

Audio-Visual Scene-Aware Dialog

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Lee, S.; Anderson, P.; Essa, I.; Parikh, D.; Batra, D.;
Cherian, A.; Marks, T.K.; Hori, C.

TR2019-048 June 29, 2019

Abstract

We introduce the task of scene-aware dialog. Given a follow-up question in an ongoing dialog about a video, our goal is to generate a complete and natural response to a question given (a) an input video, and (b) the history of previous turns in the dialog. To succeed, agents must ground the semantics in the video and leverage contextual cues from the history of the dialog to answer the question. To benchmark this task, we introduce the Audio Visual Scene-Aware Dialog (AVSD) dataset. For each of more than 11,000 videos of human actions for the Charades dataset. Our dataset contains a dialog about the video, plus a final summary of the video by one of the dialog participants. We train several baseline systems for this task and evaluate the performance of the trained models using several qualitative and quantitative metrics. Our results indicate that the models must comprehend all the available inputs (video, audio, question and dialog history) to perform well on this dataset.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Audio-Visual Scene-Aware Dialog

Huda Alamri¹, Vincent Cartillier¹, Abhishek Das¹, Jue Wang², Stefan Lee¹, Peter Anderson¹,
Irfan Essa¹, Devi Parikh¹, Dhruv Batra¹, Anoop Cherian², Tim K. Marks², Chiori Hori²

¹Georgia Institute of Technology ²Mitsubishi Electric Research Laboratories (MERL)

¹{halamri, vcartillier3, abhshkdz, steflee, peter.anderson, irfan, parikh, dbatra}@gatech.edu

²{juewangj, cherian, tmarks, chori}@merl.com

video-dialog.com

Abstract

We introduce the task of scene-aware dialog. Given a follow-up question in an ongoing dialog about a video, our goal is to generate a complete and natural response to a question given (a) an input video, and (b) the history of previous turns in the dialog. To succeed, agents must ground the semantics in the video and leverage contextual cues from the history of the dialog to answer the question. To benchmark this task, we introduce the Audio Visual Scene-Aware Dialog (AVSD) dataset. For each of more than 11,000 videos of human actions for the Charades dataset. Our dataset contains a dialog about the video, plus a final summary of the video by one of the dialog participants. We train several baseline systems for this task and evaluate the performance of the trained models using several qualitative and quantitative metrics. Our results indicate that the models must comprehend all the available inputs (video, audio, question and dialog history) to perform well on this dataset.

1. Introduction

Developing conversational agents has been a longstanding goal of artificial intelligence (AI). Some recent research has focused on designing and training conversational agents (chatbots) that are visually grounded. Das *et al.* [6] introduced the problem of *visual dialog*, in which the task is to train a model to carry out a conversation in natural language about static images. Developing visually aware conversational agents is an emerging and vibrant area of research that promises to extend the capabilities of conversational agents. However, conversing about a static image is inherently limiting. Many potential applications for conversational agents, such as a helper robot or a smart home, would benefit greatly from understanding the scene in which the agent or a human is operating. The context often cannot be captured only by a still image, as there is important information in the temporal dynamics of the scene as well as

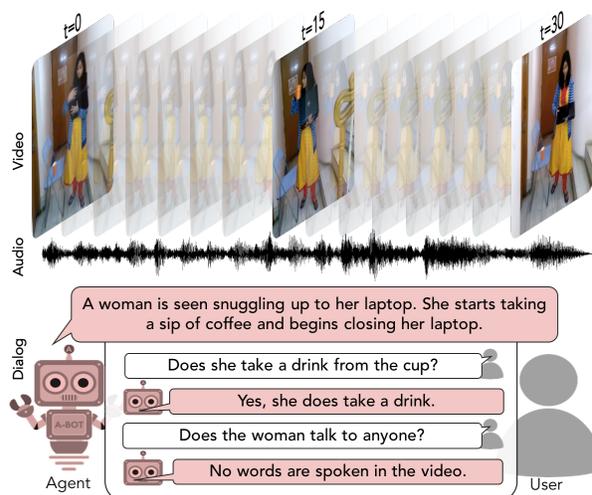


Figure 1: In Audio Visual Scene-Aware Dialog, an agent’s task is to answer in natural language questions about a short video. The agent grounds its responses on the dynamic scene, the audio, and the history (previous rounds) of the dialog.

in the audio. Our goal is to move towards chat agents that are not only visually intelligent but also aware of the sound and temporal dynamics. Such an AI agent can help answer questions such as the following: *Is there any one coming to the door? Can you hear any noise in the room? When did they leave the house? Is my cat still eating?* Answering such questions requires a holistic understanding of the visual and audio information in the scene, including its temporal dynamics.

We introduce the task of scene-aware dialog, as well as a new Audio Visual Scene-aware Dialog (AVSD) Dataset to provide a means for training and testing scene-aware dialog systems. In the general task of scene-aware dialog, the goal of the system is to carry on a conversation with a human about a temporally varying scene, such as a video or a live

scene. In the AVSD Dataset, we are addressing a particular type of scene-aware dialog. Each dialog in the dataset is a temporal sequence of question/answer (QA) pairs about a short video that includes human actions. We defined a specific task for the scene-aware dialog system to learn: Given an input video, the history (the first t QA pairs) of a dialog about the video, and a follow-up question (the $t + 1$ st question in the dialog), the system’s goal is to automatically generate complete and natural responses to the follow-up question.

We aim to use the dataset to explore the compositionality of dynamic scenes and train an end-to-end model to leverage information from the video frames, audio signals, and dialog history. The system should engage in this conversation by providing complete, natural responses to enable real-world applicability. The development of such scene-aware conversational agents represents an important frontier in artificial intelligence. In addition, it holds promise for numerous practical applications, such as retrieving video content from users’ free-form queries and helping visually impaired people understand visual content.

Our contributions include the following:

1. We introduce the task of scene-aware dialog, which is a multimodal semantic comprehension.
2. We introduce a new benchmark for the scene-aware dialog task, the AVSD Dataset, consisting of more than 11,000 conversations that discuss the content (including actions, interactions, sound, and temporal dynamics) of videos of humans.
3. We analyze the performance of several baseline systems on this new benchmark dataset.

2. Related Work

Video Datasets: In the domain of dynamic scene understanding, there is a large body of literature focused on video action classification [9, 10, 16, 21, 37]. Benchmarks like HMDB51 [20], Sports-1M [17], and UCF-101 [32] have been widely used to demonstrate the performance of several machine learning models in the task of action recognition.

To target a broader range of action categories and handle a larger quantity of videos with more realistic settings, Caba Heilbron *et al.* introduced ActivityNet [3], a dataset of 280,000 YouTube videos with more than 200 different human action classes. Another benchmark for the task of human action recognition is Kinetics [18], which consists of 500,000 videos of around 400 action classes. Both ActivityNet and Kinetics are used in the ActivityNet Large Scale Activity Recognition Challenge [13]. Sigurdsson *et al.*, presented the Charades dataset. Charades is a crowdsourced video dataset that was built by asking Amazon Mechanical Turk (AMT) workers to write some scene scripts of daily activities, then asking another group of AMT workers

to record themselves “acting out” the scripts in a “Hollywood style.” The dataset is also temporally annotated with a list of actions and objects in every temporal segment. [30]

Video Captioning: Video captioning is the task of describing the dynamic scene with a natural sentence. The generated description should capture the semantic knowledge of the video, the objects and the actions. It also expresses the spatio-temporal relationship between the dynamics in the scene [12, 29, 38]. Several datasets have been introduced to benchmark the video captioning task [15, 19, 22, 35, 41].

Visual Question Answering: Inspired by the success of image-based question answering [1, 11, 39, 42], some recent work has addressed the task of video-based question answering [22]. MovieQA by Tapaswi M. *et al.* [35] focuses on story comprehension for text and video. MovieQA consists of 14,944 questions about 408 movies. MovieQA and TVQS are challenging benchmarks and have achieved promising results. However, they only focus on the problem of one question and answer. In AVSD we focus on the problem of multiple rounds of questions and answers. Another important point is that the questions and answers in these datasets are generated from the text associated with the movies. However in AVSD, the questions are initiated by a person who does not have access to the video or the text associated with it, resulting conversation that is not biased by the associated textual information.

VisDial: Our work is directly related to the image-based dialog (VisDial) introduced by Das *et al.* [6]. Given an input image, a dialog history, and a question, the agent is required to answer the given question while grounding the answer on the input image and the dialog history. The paper introduces several networks architectures to encode the different input modalities: late fusion LSTM, hierarchical LSTM and a memory network encoder. The model responses were modeled using generative and discriminative models. In this paper, we extend the work from [6] to include more complex modality: video frames and audio signals.

3. Audio Visual Scene-Aware Dialog Dataset

A primary goal of our paper is to create a benchmark for the task of scene-aware dialog. There are several characteristics that we desire for such a dataset: 1) The dialogs should focus on the dynamic aspects of the video, actions and interactions; 2) The answers should tend toward more

| Dataset | # Video Clips | # QA Pairs | Video Source |
|--------------|------------------|------------|--------------|
| TVQA [22] | 21,793 | 152,545 | TV shows |
| MovieQA [35] | 408 | 14,944 | Movies |
| TGIF-QA [15] | 56,720 | 103,919 | Social media |
| VisDial [6] | 120,000 (images) | 1.2 M | N/A |
| AVSD (Ours) | 11,816 | 118,160 | Crowdsourced |

Table 1: Comparison with existing video question answering and visual dialog datasets.

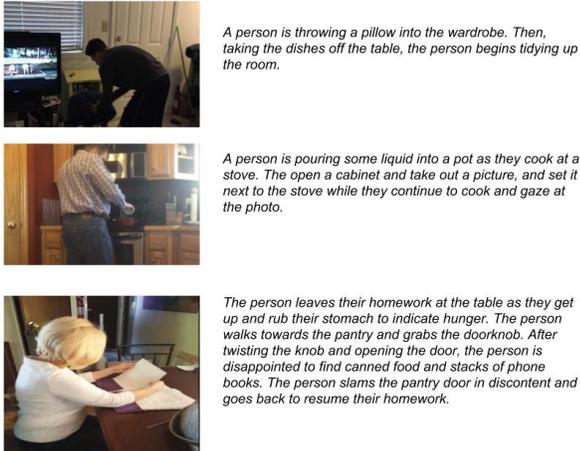


Figure 2: Examples of videos and scripts from the Charades dataset. Each video’s temporally ordered sequence of small events is a good fit for our goal to train a video-based dialog system.

complete explanatory responses rather than brief one- and two-word answers (e.g., not simply yes or no); 3) The conversations should discuss the events in the video in their temporal order.

Table 1 puts the AVSD dataset in context with several other video question answering benchmarks. While AVSD has fewer unique video clips compared to TVQA and MovieQA, which are curated from television and film, our videos are more naturalistic. Moreover, AVSD contains a similar number of questions and answers, but as a part of multi-round dialogs.

Video Content. An essential element to collecting video-grounded dialogs is of course the videos themselves. We choose to collect dialogs grounded in the Charades [30] human-activity dataset. The Charades dataset consists of 11816 videos of daily indoor human activities with an average length of 30 seconds. Each video includes at least two actions. Examples of frames and action scripts for Charades videos are shown in Figure 2. We choose the Charades dataset for two main reasons. First, the videos in this dataset are crowd-sourced on Amazon Mechanical Turk (AMT), so the settings are natural and diverse. Second, each video consists of a sequence of small events that provide AMT Workers (Turkers) with rich content to discuss.

3.1. Data Collection

We adapt the real-time chat interface from [6] to pair two AMT workers to have a conversation about a video from the Charades Dataset (Figure 2). One person, the “Answerer,” is presented with the video clip and the script, and their role is to provide detailed answers to questions about the scene. The other person, the “Questioner,” does not have access to the video or the script, and can only see three frames (one

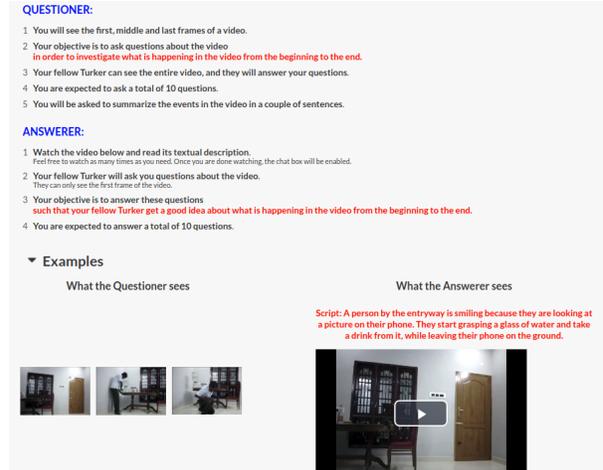


Figure 3: Set of instructions for both AMT workers about their roles of “Questioner” and “Answerer”.

each from the beginning, middle, and end) of the video. The Questioner’s goal is to ask questions to obtain a good understanding of what happens in the video scene. We make several design choices in the data collection interface in order to encourage natural conversations about the activities in the videos.

Investigating Events in Video. To help distinguish this task from previous image and video captioning tasks, our instructions direct the Questioner to “investigate what is happening” rather than simply asking the two Turkers to “chat about the video.” We find that when asked to “chat about the video,” Questioners tend to ask a lot of questions about the setting and the appearance of the people in the video. In contrast, the direction “investigate what is happening” leads Questioners to inquire more about the actions of the people in the video.

Seeding the Conversation. There are two reasons that our protocol provides the Questioners with three frames before the conversation starts: First, since the images provide the overall layout of the scene, they ensure that the conversations are centered around the actions and events that take place in the video rather than about the scene layout or the appearance of people and objects. Second, we found that providing multiple frames instead of a single frame encouraged users to ask about the succession of events. The ordering of the questions in the dialog follows the temporal events in the videos to a certain extent, (e.g., early questions in the dialog ask about events or actions in the first part of the video). Providing the Questioners with these three images achieves both criteria without explicitly dictating Questioners’ behavior; this is important because we want the conversations to be as natural as possible.

Downstream Task: Video Summarization. Once the conversation (sequence of 10 QA pairs) between the Questioner

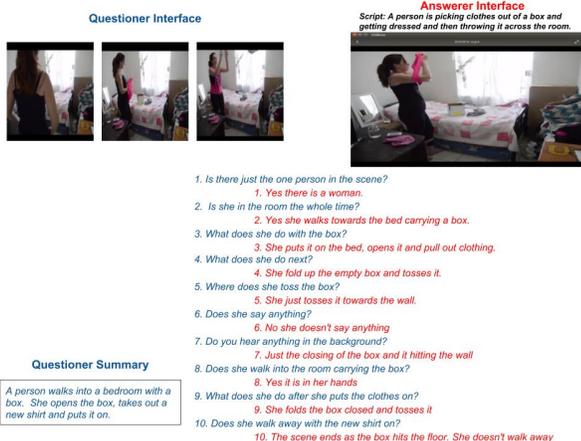


Figure 4: Example conversation between two AMT workers. The Questioner is presented with 3 static images from the video and asks a question. The Answerer, who has already watched the video and read the script, responds. Then the Questioner asks a follow-up question, the Answerer replies, and so on. After 10 rounds of QA, the Questioners provide a written summary of what they think happened in the video based on the conversation.

and Answerer is complete, the Questioners’ final task is to summarize what they think happened in the video. Knowing that this will be their final task motivates the Questioners to ask good questions that will lead to informative answers about the events in the video. In addition, this final downstream task is used to evaluate the quality of the dialog and how informative it was about the video. Figure 3 shows the list of instructions, and the examples provided to help them complete the task.

Worker Qualifications. To ensure high-quality and fluent dialogs, we restrict our tasks on AMT to Turkers with $\geq 95\%$ task acceptance rates, located in North America, and having completed at least 500 tasks already. We further restrict any one Turker from completing more than 200 tasks in order to maintain diversity. In total, 1553 unique workers contributed to the dataset collection effort.

3.2. AVSD Dataset Analysis

In this section, we analyze the new AVSD V.1 Dataset. In total, the dataset contains 11,816 conversations (7985 training, 1863 validation, and 1968 testing), each including a video summary (written by the Questioner after each dialog). There are a total of 118,160 question/answer (QA) pairs. Figure 4 shows an example of our dataset.

We compare the length of AVSD questions and answers with those from VisDial [6] in Figure 5a. As we can see, the answers and questions in AVSD are longer in average than in VisDial. The average length for AVSD questions is 7.85 words and the average answer length is 9.43. In contrast, VisDial questions average 5.12 words and are answered by

| | BLEU ₁ | BLEU ₂ | BLEU ₃ | BLEU ₄ | METEOR | ROUGE _L | CIDEr |
|---------------|-------------------|-------------------|-------------------|-------------------|--------|--------------------|-------|
| video-watcher | 0.638 | 0.433 | 0.287 | 0.191 | 0.223 | 0.407 | 0.429 |
| Questioner | 0.560 | 0.379 | 0.249 | 0.165 | 0.191 | 0.369 | 0.297 |

Table 2: Comparison on different metrics of a video-watcher summary vs the 3 other video-watcher summaries and the Questioner’s summary vs the 3 other video-watcher summaries.

2.9 words answers on average. This shows that dialogs in our set are much more verbose and conversational.

Audio-Related Questions. When the Questioners were presented with 3 frames of the video, the AMT workers asked more questions about the audio track of the video, such as whether there was any music or noise, or whether the people were talking. In 57% of the conversations, there are questions about the audio. Here are some examples of these audio-related questions from the dataset:

Does she appear to talk to anyone? Do you hear any noise in the background? Is there any music? Can you can hear him sneezing? Do the men talk to each other?

Moreover, looking at the burst diagram for questions in Figure 5b we can see questions like “Can / Do you hear ...” and “Is there any sound ...” appear frequently in the dataset.

Temporal Questions. Another common type of questions is about what happened next. As previously noted, the investigation of the temporal sequence of events was implicitly encouraged by our experimental protocol, such as providing the Questioner with three images from different parts of the video. In fact, people asked questions about what happened next in more than 70% of the conversations. Here are some examples of such questions, taken from many different conversations:

Does he do anything after he throws the medicine away? Where does she lay the clothes after folding them? What does he do after locking the door? what does he do after taking a sip? Does he do anything after he sits on the stairs?

Likewise, we see questions like “What happens ...” and “What does he do ...” style questions occur frequently in the dataset as shown in Figure 5b.

Dataset quality. In order to further evaluate dialog quality, we develop and run another study where we ask AMT workers to watch and summarize the videos from the AVSD dataset. The instruction was “Summarize what is happening in the video”. During the dialog data collection, the Questioner is asked, using the same instruction, to summarize the video based on the knowledge gathered through the conversation. We collect 4 summaries per video. We use the BLEU [26], ROUGE [23], METEOR [2] and CIDEr [36] metrics to compare the summaries collected from the video-watcher to the ones collected from the Questioners. In Table

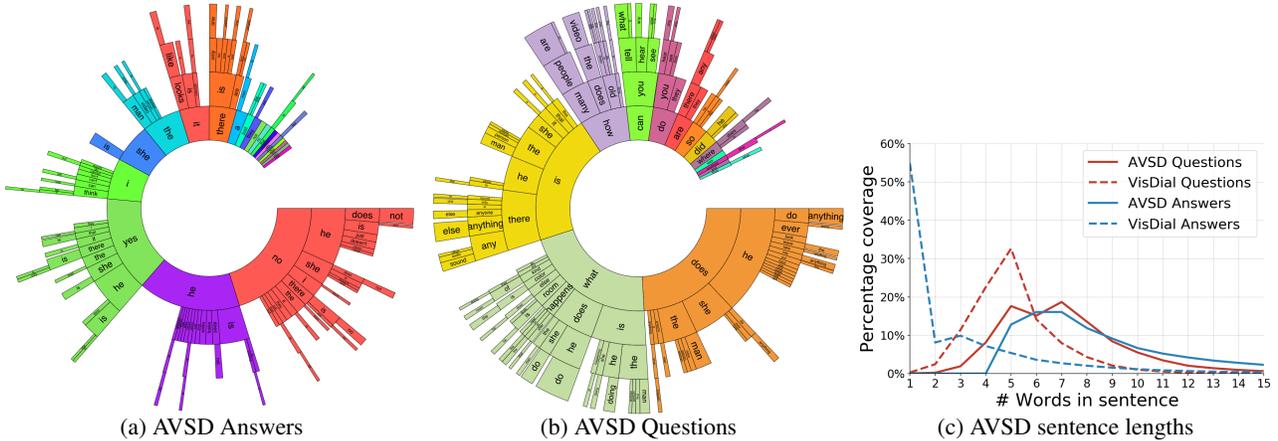


Figure 5: Distribution of first n-grams for AVSD Answers, And AVSD Questions. Distribution of lengths for AVSD questions and answers compared to VisDial.

2, the first row evaluates a randomly selected video-watcher summary vs. three others, and the second row evaluates the Questioner’s summary vs. the same three other video-watcher summaries. Both these numbers are close, demonstrating that the questioners do manage to gain an understanding of the video, similar to having watched it.

4. Model

To demonstrate the potential and the challenges of this new dataset, we design and analyze a video-dialog answerer model. The model takes as input a video, the audio track of the video, a dialog history (which comprises a ground-truth script or video caption followed by the first t QA pairs of the dialog), and a follow-up question (the $t + 1$ st question in the dialog). The model should ground the question in both the video and its audio and use the dialog history to leverage contextual information in order to answer.

Moving away from the hierarchical or memory network encoders common for dialog tasks [6], we opt to present a straightforward, discriminative late-fusion approach for scene-aware dialog that was recently shown to be effective for visual dialog [14]. This choice also enables a fair ablation study for the various input modalities, an important endeavour when introducing such a strongly multimodal task. For this class of model architecture, increases or decreases in performance from input ablation are directly linked to the usefulness of the input rather than to any complications introduced by the choice of network structure (e.g., some modalities having many more parameters than others).

An overview of our model is shown in Figure 6. At a high level, the network operates by fusing information from all of the modalities into a fixed sized representation then comparing this state with a set of candidate answers, selecting the most closely matching candidate as the output answer. In the rest of this section, we provide more details of the model and the input encodings for each modality.

Input Representations. The AVSD dataset is a challenging multimodal reasoning task including natural language, video, and audio. We describe how we represent each of these as inputs to the network. These correspond to the information that was available to the human Answerer in round t of a dialog.

- **Video Caption (C):** Each dialog in AVSD starts with a short natural language description (ground-truth script or caption) of the video contents.
- **Dialog History (DH):** The dialog history consists of the initial caption (C) and each of the question-answer pairs from previous rounds of dialog. At round t , we write the dialog history as $DH_t = (C, Q_0, A_0, Q_1, A_1, \dots, Q_{t-1}, A_{t-1})$. We concatenate the elements of the dialog history and encode them using an LSTM trained along with the late-fusion model.
- **Question (Q):** The question to be answered, also known as Q_{t+1} . The question is encoded by an LSTM trained along with the late-fusion model.
- **Middle Frame (I):** In some ablations, we represent videos using only their middle frame to eliminate all temporal information as a mean to evaluate the role of temporal visual reasoning. In these cases, we encode the frame using a pretrained VGG-16 network [31] that was trained on ImageNet [7].
- **Video (V):** Each AVSD dialog is grounded in a video that depicts people performing simple actions. We transform the video frames into a fixed sized feature using the popular pretrained I3D model [4]. I3D is a 3D convolutional network that pushed state-of-the-art on multiple popular activity recognition tasks [20, 32]. To our knowledge, we are the first to explore the use of I3D for question answering in video-grounded dialog.
- **Audio (A):** We similarly encode the audio from the video using a pretrained AENet model [34]. AENet is a convolutional audio encoding network that operates over long-

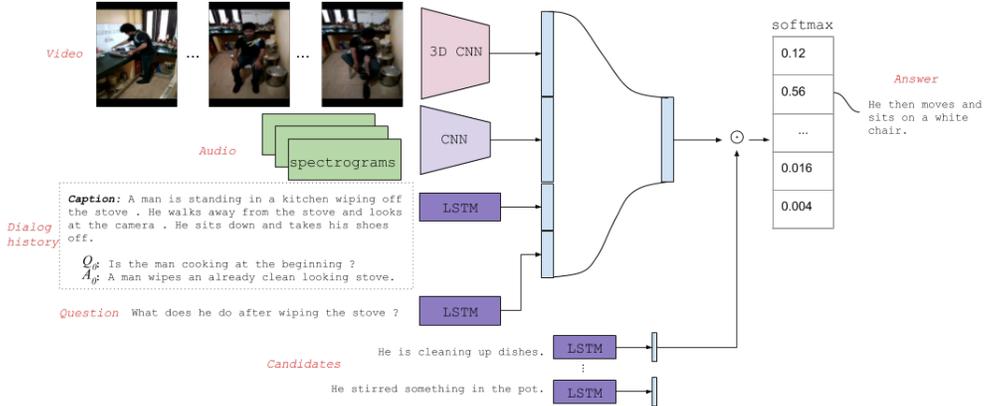


Figure 6: An overview of our late-fusion multimodal network. The encoder takes each input modality and transforms them to a state embedding that is used to rank candidate answers.

time-span spectrograms, and it has been shown to improve activity recognition when combined with video features.

Encoder Network In order to combine the features encoded from these diverse inputs, we follow recent work in visually grounded dialog [14]: we simply concatenate the features and allow fusion to occur through fully-connected layers. More concretely, we can write our network’s computation as:

$$\begin{aligned}
 h_t &= LSTM(DH) \\
 q_t &= LSTM(Q) \\
 i &= I3D(V) \\
 a &= AENet(A) \\
 z &= concat(h_t, q_t, i, a) \\
 e_n &= tanh\left(\sum_{k=1}^K w_{k,n} \times z_k + b_n\right)
 \end{aligned}$$

where h_t , q_t , i , and a are the dialog history, question, video and audio feature embeddings as described above. The embeddings are concatenated to form the vector z which is passed through a linear layer with a tanh activation to form the joint embedding vector e . For any of our ablations of these input modalities, we simply train a network excluding that input, without adjusting the linear layer output size.

Decoder Model We approach this problem as a discriminative ranking task, selecting an output from a set of candidate options, since these approaches have proven to be stronger than their generative counterparts in visual dialog [6]. (However, we note that generative variants need not rely on a fixed answer pool and may be more useful in general deployment.) More concretely, given a set of 100 potential answers $\{\mathcal{A}_t^{(1)}, \dots, \mathcal{A}_t^{(100)}\}$, the agent learns to pick the most appropriate response.

The decoder computes the inner product between a candidate answer embedded with an LSTM and the holistic input

embedding e generated by the encoder. We repeat this for all of the candidate answers, then pass the results through a softmax layer to compute probabilities of all of the candidates. At training time, we maximize the log-likelihood of the correct answer. At test time, we simply rank candidates according to their probabilities and select the argmax as the best response. We can write the decoder as:

$$\begin{aligned}
 a_{t,i} &= LSTM(\mathcal{A}_t^{(i)}) \\
 s_{t,i} &= \langle a_{t,i}, e \rangle
 \end{aligned} \tag{1}$$

where $a_{t,i}$ is the embedding vector for answer candidate $\mathcal{A}_t^{(i)}$, the notation $\langle \cdot, \cdot \rangle$ represents an inner product, and $s_{t,i}$ is the score computed for the candidate based on its similarity to the input encoding e . The vector s_t contains scores for all of the candidates and passes through a softmax during training. The model is then trained with cross-entropy loss to score the ground-truth dialog candidate highly.

Selecting Candidate Answers: Following the selection process in [6], the set of 100 candidates answers consists of four types of answers: the ground-truth answer, hard negatives that are ground-truth answers to similar questions (but different video contexts), popular answers, and answers to random questions. We first sample 50 plausible answers which are the ground-truth answers to the 50 most similar questions. We are looking for questions that start with similar tri-grams (i.e., are of the same type such as “what did he”) and mention similar semantic concepts in the rest of the question. To accomplish this, all the questions are embedded in a common vector space. The question embedding is computed by concatenating the GloVe [27] embeddings of the first three words with the averaged GloVe embedding of the remaining words in the question. We then use Euclidean distance to select the closest neighbor questions to the original question. Those sampled answers are considered as hard negatives, because they correspond to similar questions that were asked in completely different contexts (different video, audio and dialog). In addition, we select the

30 most popular answers from the dataset. By adding popular answers, we force the network to distinguish between purely likely answers and plausible responses for the specific question, which increases the difficulty of the task. The next 19 candidate answers are sampled from the ground-truth answers to random questions in the dataset. The final candidate answer is the ground-truth (human-generated) answer from the original dialog.

Implementation details: Our implementation is based on the visual dialog challenge starter code [8]. The VisDial repository also provides code and model to extract image features. We extract video features using the I3D model [4]. The repository [28] provides code and models fine-tuned on the Charades dataset to extract video features. We subsample 40 frames from the original video and feed them into the rgb pipeline of the I3D model. The frames are sampled to be equally spaced in time. For the audio features, we use the AEnet network [34]. The repository [43] provides code to extract features from an audio signal. We first extract the audio track from the original Charades videos and convert them into 16kHz, 16bit, mono-channel signals. Both the video and audio features have the same dimension (4096).

5. Experiments

Data Splits. Recall from Section 3 that the AVSDv1.0 dataset contains 11k instances split across training (8k), validation (1.5k), and testing (1.5k) corresponding to the source Charades video splits. We present results on the test set.

Evaluation Metrics. Although metrics like BLEU [26], METEOR [2], and ROUGE [23] have been widely used to evaluate dialog [25, 33, 40], there has been recent evidence that they do not correlate well with human judgment [24]. In contrast, we do what [6] does. We instead choose to evaluate our models by checking individual responses at each round in a retrieval or multiple-choice setting. The agent is given a set of 100 answer candidates (generated as described in Section 4) and must select one. We report the following retrieval metrics:

- **Recall@k [higher is better]** that measures how often the ground truth is ranked in the top k choices
- **Mean rank (MR) [lower is better]** of the ground truth answer which is sensitive to overall tendencies to rank ground-truth higher – important in our context as other candidate answers may be equally plausible
- **Mean reciprocal rank (MRR) [higher is better]** of the ground truth answer which values placing ground truth in higher ranks more heavily

We note evaluation even in these retrieval settings for dialog has many open questions. One attractive alternative that we leave for future work is to evaluate directly with human users in cooperative tasks [5].

6. Results and Analysis

In order to assess the challenges presented by the AVSDv1.0 dataset and the usefulness of different input modalities to address them, we present comprehensive ablations of our baseline model with respect to inputs. Table 3 reports the results of our models on AVSDv1.0 test. We find that our best performing models are those that can leverage video, audio, and dialog histories – signaling that the dialog collected in AVSD is grounded in multi-modal observations. In the rest of this section, we highlight noteworthy results.

Language-only Baselines. The first three lines of Table 3 show the language-only models. First, the Answer Prior model encodes each answer with an LSTM and scores it against a static embedding vector learned over the entire training set. This model lacks question information, dialog history, or any form of perception, and acts as a measure of dataset answer bias. Naturally, it performs poorly over all metrics, though it does outperform chance. We also examine a question-only model Q that selects answers based only on the question encoding as well as a question and dialog Q+DH model that also includes the dialog history. These models measure regularities between questions or dialog and answer distributions. We find that access to the question greatly improves performance over the answer prior from 28.54 mean rank to 7.63 with question alone and the addition of the dialog improves this further to 4.72.

Dialog history is a strong signal. The dialog history appears to be a very strong signal – models with it consistently achieve mean ranks in the 4-4.8 range even without additional perception modalities whereas models without dialog history struggle to get below a mean rank of 7. This makes sense, we purposely designed our data collection task to generate dialog focusing on successions of action which

| | Model | MRR | R@1 | R@5 | R@10 | Mean |
|-------------------------------|----------------|--------------|--------------|--------------|--------------|-------------|
| Language Only | Answer Prior | 7.85 | 1.66 | 8.17 | 16.54 | 28.54 |
| | Q | 36.12 | 20.01 | 53.72 | 74.55 | 7.63 |
| | Q + DH | 50.40 | 32.76 | 73.27 | 88.60 | 4.72 |
| Perception w/o Dialog Context | Q + I | 35.12 | 19.08 | 52.36 | 73.35 | 7.90 |
| | Q + V | 39.36 | 22.32 | 59.34 | 78.65 | 6.86 |
| | Q + A | 35.94 | 19.46 | 54.55 | 75.14 | 7.58 |
| | Q + V + A | 38.83 | 22.02 | 58.17 | 78.18 | 7.00 |
| | Q + C + I | 36.77 | 20.30 | 55.28 | 75.72 | 7.44 |
| Full Models | Q + DH + I | 50.52 | 32.98 | 73.26 | 88.39 | 4.73 |
| | Q + DH + V | 53.41 | 36.22 | 75.86 | 89.79 | 4.41 |
| | Q + DH + V + A | 53.03 | 35.65 | 75.76 | 89.92 | 4.39 |

Table 3: Results of model ablations on the AVSDv1.0 test split. We report mean reciprocal rank (MRR - higher is better), recall@k (R@K - higher is better), and mean rank (Mean - lower is better). We find that our best performing model leverages the dialog, video, and audio signals in order to answer questions.



Dialog history: A man with a blue T-shirt holds a white plastic bag by his couch and smiles before leaving the room through a white door. How many people are in the video? just one man in the video. What room is he in? looks like a living room. What is he doing in the living room? he is standing and holding a bag.

Question: What is inside the bag ?

Top 5 answers:

| | |
|------|---|
| 0.32 | not sure it could be takeout food |
| 0.29 | it looks like It I can not tell the make |
| 0.23 | it appears to be a broom |
| 0.05 | on the counter then picks it up and moves it to a different counter |
| 0.03 | it looks like a bedroom. |



Dialog history: The camera pans to the right to show a man sitting in a chair in front of a tv. It then pans left to a little boy walking in the room playing on a cellphone. how does the video start? A boy is sitting in chair watching television. Is the boy in the room the whole time? No he leaves once the other boy walks in .

Question: Does the boy talk with the other boy?

Top 5 answers:

| | |
|------|---|
| 0.61 | Someone says something but I do not think it is either boy. |
| 0.09 | No, he is not talking to anyone. |
| 0.08 | No but I hear someone else speaking in the background. |
| 0.05 | Yes its just the guy. |
| 0.03 | No he is not talking in the video. |

Figure 7: Example using Q+DH+V+A. The left column of the tables in each figure represents the corresponding answer probability.

naturally lead to follow up questions with answers that are strongly dependent on the prior conversation. We note that adding video and audio signals improves over just dialog – providing complementary information to ground questions.

Temporal perception seems to matter. Adding video features (V) consistently leads to improvements for all models. To further tease apart the effect of temporal perception from being able to see the scene in general, we run two ablations where rather than the video features, we encode visual perception using only the middle frame of the video. In both cases, Q+I and Q+DH+I, we see that the addition of static frames hurts performance marginally whereas addition of video features leads to improvements. It seems then that while temporal perception is helpful, models with access to just the middle image learn poorly generalizable groundings. We point out that one confounding factor for this finding is that the image is encoded with a VGG network rather than the I3D encoding for videos.

Audio provides a boost. The addition of audio features generally improves model performance (Q+V to Q+V+A being the exception). Interestingly, we see model performance even when combined with dialog history and video features (Q+DH+V+A) for some metrics, indicating there is still complementary knowledge between the video and audio signals despite their close relationship.

Audio and Temporal Based Questions. Table 4 shows mean rank on a subset of questions. We filter the questions using the two lists of keywords: audio related words {talk hear sound audio music noise} and temporal related words: {after, before, beginning, then, end, start}. We then generated answers to those questions using the three different models Q, Q+A and Q+V and compared which one would lead to higher rank of the ground truth answer.

| | Q | Q+A | Q+V |
|--------------------|------|------|-------------|
| audio questions | 6.91 | 6.69 | 6.52 |
| temporal questions | 7.31 | 7.15 | 5.98 |

Table 4: Mean rank results for the three models Q, Q+A and, Q+V for audio related questions and temporal related questions.

For the audio related questions we can see that although both the Q+A and Q+V outperform the Q model, the visual features seem more useful. This can be easily balanced as it is also unlikely that vision is unnecessary in audio questions. However, the temporal related questions shown to be better answered using the Q+V model which confirms our intuition. The Q+A helps only slightly (7.15 vs 7.31) but the Q+V yields to better improvements (5.98 vs 7.31).

Qualitative Examples. Figure 8 shows two examples using the setup Q+DH+V+A. The first column in the answer table of each figures is the answer probability. The ground truth answer is highlighted in red.

7. Conclusion

We introduce a new AI task: Audio Visual Scene-Aware Dialog, where the goal is to hold a dialog by answering a user’s questions about dynamic scenes in a natural language manner. We collected the Audio Visual Scene-Aware Dialog dataset through a two-person chat protocol on more than 11,000 videos of human actions. We also developed a model and experimented on many ablation studies to highlight the quality and complexity of the data collected. Our results show that the dataset is very rich having all the different modalities playing a role in tackling this task. We believe our dataset can serve in evaluating progress in audio and visual intelligence agents.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2
- [2] S. Banerjee and A. 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. 4, 7
- [3] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2
- [4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017. 5, 7
- [5] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2017. 7
- [6] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017. 1, 2, 3, 4, 5, 6, 7
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5
- [8] K. Desai, A. Das, D. Batra, and D. Parikh. Visual dialog challenge starter code. <https://github.com/batra-mlp-lab/visdial-challenge-starter-pytorch>, 2018. 7
- [9] C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*, pages 3468–3476, 2016. 2
- [10] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016. 2
- [11] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015. 2
- [12] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 2
- [13] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Khrisna, S. Buch, and C. D. Dao. The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*, 2018. 2
- [14] U. Jain, S. Lazebnik, and A. G. Schwing. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5, 6
- [15] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. Tgifqa: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawaii*, pages 2680–8, 2017. 2
- [16] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013. 2
- [17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2
- [18] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset, 2017. 2
- [19] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2
- [20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011. 2, 5
- [21] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011. 2
- [22] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2
- [23] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 4, 7

- [24] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016. 7
- [25] R. Lowe, N. Pow, I. V. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 285, 2015. 7
- [26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 4, 7
- [27] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6
- [28] A. Piergiovanni. I3d models trained on kinetics. <https://github.com/piergiaj/pytorch-i3d>, 2018. 7
- [29] R. Shetty and J. Laaksonen. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. *arXiv preprint arXiv:1512.02949*, 2015. 2
- [30] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2, 3
- [31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [32] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 5
- [33] A. Sordani, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, 2015. 7
- [34] N. Takahashi, M. Gygli, and L. Van Gool. Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, 20(3):513–524, 2018. 5, 7
- [35] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 2
- [36] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4
- [37] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011. 2
- [38] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016. 2
- [39] L. Yuu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2461–2469, 2015. 2
- [40] K. Zhou, S. Prabhume, and A. W. Black. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, 2018. 7
- [41] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017. 2
- [42] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016. 2
- [43] znaoya. Aenet: audio feature extraction. <https://github.com/znaoya/aenet>, 2017. 7

Appendix Overview

This supplementary document is organized as follow:

- Sec. **H** qualitative examples from AVSD.
- Sec. **I** snapshots of our Amazon Mechanical Turk interface that served collecting the video summaries along with some examples.

H. Qualitative examples from our dataset.

In this section, we discuss the model responses to several types of challenging and interesting questions in our dataset. In a video-based dialog, questions can be about audio, visual appearance, temporal information or actions. We examine the model responses for these question based on different input modalities, examples are randomly select from the test set.

H.1. Examples with Q+DH+V+A

Figures **8a** and **8b** show examples of audio related questions. In figure **8a**, although model ranked the ground truth answer at the third position, the two top ranked answers can also be valid answers to the given question "Dose he say any thing?" . In **8b**, 3 out of the top 4 ranked answers can be a valid answers as well. They all answered 'no' to the question. This highlights the deep understanding of the question and context. Figures **8c**, **8d** and **8e** are examples of visual-related questions. In figure **8e**, the model must determine a person's age by leveraging visual cues from the video frames. An important type of questions in video-based dialog is the temporal-based question. Examples of this type are shown in Figure **8d** and **8f**. Figures **8g** and **8h** show interesting and challenging questions about the general scene. In our dataset, there are no binary answers "yes"

or "no", the Answers were asked to provide further details about their responses.

H.2. Examples comparing setups Q, Q+V, Q+A and Q+DH+V+A

Figure **9** shows examples comparing results between models Q, Q+V, Q+A and Q+DH+V+A. The GT rank is the rank of the ground truth answer for the corresponding model. The top answer is the first ranked answer for the corresponding model. The red highlights the best model. In figure **9a** the question is audio related question and the Q+A model performs better. The question from the example in figure **9b** is visual related question and the Q+V model performs best. Figure **9c** presents a temporal related question best answered by the Q+V model. This highlights the value of each modality in the dataset.

I. Summaries Interface.

The data collection process of AVSD included a downstream task, where the Questioners had to write a summary of what they think happened in the video, based on the conversation they had about it. To evaluate the quality of these conversations, we ran a separate study case on AMT. We asked 4 people to watch the video and write a summary describing all the events in the video. Figure **10** shows the interface for this task. People where presented with example of the video and the script for that video. We then compared these summaries with the one written by the questioner.

Figure **11** shows some examples of the 4 summaries collected in this study (first four rows) and the summary written by the Questioner at the end of the dialog (last row). In these examples, we see that the summary written by the Questioner captures most of the events described in the 4 summaries.



Dialog history: A man walks into a hallway. He first picks a camera up off of the shelf and then sets it back down. Then the man walks over to a vacuum cleaner and begins to vacuum the hallway. How does the video start? Man gets into the room. Is he holding anything? No, but he goes to check the camera. What else does he do? Picks it up, looks at it then starts vacuuming.

Question: Does he say anything?

Top 5 answers:

| | |
|------|--|
| 0.29 | No this man never spoke a single word. |
| 0.21 | Nothing to say anything and any sounds not come. |
| 0.16 | No, i didn't hear anything. |
| 0.06 | He is saying something in a foreign language. |
| 0.05 | No, I can not hear him say anything. |

(a)



Dialog history: A man walks into a hallway. He first picks a camera up off of the shelf and then sets it back down. Then the man walks over to a vacuum cleaner and begins to vacuum the hallway. How does the video start? Man gets into the room. Is he holding anything? No, but he goes to check the camera. What else does he do? Picks it up, looks at it then starts vacuuming. Does he say anything? No, I didn't hear anything. Does he do anything else? No, he does not do anything else. What is the man wearing? He is wearing tank top and shorts.

Question: Is there anything in the room?

Top 5 answers:

| | |
|------|--|
| 0.46 | A shelf where is the camera. |
| 0.17 | Yes there is a closet in the room. |
| 0.08 | Yes, when he wipes the sweat, he continues to play. |
| 0.05 | It look like a man. |
| 0.04 | He pushes the table to the side and puts a chair in its place. |

(c)



Dialog history: A woman is standing at the table reading a book. She is drinking a cup of coffee and eating something on the table. How many people are there? Just one person is what I see. What do they do? They appear to drink out of a cup. Can you tell what they are drinking? no, I cannot tell what they are drinking. Do they talk? No, I do not see them talk at all. Are they reading? Yes, I see them reading a book. Do they look at the camera? No, I do not believe they do. Is this in a kitchen? No, I do not think it is in a kitchen.

Question: Is the girl a teenager?

Top 5 answers:

| | |
|------|---|
| 0.81 | I do not think she is a teenager. |
| 0.11 | No this is an adult. |
| 0.02 | Maybe 30s or late 20s. |
| 0.01 | Yes from what I can see. |
| 0.01 | It looks like a controller of some kind but because it is so dark it is hard to tell. |

(e)



Dialog history: A man is watching tv as he grabs a piece of bread and takes a bit. He grabs a cup and drinks from it as he continues to watch tv. What is happening in this video? A young man is watching TV. What is he watching on TV? I can not tell. It only shows the tv for a second. How old is this young man? this young man is 21. What room is he in? He is in his bedroom. Does he speak at all in this video? He does not say anything. What color is his shirt? He is wearing a purple shirt. Is he eating any food during the TV show? He picks up a piece of bread and eats it. Is it daytime or night where he is? I can not tell, there is no clock or window. What nationality would you guess he was? He appears to be Indian.

Question: Is there anything else going on there I should know about?

Top 5 answers:

| | |
|------|---|
| 0.59 | After he sneezes a few times, he closes the medicine cabinet, and the video ends. |
| 0.09 | The only other thing is that he takes a sip of water from a cup at the end. |
| 0.08 | Guy is in front of large open curtain window standing while pulling off sweatshirt. |
| 0.04 | He says hello, UNK it going. |
| 0.04 | Nothing else is happening in this video. |

(g)



Dialog history: A man with a blue T-shirt holds a white plastic bag by his couch and smiles before leaving the room through a white door. How many people are in the video? Just one man in the video. What room is he in? Looks like a living room. What is he doing in the living room? He is standing and holding a bag. What is inside the bag? Not sure it could be takeout food. What else does he do? He stands for sometime before walking into another room. Does he start off in the room? He starts off walking into living room.

Question: Is there any sound?

Top 5 answers:

| | |
|------|----------------------------------|
| 0.27 | There is no sound in video. |
| 0.13 | No don't know why he does that. |
| 0.11 | No there is no sound. |
| 0.10 | There is no audio of importance. |
| 0.09 | Yes there is sound in the video. |

(b)



Dialog history: The man is walking into the bathroom and he closes the door. The man is fixing his clothing in the mirror.

Question: Where in the house does the video begin?

Top 5 answers:

| | |
|------|--|
| 0.41 | It starts with a man walking into the kitchen. |
| 0.36 | The door in the room looks like the front door, so it may be the entry room. |
| 0.11 | In a hallway with closets. |
| 0.06 | He walks into the bathroom. |
| 0.02 | It looks like a stairway down. |

(d)



Dialog history: A guy walks into a house and removes his shoes before entering. He closes the door behind himself and smiles at the door. Is the one man the only person in the video? There is only one man in the video. What does the man do? He walks towards a door and takes his shoes off before he steps inside.

Question: Then what does he do?

Top 5 answers:

| | |
|------|--|
| 0.59 | He washes men's shirts and put them up to dry on the back of the chair. |
| 0.09 | He grabs a broom and starts sweeping the floor. |
| 0.08 | He steps inside and he closes the door slowly. |
| 0.04 | To the other side of the place. |
| 0.04 | He is staring down the entire thing, when he takes off his coat he neatly folds it and places. |

(f)



Dialog history: A man is looking at his laptop and then presses a button. The man continues using the laptop while standing and eating. How many folks are in this scene today? There is only one person. What room is he in? I think he is in the dining room. What happens in this scene today? The man is using his laptop while standing and eating something. What does he do next? That is the only thing he does. How old would you say he is? I think he is around 20. Do you think he is working or just messing around? I think he is just messing around. Does he look hurried or nervous? No, he does not look nervous at all.

Question: Is the room clean?

Top 5 answers:

| | |
|------|--|
| 0.19 | Yes very clean and organized. |
| 0.12 | It looks clean enough. It is hard to see. |
| 0.11 | It looks pretty clean and organized. |
| 0.10 | Yes, the closet looks clean, except for the pillow that was on the ground. |
| 0.10 | Yeah a bit with stuff on the floor. |

(h)

Figure 8: Examples using Q+DH+V+A. The left column of the tables in each figure represents the corresponding answer probability.



Dialog history: A man is fixing the comforter on a bed, then inspects a picture frame on the nightstand. He then takes 2 pillows out of a cardboard box and places them on the bed. How many people are in the clip? There is just one guy. Is he there in the beginning? Yes, he is there the whole time. What does he start off doing? He is fixing his comforter on the bed. What does he do when he's done making his bed? He checks the picture frame on the nightstand. What's the guy's next move? Then he takes two pillows out of a box by the bed. What does he do with them? He puts them on the bed. Does he do anything else after that? No, that is all that he does. Does he speak at all? There is no talking in the clip.

Question: Is there any background noise?

GT answer: There is no noise either.

| | GT rank | Top answer |
|----------------|---------|--|
| Q | 6 | Mostly static noise in the background other than coughing. |
| Q + V | 9 | There is no background noise of note. |
| Q + A | 3 | Just the noises of running around. |
| Q + DH + V + A | 1 | There is no noise either. |

(a)



Dialog history: A man is in the hallway talking to someone on the phone. The man reaches his hand out, and grabs the door handle next to him and opens the doors one side at a time. What is happening in the video? A man is talking on a phone and opening a closet door where washer and dryer are located. Can you hear what he is saying? He is talking low, but it sounds like laundry he is talking about. Is that all he do? He opens the dryer to see if any laundry is in there. Is there anyone else in the video? No, he is the only one in the video. So this is like a laundry room or is it like a general area? It is a hallway and the washer dryer are behind a door. Does he walk around? He stays in front of the doors while talking on the phone.

Question: Is he there from the beginning of the video or he walks in?

GT answer: He is there in the hall from the beginning.

| | GT rank | Top answer |
|----------------|---------|---|
| Q | 12 | Yes, he does walk inside. |
| Q + V | 2 | He is standing there from the beginning. |
| Q + A | 13 | Yes, he walks towards the door. |
| Q + DH + V + A | 1 | He is there in the hall from the beginning. |

(b)



Dialog history: A man reaches high up and puts something on a wall. After he is done he starts sweeping the floor with a broom. Is there only one person in the scene? Yes, he is the only person present in the video. What happens at the beginning? The man is watching at something he's holding in his hands. Do you hear any noise throughout the scene? No, there is audio but it's mostly silent.

Question: Great, so he fiddles with something in his hands then what happens?

GT answer: The man puts it in the wall, it looks like its a lightbulb but not really sure.

| | GT rank | Top answer |
|----------------|---------|--|
| Q | 10 | He picks the book up and flips through it. |
| Q + V | 2 | He opens the door and gazes out of it. |
| Q + A | 16 | He picks the book up and flips through it. |
| Q + DH + V + A | 4 | He carries it out of the room. |

(c)

Figure 9: Comparison between models Q, Q+V, Q+A and Q+DH+V+A. The GT rank is the rank of the ground truth answer for the corresponding model. The top answer is the first ranked answer for the corresponding model. The red highlights the best model.

Summarize a Short Video.

▼ **Instructions**

In this task, you will provide a descriptive summary of what happens in a short video.

- 1 You will watch a short video.
- 2 Your objective is to summarize the video in couple of sentences describing what is happening in the video
- 3 Your summary should describe the sequence of actions that take place in the video, rather than people or objects appearances.

▼ **Examples**



Summary: A person by the entryway is smiling because they are looking at a picture on their phone. They start grasping a glass of water and take a drink from it, while leaving their phone on the ground.

Waiting to Accept the HIT...

Figure 10: Summaries data collection interface on AMT.



| |
|--|
| A person runs up a flight of steps holding a pillow. Another person walks down the steps holding something in his hand. |
| A young man begins to run up the stairs with a pillow in his hands, crossing paths with an older gentleman coming down the stairs. |
| A young man comes running up a spiral staircase with a pillow in his hand, while an older man comes down the stairs, carrying an object in his right hand. |
| A kid comes running in with loud flip flops. He's carrying a pillow and runs up the stairs. Meanwhile another guy is coming down the stairs. |
| An elder man is climbing the stairs and passing the boy and he is holding a sandwich in his hands. |

(a)



| |
|---|
| A woman is standing in the bathroom in front of her laptop. The woman works on the laptop and is by herself. |
| A girl is in the bathroom looking at her laptop. She sets it down and just begins typing away at it. She is staring at her laptop. |
| An person is typing on a computer in an bathroom. she stands up and replants the device on the counter. she continues to tap the keys. |
| A woman is standing in a bathroom holding an open laptop. She then places the laptop on the counter and begins to type on the keyboard. |
| A person uses their laptop next to the sink. |

(b)



| |
|---|
| Two ladies are standing outside. One has a plate in her hand and is drinking. The other is sweeping the doorway. The first one walks in the house while sipping drink. |
| A woman is sweeping off steps while a red headed woman, carrying a glass of water and a plate with a piece of bread takes one step up at a time, while taking a drink each time. |
| A woman stands near another woman who is cleaning the floors. The woman who is not cleaning is carrying water and a plate of food into her walkway. |
| Two women are outside, the one woman is sweeping off the step while the other lady is holding a plate and a cup. The lady holding the plate and cup walks up the steps while the other lady continues to sweep, she then drinks from the cup. |
| While a woman is sweeping her front steps and entry a woman in a Blue Sari approaches carrying bread and drinking water. She passes the sweeping woman and enters. |

(c)

Figure 11: Comparison between different video summaries. The first 4 rows are summaries written by people after watching the entire video. Last row is summary written by the questioner who did not watch the video.