that the selection ratio for questions is generally higher than the ratio for answers. The selection ratios also increase as more ground truth QA pairs are provided as input (*i.e.*, with larger starting round number), as in Figure 7, the ground truth selection ratios with the starting round number 8 are generally higher than the starting round number 2.

## APPENDIX D

### MORE QUALITATIVE RESULTS

We provide more qualitative results and analysis in this section.

**Additional Qualitative Examples.** We present in Figure 8 more qualitative examples.

**Question Evaluation.** Although the final objective of our work is for *Q-BOT* to describe an unseen video, the ability of *Q-BOT* to ask meaningful questions is also very important.

For the generative setting, there is no explicit loss function for the question generation process imposed on the *Q-BOT* during the training, therefore, the model tends to ask repetitive questions with a relatively high score of Self-BLUE4 metric [?] of 0.82. We then proceed to introduce the discriminative setting to reduce the possible bias learned from the generative setting.

Figure 5 shows the distributions of the clusters for question and answer candidates in inference for the discriminative dialog setting.

**Human Evaluation.** Considering that the intended practical application scenario for our proposed task involves the interactions between the AI systems (*i.e.*, *Q-BOT*) and real human users (*i.e.*, *A-BOT*), we perform an extra set of human evaluation test to provide a more thorough analysis of our work.

Figure 9 shows qualitative examples of the human evaluation test corresponding to the qualitative examples in the main paper. During the test, we replace the role of *A-BOT* by human participants and provide the real-time answers according to the generated/selected questions. It is worth mentioning that there are several difficulties during the human evaluation test. The most challenging problem is that the questions asked by *Q-BOT* are not always reasonable. Specifically, there are questions that are irrelevant to the actual video. We intentionally define that for those questions, the participants can always provide the answer as "I don't know".

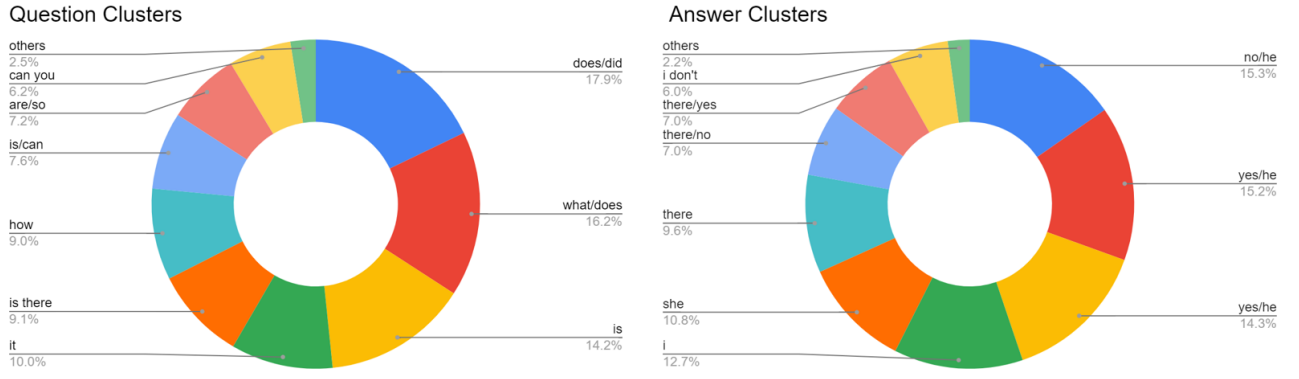


Fig. 5. Distributions of the clusters for question and answer candidates in inference. We roughly show the first n-grams for the majority of questions and answers in each cluster. It is worth noting that there are possibly several clusters with similar first n-grams due to the fact that we embed the entire sentence for clustering.

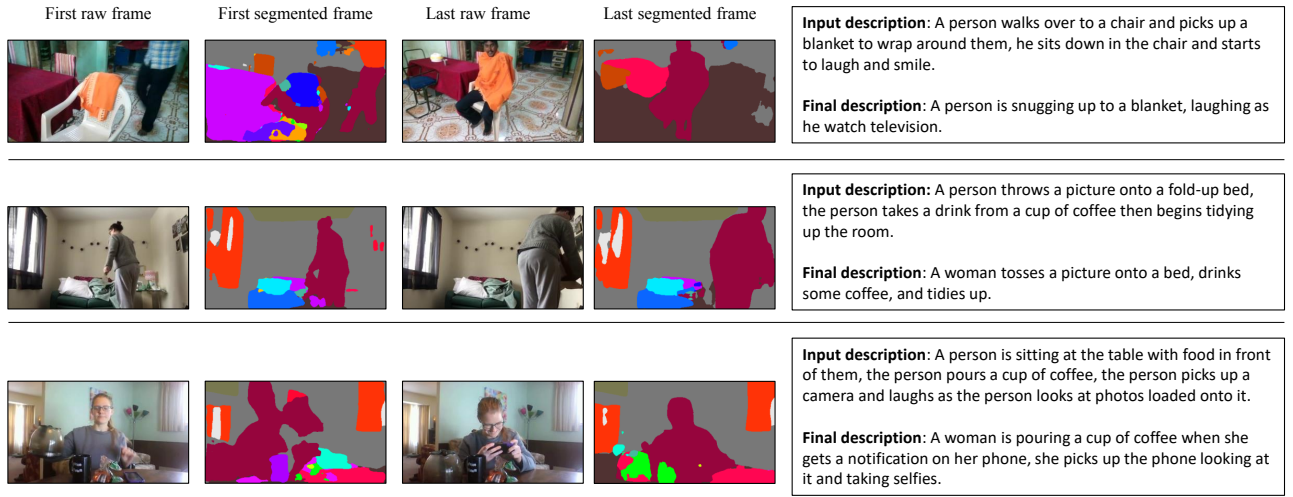


Fig. 6. More qualitative examples from the AVSD datasets with input and final descriptions. We observe that the final descriptions given by human annotators without seeing the entire videos miss certain details compared to the input descriptions, despite the dialog interactions help to provide more video information that are not revealed in the static frames.

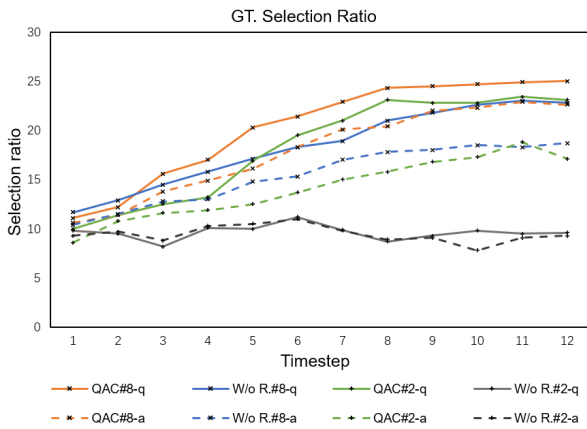


Fig. 7. Ground truth question and answer selection ratio during training. We plot the selection ratios with the starting round number 2 and 8 as examples. The solid lines represent the ground truth selection ratios for questions, the dotted lines are the ratios for answers.

TABLE 5

Details about the model components. The column of agent without specification of **Generative** or **Discriminative** means the component is the same for both settings.

Agent	Component	Functions	Details
Q&A	MM module	cross-modal attention	[49]
Q&A	history encoder	process the existing dialog history	linear + 1-layer LSTM with size equals to the dimension of the history embedding 256
Q	visual LSTM	process the segmented visual input	LSTM with 2 units with size equals to the dimension of the attended visual embedding 128
Q(G)	question decoder	generate questions to ask	LSTM-based generator with size equals to the dimension of the question 128
Q(D)	question decoder	select questions to ask	linear + dot product + softmax selection
Q(D)	candidates encoder	process the question candidates	1-layer LSTM with size equals to the dimension of question embedding 128
Q	description generator	generate the final video descriptions	LSTM-based generator with size equals to the overall history embedding 256
A	audio-visual LSTM	process the audio and video information	LSTM with 5 units with size equals to the dimension of attended audio + visual embedding 256
A	input description encoder	process the input description	linear + 1-layer LSTM with size equals to the dimension of the input description embedding 256
A(G)	answer decoder	generate answers for the questions raised by Q-BOT	LSTM-based generator with size equals to the dimension of question embedding + history embedding + input description embedding
A(D)	answer decoder	select answers for the questions picked by Q-BOT	linear + dot product + softmax
A(D)	candidate encoder	process the answer candidates	1-layer LSTM with size equals to the dimension of answer embedding 128
Q&A	dynamic history update	update the existing history and emphasize the newly generated information	linear + concatenation

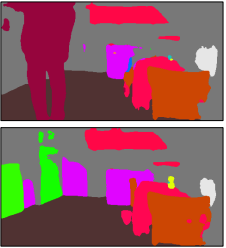
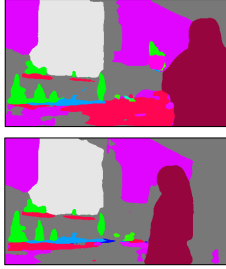

Input of Q-BOT	Video descriptions	Ground Truth Dialog	Generative Dialog	Discriminative Dialog
 <p>Q1: is there a person in the video? A1: yes, there is <b>one person</b> in the video. Q2: is it a <b>male</b> or <b>female</b>? A2: the person is <b>male</b>.</p>	<p>Human(GT): a man is vacuuming the kitchen floor. when he finishes he picks up some garments and leaves the kitchen.</p> <p>Q-Basic: a man walks into a room and takes off his shirt. he then walks away.</p> <p>Q-Strong: a man walks into a room and takes off his jacket. he then takes a bite of a sandwich and then walks away.</p> <p>A-BOT: a man is vacuuming the floor. he then goes to the living room and grabs a vacuum cleaner.</p> <p><b>Ours (G): a man is in the kitchen. he is holding a cup of coffee. he takes a drink from a cup and walks out of the room.</b></p> <p><b>Ours (D): a man walks into a room and puts on a jacket, he then picks up a broom and leaves the room.</b></p>	<p>Q3: does more than one person appear in the video? A3: there is only one person. Q4: is he in the kitchen? A4: yes, did you see the machine he uses. Q5: is he holding something? A5: yes, he is holding a handheld vacuum. Q6: is he cleaning? A6: yes he is, he is vacuuming the floor. Q7: does he pick up anything else? A7: he does very good, he picked up a garment. Q8: does he walk out the kitchen? A8: yes he walks out after picking up the garment. Q9: where does he pick up the clothes from? A9: towards the back of the kitchen behind the table. Q10: does he come back into the kitchen after taking the clothes? A10: no he does not.</p>	<p>Q3: is there one person? A3: yes, there is. Q4: is there any person in pics? A4: yes, there is one person. Q5: is there anything in the kitchen? A5: yes. Q6: is there anything prepared? A6: yes, i think so. Q7: is there anything crossed? A7: no, there is one person. Q8: is there any person in pics? A8: yes, i think so. Q9: is there any person in pics? A9: yes, i think so. Q10: is there any person in pics? A10: yes.</p>	<p>Q3: does he pick up the bag ? A3: not true, maybe his twenties. Q4: what room does this take place ? A4: by having him go through a box of dishes. Q5: can you hear sound or any speaking in the video ? A5: yes he does look around some for a short time. Q6: so a man in a room, picks up a plate of food and starts to eat. What is next ? A6: he does some kind of jogging dance. Q7: is his phone on as he looks at it ? A7: yes, i think so. Q8: where does he go in the video ? A8: yes, he stays in the room. Q9: how many people are in the whole video ? A9: yes it is just him sitting. Q10: does anyone else appear in the video ? A10: yes, they seem to be alone.</p>
 <p>Q1: what is she doing in the kitchen ? A1: she walks up to the <b>counter</b> and grabs a pan off the <b>stove</b> while holding a black cat.</p>	<p>Human(GT): a woman holding a cat cooks in the kitchen and looks out a window.</p> <p>Q-Basic: a man walks into a room and takes off his shoes , he then walks over to a door and sits down.</p> <p>Q-Strong: a woman walks into a kitchen , opens a cabinet, takes something out of the fridge , puts them in the fridge , then walks away .</p> <p>A-BOT: a woman is in the kitchen cooking something on the stove , she then goes to the kitchen to turn off the light .</p> <p><b>Ours (G): a woman walks into a kitchen and opens a cabinet and pulls out a bottle of medicine. she then leaves the room.</b></p> <p><b>Ours (D): a woman is cooking and stirring a pot on the stove , she is cooking something on a stove .</b></p>	<p>Q2: is she in the kitchen the entire time ? A2: yes , she 's in the kitchen the entire clip . Q3: is the cat always held ? A3: she puts the cat down towards the end of the video . Q4: where does she put the cat ? A4: she puts the cat on the floor . Q5: what does she do with the pan ? A5: after she takes it off the stove she sets it on the counter and just moves it around a little bit . Q6: does she do anything with the stove itself ? A6: no , just takes the pan off of it . Q7: why does she look out the window ? A7: you can 't see anything out the window so i 'm not sure . Q8: could she have heard something ? A8: the only thing that can be heard is the pan hitting the counter. Q9: does she appear to be alone ? A9: yes , she appears to be alone . Q10: why does she move the pan around after she sets it on the counter ? A10: i 'm not sure , she just pushes it around the counter slightly .</p>	<p>Q2: is she the only one ? A2: yes, she is. Q3: does she cross the table? A3: no he sits down on the floor. Q4: does she cross and grasp it from the table ? A4: no she is not. Q5: is there only one person ? A5: yes i think so . Q6: does he eat anything ? A7: no he is not . Q7: does he eat anything ? A8: no he is not . Q8: does he eat anything ? A8: no he is not . Q9: does he eat anything ? A9: no he is not . Q10: does he eat anything ? A10: no he is not.</p>	<p>Q2: is there only one person? A2: yes , there is only one person. Q3: does she pick up anything? A3: no i do not see that. Q4: is there just one person? A4: there is no music playing. Q5: do they make eye contact? A5: yes did you see the machine he uses. Q6: is she cleaning? A6: basic yes she turns it on and walks off . Q7: does he put the glass in the sink? A7: in the kitchen. Q8: where does she pick up the clothes from? A8: looks like the mens / ladies room. Q9: does he come back into the kitchen after taking the clothes ? A9: there was nothing else interesting but before she drops the towel on the floor she stretches for a while . Q10: is that when it ends ? A10: there is nobody else with him .</p>
 <p>Q1: is there just <b>one person</b> in the video ? A1: yes , just <b>one person</b> in the video . Q2: is there any <b>sound</b> ? A2: yes , there is <b>sound</b> in the video . Q3: does the man <b>speak</b> or is the sound from the <b>tv</b> ? A3: there is sound from the tv and him <b>sneezing</b> . Q4: is the man <b>standing</b> the whole time ? A4: no he starts out <b>sitting down</b> on the <b>couch</b> . Q5: is he <b>holding</b> something in his hand ? A5: a <b>sandwich</b> and then he gets a <b>drink</b> .</p>	<p>Human(GT): : a man is sitting watching tv and eating a sandwich , then gets up and gets a drink .</p> <p>Q-Basic: a man walks into a room and sits down on a chair, he then walks to the door and leaves the room .</p> <p>Q-Strong: a man is sitting on a couch watching tv, he then gets up and picks up a cup, he then sits down and eats a sandwich .</p> <p>A-BOT: a man is sitting on a couch eating a sandwich and drinking a glass of water .</p> <p><b>Ours (G): a man is sitting on a couch eating a sandwich, he gets up and leaves the room .</b></p> <p><b>Ours (D): a man is sitting on a couch reading a book, he gets up and looks at a picture on the wall.</b></p>	<p>Q6: do you see him eat the sandwich ? A6: yes , he takes a few bites of it . Q7: do you see him drink his drink ? A7: yes , he takes a drink from the glass . Q8: does he do anything else , besides what you have already told me ? A8: he walks over to get a drink and walks back across the room . Q9: does he sit back down on the sofa ? A9: no , he does not sit back down . Q10: is he drinking his drink when the video ends ? A10: no , he is taking a bite of sandwich .</p>	<p>Q6: is he the only person? A7: yes, i think so . Q7: is he the only person in the video ? A8: yes, i think so . Q8: does he go and leave ? A8: no he does not. Q9: does he go and leave ? A9: no . Q10: does he go and leave ? A10: no .</p>	<p>Q6: why does he look out the window ? A6: you can't see anything out the window so i'm not sure . Q7: how many people are in the video ? A7: no , i don't believe this man to be that young . Q8: does he eat or drink anything ? A8: yes , after that he walks over and drinks from a water bottle . Q9: is there any audio in the clip ? A9: yes , you can hear the tv in the background . Q10: what does he have around the neck ? A10: they are chatting and laughing but i can not understand what they are saying .</p>

Fig. 8. More qualitative results for the proposed task in addition to the example shown in the main paper.

Video information	Input of Q-BOT	Simulated human test	Human evaluation (G)	Human evaluation (D)
 <p>Input description: a person is running a vacuum while sneezing and then picking up clothes in a basement.</p>	 <p>Q1: is there a person in the video? A1: yes, there is <b>one person</b> in the video. Q2: is it a <b>male or female</b>? A2: the person is <b>male</b>.</p>	<p>Q3: does he pick up the bag? A3: he picks up his bag, places it in it, closes it, then places it by the stairs. Q4: is he in the scene from the beginning? A4: he's walking into frame right as the video starts. Q5: does he ever react to what he sees on his phone? A5: he does not have a reaction. Q6: does he pick up anything in the bathroom? A6: no, he does turn on the light though. Q7: where does he get the drink? A7: he gets the drink off of the bookshelf. Q8: does he ever get up? A8: no he does not get up. Q9: where does he pick up the clothes from? A9: toward the back of the kitchen behind the table. Q10: does he drink from the mug? A10: no, i never see him drinking from it.</p> <p>Descriptions: a man walks into a kitchen and puts on a jacket, he then takes a drink from a cup and leaves the room.</p>	<p>Q3: is there one person? A3: yes, there is only one person in the video. Q4: is there anything on the floor? A4: yes, there is a table and some chairs on the floor. Q5: is there anything? A5: I am not sure what you ask. Q6: is there anything placed on the floor? A6: yes, a table and some chairs. Q7: is there anyone in the video? A7: yes, there is one person. Q8: is there any person in the video? A8: yes, there is one person. Q9: is there any person in the video? A9: yes, there is one person. Q10: is there any person in the video? A10: yes, there is one person.</p> <p>Description: a man walks into a room, he takes a drink from a cup and leaves the room.</p>	<p>Q3: does he pick up the bag? A3: I don't know. Q4: can you hear sound or any speaking in the video? A4: yes, there is sound of vacuum machine in the video. Q5: how many people are there in the video? A5: there is only one person. Q6: does he smile or laugh in the video? A6: no, he does not. Q7: does he say anything or make any sound? A7: he makes sound while using the vacuum machine? Q8: does he make any sound? A8: he makes sound while using the vacuum machine. Q9: does he leave the room? A9: yes, he leaves the room at the end of the video. Q10: does he walk fast or slow? A10: he walks slow.</p> <p>Description: a man walks into a kitchen and picks up a cup, he then takes a drink and leaves the room.</p>
 <p>Input description: a person is in the kitchen cooking by the window in the light while snuggling with their kitten</p>	 <p>Q1: what is she doing in the kitchen? A1: she walks up to the counter and grabs a pan off the <b>stove</b> while holding a black cat.</p>	<p>Q2: is there only one person? A2: yes, there is only one man. Q3: how old is the woman? A3: the woman appears to be in her late 20s. Q4: does she pick it back up? A4: she just stands there looking down at it / the camera without picking it up. Q5: where does he go after he sets up the camera? A5: he backs up and looks at the camera then picks up a cup and drinks from it. Q6: is there anything you think that i am missing? A6: no that would be all the main points. Q7: what kind of noise does she make? A7: she makes some noise when she touches or moves things. Q8: what does she take from the fridge? A8: she takes out a beverage of some kind. i can't tell what it is. Q9: what room is he in? A9: he is in his bedroom. Q10: who is in the video? A10: there is just one man in this video.</p> <p>Descriptions: a woman walks into a kitchen, she cooks the stove and opens a cabinet.</p>	<p>Q2: is she the only one? A2: yes, she is the only person in the video. Q3: does he walks into the room? A3: no, she was there from the beginning. Q4: is he alone in the video? A4: yes, she is the only person in the video. Q5: is he the only one? A5: yes, she is the only one. Q6: does he eat and drink? A7: no, she does not. Q7: does he drink? A8: no, she does not. Q8: is she alone? A8: yes, she is. Q9: does she crossed the table? A9: i don't know. Q10: does she crossed and leave? A10: I don't know</p> <p>Descriptions: a woman walks into a kitchen and takes a drink, he then gets up and walks out of the room.</p>	<p>Q2: is there only one person? A2: yes, there is only one woman in the video. Q3: can you see anyone else in the video? A3: no, there is no one else in the video. Q4: where does she place her coat? A4: I don't know. Q5: she is the only one in the video the entire video? A5: yes, that is correct. Q6: how long is the video? A6: about 30 seconds. Q7: what time of day is it in the video? A7: it looks like to be the daytime. Q8: is she washing clothes? A8: no, I don't think so. Q9: does she pick it back up? A9: I don't know. Q10: are they in the room when the video starts? A10: yes, I think so.</p> <p>Descriptions: a woman is cooking and stirring a pot on the stove, she washes the dishes and leaves the room.</p>
 <p>Input description: a sneezing person is in a living room eating a sandwich watching television. they pour a drink in a glass and drink it with their sandwich.</p>	 <p>Q1: is there just <b>one person</b> in the video? A1: yes, just <b>one person</b> in the video. Q2: is there any <b>sound</b>? A2: yes, there is <b>sound</b> in the video. Q3: does the man <b>speak</b> or is the sound from the <b>tv</b>? A3: there is sound from the tv and him <b>sneezing</b>. Q4: is the man <b>standing</b> the whole time? A4: no he starts out <b>sitting down</b> on the <b>couch</b>. Q5: is he <b>holding</b> something in his hand? A5: a <b>sandwich</b> and then he gets a <b>drink</b>.</p>	<p>Q6: why does he look out the window? A6: you can't see anything out the window so i'm not sure. Q7: how many people are in the video? A7: i just see one man in this video. Q8: does he make any sound? A8: he does not make any sound. Q9: does he eat anything? A9: yes, he started eating the food and got up. Q10: does he eat or drink anything? A10: it looks like he may be chewing gum.</p> <p>Descriptions: a man is sitting on a couch eating a sandwich, he gets up and takes a drink.</p>	<p>Q6: is he the only person? A7: yes, he is the only person in the video. Q7: is he alone? A8: yes, he is. Q8: is he the only person? A8: yes, he is. Q9: does he drink? A9: yes, he drinks something from a glass cup. Q10: does he drink and go? A10: yes, he takes a drink.</p> <p>Descriptions: a man is sitting on a couch eating a sandwich, he gets up and walks out of the room.</p>	<p>Q6: why does he look out the window? A6: I don't know. Q7: does he eat or drink anything? A7: yes, he eats a sandwich and takes a drink from a glass cup. Q8: and then he takes a drink out of a coffee cup? A8: he takes a drink from a glass cup. Q9: does he take a drink of anything? A9: yes, he drinks from a cup. Q10: is he the only person in this video? A10: yes, he is the only person in this video.</p> <p>Descriptions: a man is in a living room eating a sandwich, he then takes a drink from a cup and leaves the room.</p>

Fig. 9. More qualitative results for the simulated human test and real human evaluations.