Audio Visual Scene-Aware Dialog

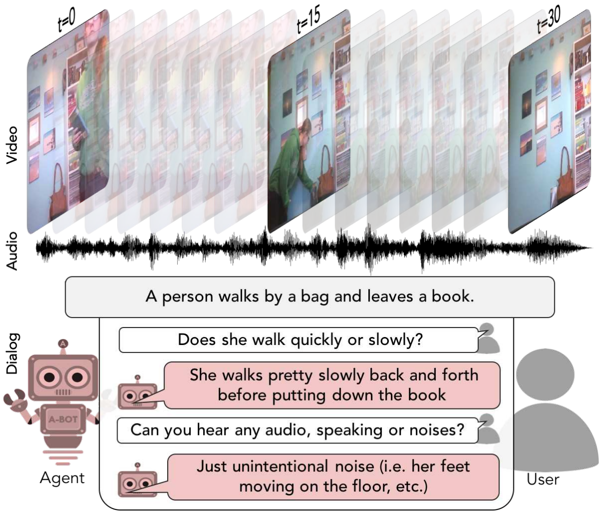
Huda Alamri[1], Vincent Cartillier[1], Abhishek Das[1], Jue Wang[2] , Anoop Cherian[2], Irfan Essa[1],
Dhruv Batra[1], Tim K. Marks[2], Chiori Hori[2], Peter Anderson[1], Stefan Lee[1], Devi Parikh[1]
[1]Georgia Institute of Technology [2]Mitsubishi Electric Research Laboratories (MERL)
[1]{halamri, vcartillier3, abhshkdz, irfan, dbatra, peter.anderson, steflee, parikh}@gatech.edu
[2]{juewangj, cherian, tmarks, chori}@merl.com
video-dialog.com

Abstract

We introduce the task of scene-aware dialog. Our goal is to generate a complete and natural response to a question about a scene, given video and audio of the scene and the history of previous turns in the dialog. To answer successfully, agents must ground concepts from the question in the video while leveraging contextual cues from the dialog history. 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 from 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 both qualitative and quantitative metrics. Our results indicate that models must utilize all the available inputs (video, audio, question, and dialog history) to perform best on this dataset.

1. Introduction

Developing conversational agents has been a longstanding goal of artificial intelligence (AI). For many human-computer interactions, natural language presents an ideal interface, as it is fully expressive and requires no user training. One emerging area is the development of *visually aware* dialog systems. Das *et al*. [6] introduced the problem of *visual dialog*, in which a model is trained to carry out a conversation in natural language, answering questions about an image. For a given question, the system has to ground its response in the input image as well as the previous utterances. However, conversing about a static image is inherently limiting. Many potential applications for conversational agents, such as virtual personal assistants and assistive technologies for the visually impaired, would benefit greatly from understanding the scene in which the agent is operating. This context often cannot be captured in a single image, as there is important information in the temporal dynamics of the scene as well as in the audio.



A person walks by a bag and leaves a book.

Does she walk quickly or slowly?

She walks pretty slowly back and forth before putting down the book

Can you hear any audio, speaking or noises?

Just unintentional noise (i.e. her feet moving on the floor, etc.)

Figure 1: In Audio Visual Scene-Aware Dialog, an agent's task is to answer 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, dialog history, which begins with a short script of the scene.

Our goal is to move towards conversational agents that are not only visually intelligent but also aware of temporal dynamics. For example, a security guard (G) might have the following exchange with an AI agent: "G: Has there been anyone carrying a red handbag in the last week in Sector 5? AI: Yes, a woman in a blue suit. G: Do any of the exit cameras show her leaving with it? AI: No. G: Did anyone else pick it up?". Answering such questions requires a holistic understanding of the visual and audio information in the scene, including temporal dynamics. Since human communication is rarely single-shot, an understanding of sequential dialog (e.g., what *her* and *it* refer to) is also required.

We introduce the task of scene-aware dialog, as well as the

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. In the AVSD Dataset, we are addressing a particular type of scene-aware dialog. Each dialog in the dataset is a sequence of question/answer (QA) pairs about a short video; each video features a person performing everyday activities in a home environment.

We defined a specific task for the scene-aware dialog system to learn: Given an input video, the history of a dialog about the video (consisting of a short script plus the first $t-1$ QA pairs), and a follow-up question (the tth question in the dialog), the system's goal is to generate a correct response to the follow-up question. We aim to use the dataset to explore the compositionality of dynamic scenes and to 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 and 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 video retrieval 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 task.
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 human-centered activities.
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, a large body of literature focuses on action classification. Some important benchmarks for video-action recognition and detection are: HMDB51 [15] Sports-1M [12] and UCF-101 [27]. The ActivityNet [3] and Kinetics [13] datasets target a broader range of human-centered action categories. Sigurdsson *et al.* [25] presented the Charades dataset. Charades is a crowd-sourced 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." In our work, Charades videos were used to collect conversations about the activities in the videos. Video-based captioning is another exciting research area, and there are several datasets introduced to benchmark and evaluate this task [14, 24].

Video-based Question Answering: Inspired by the success of image-based question answering [1, 9, 32, 34], recent work has addressed the task of video-based question answering [11,20,30]. MovieQA [16] and TVQA [30] evaluate the video-based QA task by training end-to-end models to answer multiple-choice questions by leveraging cues from video frames and associated textual information, such as scripts and subtitles. While this one-shot question answering setting is more typical in existing work, we find this structure to be unnatural. Our focus in AVSD is on settings involving multiple rounds of questions that require natural free-form answers.

Visual Dialog: Our work is directly related to the image-based dialog task (VisDial) introduced by Das *et al.* [6]. Given an input image, a dialog history, and a question, the agent is required to answer the question while grounding the answer on the input image and the dialog history. In [6], several network architectures are introduced to encode the different input modalities: late fusion, hierarchical recurrent encoder, and memory network. In this work, we extend the work from [6] to include additional complex modalities: video frames and audio signals.

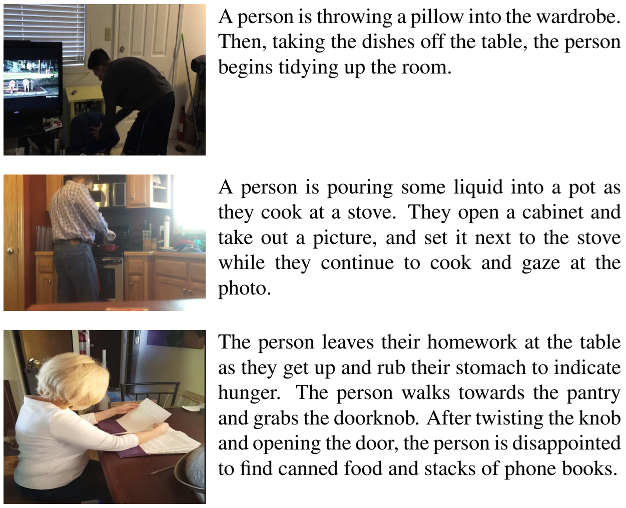
3. Audio Visual Scene-Aware Dialog Dataset

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 (i.e., actions and interactions); 2) The answers should be complete explanatory responses rather than brief one- or two-word answers (e.g., not simply yes or no); 3) The dialogs should discuss the temporal order of events in the video.

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 [25] human-activity dataset. The Charades dataset consists of 11,816 videos of everday indoor human activities with an average length of 30 seconds. Each video includes at least two actions. Examples of frames and scripts for Charades videos are shown in Figure 2. We choose the Charades dataset for two main reasons. First, the videos in this dataset

Dataset	# Video Clips	# QA Pairs	Video Source	Answers
TVQA [16]	21,793	152,545	TV shows	Multi-Choice
MovieQA [30]	408	14,944	Movies	Multi-Choice
TGIF-QA [11]	56,720	103,919	Social media	Multi-Choice
VisDial [6]	120K (images)	1.2 M	N/A	Free-Form
AVSD (Ours)	11,816	118,160	Crowdsourced	Free-Form

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



A person is throwing a pillow into the wardrobe. Then, taking the dishes off the table, the person begins tidying up the room.

A person is pouring some liquid into a pot as they cook at a stove. They open a cabinet and take out a picture, and set it next to the stove while they continue to cook and gaze at the photo.

The person leaves their homework at the table as they get up and rub their stomach to indicate hunger. The person walks towards the pantry and grabs the doorknob. After twisting the knob and opening the door, the person is disappointed to find canned food and stacks of phone books.

Figure 2: Examples of videos and scripts from the Charades [25] dataset. Each video's temporally ordered sequence of small events is a good fit for our goal to train a video-based dialog system.

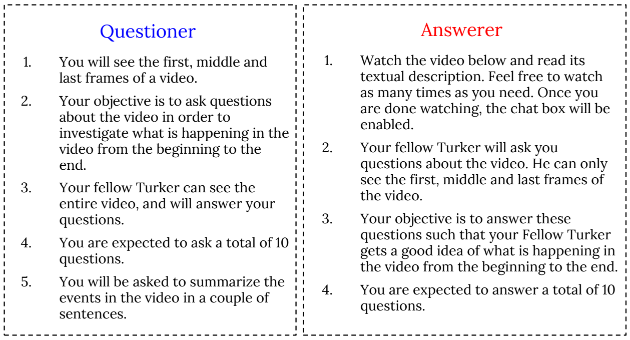
are crowd-sourced on Amazon Mechanical Turk (AMT), so the settings are natural and diverse. Second, each video consists of a sequence of small events that provide AMT Workers (Turkers) with rich content to discuss.

3.1. Data Collection

We adapt the real-time chat interface from [6] to pair two AMT workers to have an English-language 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 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 considered several design choices for the chat interface and instructions, in order to encourage natural conversations about events 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 sequence of events. 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.



Questioner

1. You will see the first, middle and last frames of a video.
2. Your objective is to ask questions about the video in order to investigate what is happening in the video from the beginning to the end.
3. Your fellow Turker can see the entire video, and will answer your questions.
4. You are expected to ask a total of 10 questions.
5. You will be asked to summarize the events in the video in a couple of sentences.

Answerer

1. Watch the video below and read its textual description. Feel free to watch as many times as you need. Once you are done watching, the chat box will be enabled.
2. Your fellow Turker will ask you questions about the video. He can only see the first, middle and last frames of the video.
3. Your objective is to answer these questions such that your Fellow Turker gets a good idea of what is happening in the video from the beginning to the end.
4. You are expected to answer a total of 10 questions.

Figure 3: Instructions provided to AMT workers explaining the roles of "Questioner" and "Answerer."

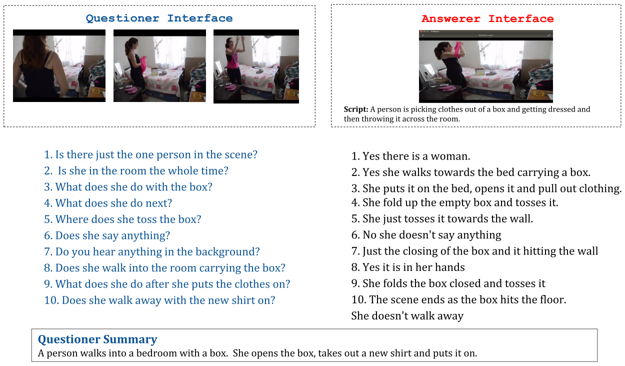
Downstream Task: Video Summarization. Once the conversation (sequence of 10 QA pairs) between the Questioner and Answerer is complete, the Questioner's final task is to summarize what they think happened in the video. Knowing that this will be their final task motivates the Questioner to ask good questions 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 provided to AMT workers.

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, who have 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.

Table 1 puts the Audio Visual Scene-aware Dialog (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.

3.2. AVSD Dataset Analysis

In this section, we analyze the new AVSDv1.0 Dataset. In total, the dataset contains 11,816 conversations (7,985 train-



Script: A person is picking clothes out of a box and getting dressed and then throwing it across the room.

1. Is there just the one person in the scene?
2. Is she in the room the whole time?
3. What does she do with the box?
4. What does she do next?
5. Where does she toss the box?
6. Does she say anything?
7. Do you hear anything in the background?
8. Does she walk into the room carrying the box?
9. What does she do after she puts the clothes on?
10. Does she walk away with the new shirt on?

1. Yes there is a woman.
2. Yes she walks towards the bed carrying a box.
3. She puts it on the bed, opens it and pull out clothing.
4. She fold up the empty box and tosses it.
5. She just tosses it towards the wall.
6. No she doesn't say anything
7. Just the closing of the box and it hitting the wall
8. Yes it is in her hands
9. She folds the box closed and tosses it
10. The scene ends as the box hits the floor.
She doesn't walk away

Questioner Summary
A person walks into a bedroom with a box. She opens the box, takes out a new shirt and puts it on.

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. After 10 rounds of QA, the Questioners provides a written summary of what they think happened in the video based on the conversation.

ing, 1,863 validation, and 1,968 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 from our dataset. More examples can be found in the supplementary section.

Lengths of Questions and Answers. We compare the length of AVSD questions and answers with those from VisDial [6] in Figure 5c. Note that the answers and questions in AVSD are longer on average. The average length for AVSD questions and answers is 7.9 and 9.4 words, respectively. In contrast, VisDial questions average 5.1 words and are answered in 2.9 words on average. This shows that the dialogs in our dataset are more verbose and conversational.

Audio-Related Questions. In 57% of the conversations, there are questions about the audio, such as whether there was any music or noise, or whether the people were talking. 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? Is there any other noise like a TV or music?*

Moreover, looking at the burst diagram for questions in Figure 5b we can see that 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*. In fact, people asked questions about what happened next in more that 70% of the conversations. 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 image frames from different parts of the video. Here are some examples of such questions, taken from different conversations:

	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$	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.

> *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?*

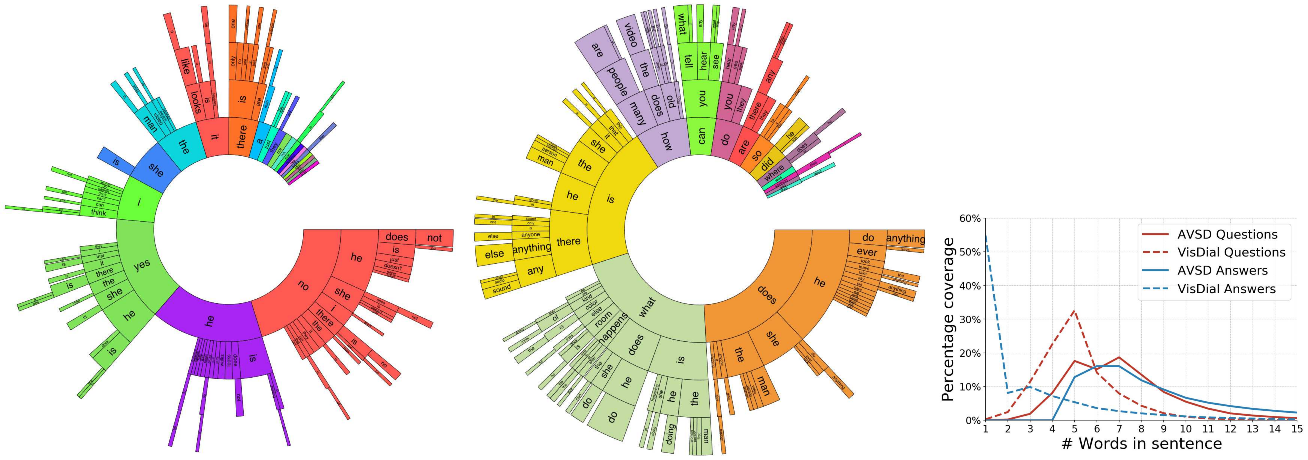
Likewise, we see that questions such as "What happens ..." and "What does he do ..." occur frequently in the dataset, as shown in Figure 5b.

Dataset Quality. In order to further evaluate dialog quality, we ran another study where we asked AMT workers to watch and summarize the videos from the AVSD Dataset. The instruction was "Summarize what is happening in the video". We collected 4 summaries per video and used the BLEU [21], ROUGE [17], METEOR [2] and CIDEr [31] metrics to compare the summaries collected from the video-watcher to the ones provided by the questioners at the end of each conversation. In Table 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 gain an understanding of the scene from the dialog that is comparable to having watched the video.

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 the ground-truth script from the Charades dataset and the first $t-1$ QA pairs of the dialog), and a follow-up question (the tth 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 [10]. 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).



(a) AVSD Answers (b) AVSD Questions (c) AVSD sentence lengths

Figure 5: Distribution of first n-grams in the AVSD Dataset for (a) AVSD answers and (b) AVSD questions. (c) Distribution of lengths for questions and answers in AVSD compared to those in VisDial [6].

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-size 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 provides 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 Script (S)**: Each dialog in AVSD starts with a short natural language description of the video contents (i.e., the Charades ground-truth script).
- **Dialog History (DH):** The dialog history consists of the initial video script (S) and each of the question-answer pairs from previous rounds of dialog. At round t, we write the dialog history as DH$_t$=$(S, Q_0, A_0, Q_1, A_1, \ldots Q_{t-1}, A_{t-1})$. 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. 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, in order to evaluate the role of temporal visual reasoning. In these cases, we encode the frame using a pretrained VGG-16 network [26] 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 achieved state-of-the-art performance on multiple popular activity recognition tasks [15, 27].
- **Audio (A):** We similarly encode the audio track from the video using a pretrained AENet [29]. AENet is a convolutional audio encoding network that operates over long-time-span spectrograms. It has been shown to improve activity recognition when combined with video features.

Encoder Network. In order to combine the features from these diverse inputs, we follow recent work in visually grounded dialog [10]: simply concatenate the features, and allow fusion to occur through fully-connected layers. More concretely, we can write our network as:

$$
\begin{aligned}
h_t &= \text{LSTM}(\text{DH}) \qquad & q_t &= \text{LSTM}(\text{Q}) \\
i &= \text{I3D}(\text{V}) \qquad & a &= \text{AENet}(\text{A})
\end{aligned}
$$

$$
\begin{aligned}
z &= \text{concat}(h_t, q_t, i, a) \\
e_n &= \tanh\left(\textstyle\sum_k w_{k,n} \times z_k + b_n\right),
\end{aligned}
$$

where h_t, q_t, i, and a are the dialog history, question, video, and audio feature embeddings 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. (Here k and n respectively index elements of the vectors z and 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

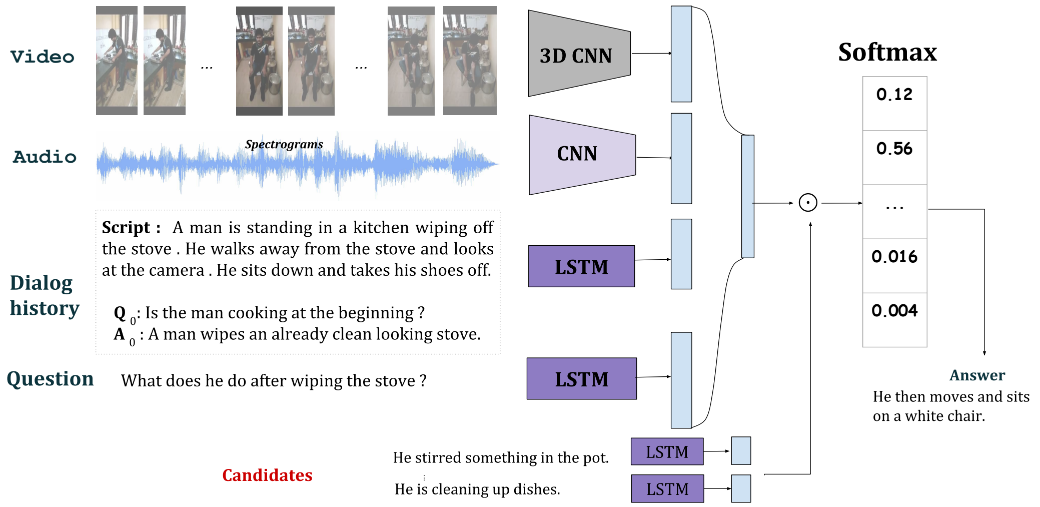

Video
...
...
Audio
Spectrograms
Script : A man is standing in a kitchen wiping off the stove . He walks away from the stove and looks at the camera . He sits down and takes his shoes off.
Dialog history
Q 0: Is the man cooking at the beginning ?
A 0 : A man wipes an already clean looking stove.
Question
What does he do after wiping the stove ?
3D CNN
CNN
LSTM
LSTM
⊙
Softmax
0.12
0.56
...
0.016
0.004
Answer
He then moves and sits on a white chair.
Candidates
He stirred something in the pot.
He is cleaning up dishes.
LSTM
LSTM

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.

general deployment.) More concretely, given a set of 100 potential answers $\{\mathcal{A}_t^{(1)}, \ldots, \mathcal{A}_t^{(100)}\}$, 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 can write the decoder as:

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

where $a_{t,i}$ is the embedding vector for answer candidate $\mathcal{A}_t^{(i)}$, the notation $< \cdot, \cdot >$ represents an inner product, and $s_{t,i}$ is the score computed for the candidate based on its similarity to the input encoding e. 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.

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 [22] 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]. Repository [23] provides code and models fine-tuned on the Charades dataset to extract I3D 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 [29]. The repository [35] 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 11,816 instances split across training (7,985), validation (1,863), and testing (1,968) corresponding to the source Charades video splits. We present results on the test set.

Evaluation Metrics. Although metrics like BLEU [21], METEOR [2], and ROUGE [17] have been widely used to evaluate dialog [19, 28, 33], there has been recent evidence suggesting that they do not correlate well with human judgment [18]. Like [6], we instead 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 (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 (Mean) [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 that 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 four 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, caption, 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, a question and a caption model `Q+C`, as well as a question and dialog history `Q+DH` model that also includes the caption. These models measure regularities between questions, dialogs, 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. While caption encoding has no significant impact on the model performance, adding the dialog history provides the best language-only model performance of 4.72 mean rank.

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, as dialogs are self-referential; in the AVSD dataset, 55.2% of the questions contain co-reference words such as *her*, *they*, and *it*. Such questions strongly depend on the prior rounds of dialog, which are encoded in the DH.

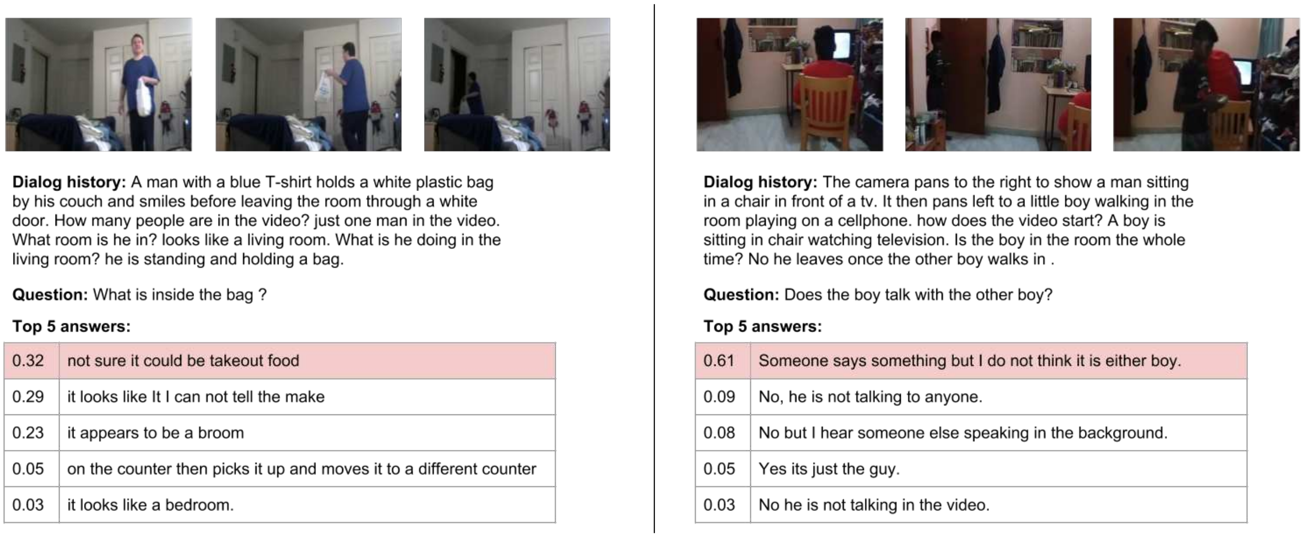
We note that adding video and audio signals improves over dialog history alone, by providing complementary information to ground questions.

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. Thus, it seems that whereas 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 used 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 that model performance improves even more when combined with dialog history and video features (`Q+DH+V+A`) for some metrics, indi-

	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 + C	37.42	20.95	56.17	76.60	7.27
	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
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 receiprocal rank (MRR), recall@k (R@K), and mean rank (Mean). We find that our best performing model leverages the dialog, video, and audio signals 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. The ground truth answer is highlighted in red. In both of these examples, the model ranked the ground truth answer at top position.

	Model	MRR	R@1	R@5	R@10	Mean
Original Setting	Q + DH	50.40	32.76	73.27	88.60	4.72
	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
Shuffled	Q + DH	49.03	31.55	71.28	86.90	5.03
	Q + DH + V	51.47	34.17	74.03	88.40	4.72
	Q + DH + V + A	50.74	33.22	73.20	88.27	4.76

Table 4: Shuffling the order of Questions (Q/A pairs). *Original Settings:* Original results. *Shuffled:* Results on shuffled dialogs.

cating there is still complementary knowledge between the video and audio signals despite their close relationship.

Temporal and Audio-Based Questions. Table 5 shows mean rank on subsets of questions. We filter the questions using the two lists of keywords: audio-related words {talk, hear, sound, audio, music, noise} and temporal 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 5: Mean rank results for the three models Q, Q+A, and Q+V for audio-related questions and temporal 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, answers to the temporal questions were much better using the Q+V model, which confirms our intuition. The Q+A model helps only slightly (7.15 vs 7.31), but the Q+V model yields more significant improvement (5.98 vs 7.31).

The order of the questions/answers is important. An important question to ask is whether the questions and the answers in the dialog are a set of independent question/answer (QA) pairs, or are they strongly co-dependent? To answer this question, we ran an experiment in which we tested the trained model on a shuffled test set containing randomly ordered QA pairs. The top section of Table 4 shows the results on the original test set (ordered), with the results on the shuffled test set below. We observe a difference of $\sim$1.87 for R@k averaged across k and models, and $\sim$0.33 for the mean rank averaged across models, indicating that the order of the QA pairs indeed matters.

Qualitative Examples. Figure 7 shows two examples using the setup Q+DH+V. The first column in the answer table of each example 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 using natural language. We collected the Audio Visual Scene-Aware Dialog (AVSD) Dataset, using a two-person chat protocol on more than 11,000 videos of human actions. We also developed a model and performed many ablation studies, highlighting the quality and complexity of the data. Our results show that the dataset is rich, with all of the different modalities of the data playing a role in tackling this task. We believe our dataset can serve as a useful benchmark for evaluating and promoting progress in audiovisual intelligent agents.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[2] 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. 4, 7

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan 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] Joao Carreira and Andrew 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, 6

[5] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi 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] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv 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] Karan Desai, Abhishek Das, Dhruv Batra, and Devi Parikh. Visual dialog challenge starter code. urlhttps://github.com/batra-mlp-lab/visdial-challenge-starter-pytorch, 2018. 6

[9] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei 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

[10] Unnat Jain, Svetlana Lazebnik, and Alexander 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. 4, 5

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: 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

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li 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

[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 2

[14] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas 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

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 4, 7

[18] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle 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

[19] Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle 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

[20] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron C Courville, and Christopher Joseph Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, pages 7359–7368, 2017. 2

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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

[22] Jeffrey Pennington, Richard Socher, and Christopher 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

[23] AJ Piergiovanni. I3d models trained on kinetics. urlhttps://github.com/piergiaj/pytorch-i3d, 2018. 6

[24] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195. Springer, 2014. 2

[25] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2, 3

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*

preprint arXiv:1409.1556, 2014. 5

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 5

[28] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill 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

[29] Naoya Takahashi, Michael Gygli, and Luc Van Gool. Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, 20(3):513–524, 2018. 5, 6

[30] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 2

[31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4

[32] Licheng Yuu, Eunbyung Park, Alexander C Berg, and Tamara 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

[33] Kangyan Zhou, Shrimai Prabhumoye, and Alan 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

[34] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li 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

[35] znaoya. Aenet: audio feature extraction. url-https://github.com/znaoya/aenet, 2017. 6