Interpretable Multimodal Dialogue System with Natural Language-Based Multimodal Integration

Yoonseok Heo,[1*] Gyunyeop Kim,[2*] Eunseok Yoo,[2*]
Seungsoo Lee,[2] Eunseo Jeong,[2] Sangwoo Kang[2†]
Sogang University,[1] Gachon University[2]
yoonsuk_419@naver.com, gyop0817@gachon.ac.kr, sunnycloud0144@gmail.com,
wlstjddl5916@gachon.ac.kr, esjeong153@naver.com, swkang@gachon.ac.kr

Abstract

A multimodal dialogue system aims to achieve the comprehensive understanding of various types of information in a human-like manner and to interact properly with users; this is one of the ultimate goals of artificial intelligence. A new answer reasoning task was added to the 10th Dialog System Technology Challenge in addition to the existing audio-visual scene-aware dialog task to find the temporal segments of the scene referenced by the system as supporting evidence for answer generation. To this end, we propose an interpretable multimodal dialogue system comprising a multimodal-based answer-generation and a moment localization. We introduce a methodology that transforms a variety of information in a scene detectable by each modality into the form of natural language and leverages a transformer-based language model to integrate them and generate an appropriate answer for a user query. We also introduce a modality-specific moment localization method that can identify video segments semantically similar to a given query and system-generated answer. Our system has exhibited significant performance improvements over the baseline released by the organizers; particularly in the answer reasoning task, we have achieved state-of-the-art performance among the reported results.

Introduction

A multimodal dialogue system(Hori et al., 2018; Liao et al., 2018; Nie et al., 2019) can interact with humans, via a comprehensive understanding based on multi-form information. This is a very challenging task because it requires convergence between multiple fields of artificial intelligence and has therefore been the focus of several studies. As a representative example, video-grounded dialogue generation involves the generation of an answer to a user query about a given video. The premise of this task is the understanding of accurate visual information and of the various methodologies developed to generate accurate system utterances based on multimodal data comprising visual and language information. In particular, recent studies(Huang et al., 2021; Li et al., 2021) have mainly focused on a method of leveraging transformer-based language models to integrate individual information obtained from modality-specific feature extractors. To more closely approximate human perception, the organizers proposed an audio-visual scene-aware task (AVSD)(Hori et al., 2017), which is to generate an appropriate answer to a user's query on a given audio-visual scene containing visual, textual, and even auditory information. (Pasunuru and Bansal, 2019) exploited a dual attention mechanism to fuse information from multiple modalities.

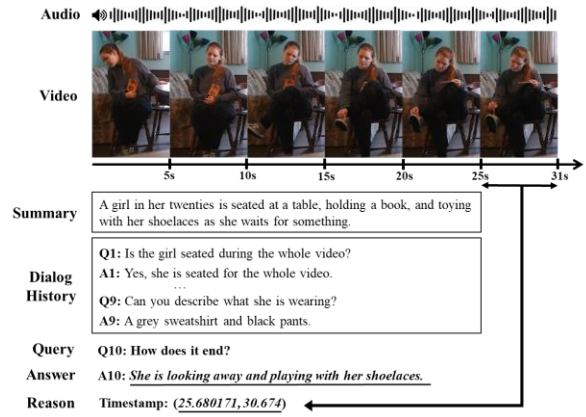

Figure 1. An illustration of the DSTC10-AVSD dataset. The two underlined results are what a system should generate for this task. The former is the system-generated answer(A10) to a user query(Q10) on a given scene and dialog history. The latter is a temporal segment (timestamp) of the scene that the system exploits as the basis for generating the answer(A10).

[*] These authors contributed equally to this work
† Corresponding Author

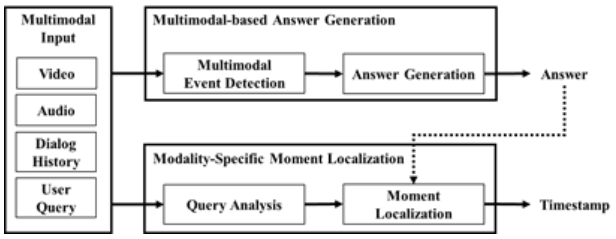

Figure 2. An overview of our system

(Li et al., 2021) proposed a transformer-based multimodal dialogue generation framework, which can integrate all the modality information in a language model.

In this work, we address an extended version of the existing AVSD task organized by the 10th Dialog System Technology Challenge(DSTC10), which includes a new topic on model interpretability. As described in Figure 1, the data consists of an audio-visual scene and a dialog history of nine pairs of questions and answers between the user and the system. In addition, a summary containing sufficient information about the scene is only available during the training phase of the system. This task has two main purposes. The first is video QA dialog (task1) in which the system sufficiently understands the given multimodal information to generate an appropriate answer(A10) to the user query(Q10). The second is the answer reasoning (task2), which aims to find the temporal segment of the given scene that the system used as supporting evidence for generating the answer in task1, which is contributing to improving the interpretability of the system.

In this paper, we propose an interpretable multimodal dialogue system comprising two modules—namely, multimodal-based answer generation and moment localization, as described in Figure 2. Inspired by the fact that a dialogue is mainly about a set of events in the scene, such as entity's actions(e.g., crossing legs, sitting) or audio information(e.g., sneezing, bell ringing), we introduce a methodology that transforms multiple events in a scene that can be detected by individual modalities into the form of natural language and leverages a transformer-based language model to integrate them and generate the appropriate answer to a given query. We also propose a multi-task learning method using the caption generation problem as an auxiliary task to better understand multimodal information and generate a more robust response. Second, we introduce a modality-specific moment localization that can identify the timestamp of a scene semantically similar to a given query and the system-generated answer. To increase the accuracy of localization, we propose a query analysis method that heuristically determines whether to solely focus on visual information or auditory information in a scene. As a result, compared to the baseline released by the organizer, our system achieved 20% and 40% improvements on the BLEU-4(Papineni et al., 2002) and CIDEr(Vedantam et al., 2015) scores in the case of task1,and 40% and 43% improvement on IOU-1 and IOU-2 scores in the case of tasks2. In particular, our system showed state-of-the-art performance among the reported results in DSTC10, in the answer-reasoning task.

Related Work

Most recent studies on multimodal dialogue systems are accompanied by a transformer-based network. (Le et al., 2019) proposes a multimodal transformer network that obtains individual information from feature extractors for each modality and fuses them using a text-based cross-modal attention mechanism. (Li et al., 2021) also proposes a transformer-based generative framework that integrates the whole modalities by encoding features into the system and generates better multimodal-based system responses with multi-task learning methods. (Chu et al., 2020) describes a consecutive multimodal fusion strategy using joint-modal attentions under the conversation progress. Although these approaches

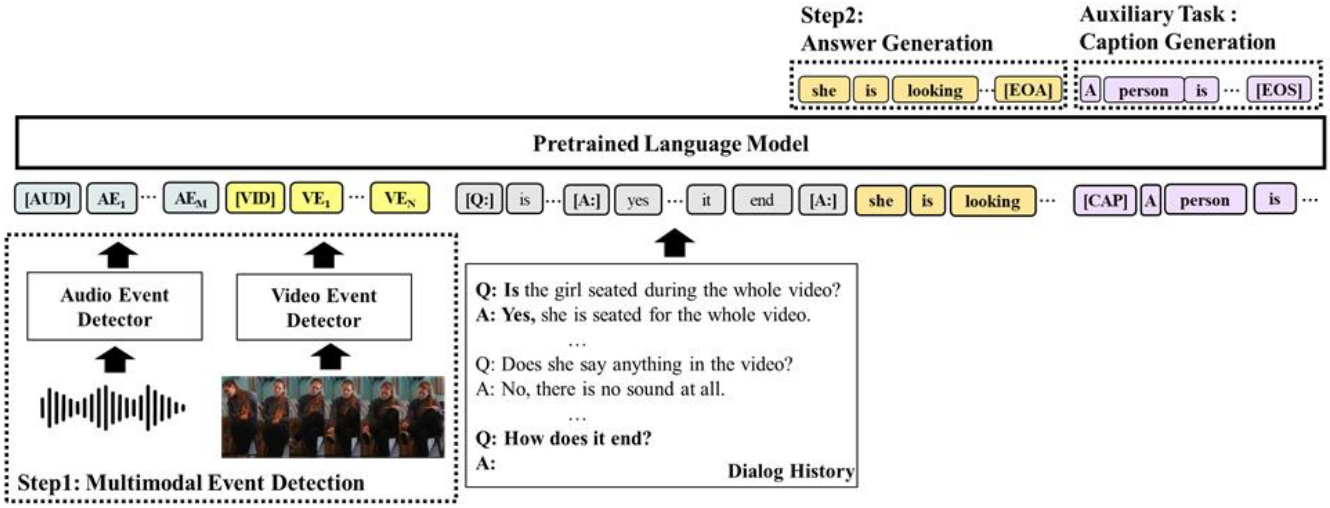

Figure 3. Our proposed method for the multimodal-based answer generation

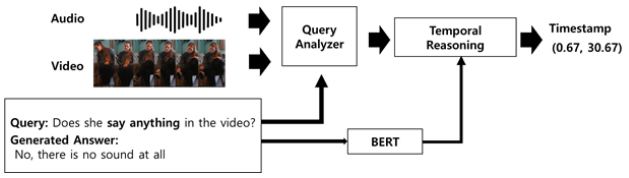

Figure 4. Our proposed Method for the answer reasoning

show meaningful performance, they still have two limitations. One limitation is that the overall system performance is considerably dependent on the usage of summaries during the training phase(Shah et al., 2021). The other is that effective multimodal integration strategies with audio have not been sufficiently proven yet.

Our Methodology

In this section, we introduce an interpretable multimodal dialogue system consisting of two modules such as a multimodal-based answer generation and a modality-specific moment localization. The first module is a multi-modal-based answer generation using a pre-trained language model consisting of two steps such as multimodal event detection and answer generation as shown in Figure 3. The second module is a modality-specific moment localization consisting of query analysis and temporal reasoning as shown in Figure 4. We will describe each module in detail.

Multimodal Event Detection

The scene-aware conversations in this study mainly encompass events that appear in a video or audio modality. Therefore, the understanding of multimodal information is directly related to the understanding of the event shown in the scene. Inspired by this fact, we employ pre-trained event detectors specialized for each modality to extract various events that occur in the video and regard them to be the information from each modality. Figure 3 shows an example of a woman sitting in a chair with a book and playing with her shoelaces. The video does not contain any specific audio information. In this case, the video event detector predicts event categories such as “holding” and “sitting” with high probability. The top N video event categories correspond to information estimated to have appeared in the scene with a high probability. Therefore, we use these categories as the direct information obtained from the visual modality. This approach has an advantage in terms of multimodal understanding in that it uses more explicit natural language information than that in previous studies that used feature embedding for each modality itself. The AVSD data we deal with includes both audio and video. Therefore, in this study, we use a pre-trained transformer-based event classification model that is open to the public to extract event information for each modality.

Audio Event Detector

In this study, we adopt the audio spectrogram transformer (AST)(Gong et al., 2021) model to detect events based on speech. AST is the first transformer-based model proposed for audio event classification problems. This model constructs an encoder model based on self-attention and feed-forward layers. The input speech is converted into a sequence of 128-dimensional log-mel spectrograms, which are used as the model inputs. Each spectrogram is divided into patches of a fixed size, and the model generates encoded results in units of patches as the output. We use the output embedding of “[CLS],” the first input token of the model, as the entire embedding information of the audio spectrogram. This embedding is also used as the input vector of the audio classification layer, which is a downstream task of the model. The AVSD data used in this study do not provide audio classification labels for the audio in the scene. We adopt the open-public AST model, which is a transformer-based encoder fine-tuned with Audio Set(Gemmeke et al., 2017) composed of 527 audio event categories. We do not perform any additional training on this model. In addition, the M audio event category results with high probability values for input speech are considered as events detected from speech. In practice, we adopt 5 audio event categories as a pivot.

Video Event Detector

In this study, we exploit a video swin transformer (VST) (Liu, Ning, et al., 2021), which exhibits high performance in video action recognition tasks, to detect events appearing in the scene. VST is a transformer-based video understanding model, which is a variant of the existing swin transformer (Liu, Lin, et al., 2021) and is suitable for application in the video domain. The Swin transformer uses a dynamic patch generation method to effectively learn the spatial information of an image. The VST samples 32 frames out of all frames of the video and uses them as input. The model comprises several layers of swin transformer blocks. VST uses a large-sized patch when passing through layers, and the self-attention between various patches is performed by changing the position of the window for each layer. In addition, the model is able to reduce the computation cost, and at the same time, sufficiently learn the context of the entire video by performing self-attention only between patches located inside a window of a fixed size in each transformer block. As a result, the output embedding of the [CLS] token can be used as a context for the entire video. Then it is applied as inputs to the linear classification layer for action recognition. In this study, we exploit the smallest VST model fine-tuned on kinetics-400(Kay et al., 2017), a large-scale human action dataset for 400 human action categories. In fact, the AVSD data used in this study contains no action labels from the scene. Therefore, we regard 400 predefined human action categories in kinetics-400 as events that can

occur in the videos. In addition, we regard N action categories with highest probabilities in the action recognition layer of the model for the input video as the detected events in the video. In practice, we adopt 8 action categories as a pivot.

Answer Generation

Each modality has a set of event labels in a natural language in a given audiovisual scene. This enables the integration of multimodal information by encoding the information from each modality directly into the language model. We sequentially encode the M audio event labels, N video event labels obtained earlier, the conversation history and the last user query into the language model. In this paper, we exploit as a language model GPT2(Radford et al., 2019), which shows good performance in various generative tasks. Specifically, the input configuration of the model is as follows:

$$F = (\ [AUD],\ AE,\ [VID],\ VE,\ D\ ) \tag{1}$$

where AE is a sequence of M audio event labels, VD is a sequence of N video event labels, [AUD] and [VID] refer to the special separator tokens for audio and video event labels respectively, and D is a sequence of words in the dialog history. In particular, we add two special separator tokens [Q:] and [A:] to the beginning of every question and answer. The input representation F is now encoded into the model. It can generate an answer(A) in an autoregressive way, conditioned on the encoded information as follows:

$$P(A \mid F;\ \theta) = \prod_{j=1}^{K} P\left(a_j \mid F, a_{<j};\ \theta\right) \tag{2}$$

where A is a sequence of words representing the generated answer and is the trainable parameters.

Moreover, we also propose a multi-task learning method using the caption generation problem as an auxiliary task to better understand multimodal information and generate a more robust response. Similar to the answer generation task, we train the model to generate a caption(C) given a set of audio and video event labels, dialog history, and system-generated answer(A) as follows:

$$P(C \mid A, F;\ \theta) = \prod_{j=1}^{K} P\left(c_j \mid A, F, c_{<j};\ \theta\right) \tag{3}$$

where C is a sequence of words representing the generated caption. In a multitask learning setting, we use the sum of the losses for each task as the overall loss to train our model. We apply a cross-entropy loss function to each task.

Answer Reasoning

A system-generated answer should refer to the scene segments near the occurrence of the event related to the user's query. More specifically, the system identifies a modality from which a hint of the event can be obtained, analyzes the features from the modality, and uses it to generate an answer to the query. Let us take Figure 4 as an example. The user asks whether the woman in the video is talking. In this case, the system needs the voice information of the woman in the video. That is, the system must detect the temporal segment in which the woman is talking from the auditory modality. Motivated by this observation, we propose a modality-specific moment localization method that can identify a temporal segment of a given scene that is semantically similar to a given query and system-generated answer. We introduce a query analysis method that heuristically determines the modality that is presumed to contain the basis of an answer. In particular, we address how to find a query that can only be answered with the auditory modality. In addition, we introduce a moment localization method that can be independently processed according to modality. The following describes each method in detail.

Query Analyzer

In order to enhance the accurate moment localization, we add a query analysis step to heuristically identify a specific modality that is highly likely to contain the rationales for the user query. We have observed that it is sometimes unnecessary for the system to generate an answer using all the information from different modalities. Instead, it is beneficial to use information only from a single modality in some user queries. As shown in Figure 4, if the system focuses only on the information from the auditory modality, it will be able to find the supporting evidence for the answer more accurately. As a result, we address queries that can be answered with only a single modality, and in particular, we heuristically determine keywords that frequently appear in queries that can be answered only using information from the auditory modality. In the actual moment localization phase, we solely use auditory information for the queries containing the aforementioned keywords. Otherwise, the video stream on the scene is used. The effectiveness of the query analyzer can be verified through the ablation experiment in the next section.

Moment Localization

We introduce a modality-specific moment localization network that can identify the temporal moment of the scene that is semantically similar to a given user query and the system-generated answer. To be more specific, we exploit a variant of a 2D temporal adjacent network(2D-TAN) proposed by (Zhang et al., 2020). We independently train two networks according to each of the visual and auditory modalities. An audio-based 2D-TAN is to find a temporal segment on the audio signal that is semantically similar to the given query and answer when it is determined in the query analysis step that the query can be solved only with the audio information.

Otherwise, a video-based 2D TAN is adopted that finds a temporal video moment in the video stream.

The audio-based 2D-TAN consists of three steps: natural language encoding, audio signal encoding, and temporal moment prediction. First, we employ BERT to obtain semantic information of the user query and the system-generated answer. In this work, we concatenate them into one sentence, encode it using BERT, and in particular, the output embedding of the [CLS] token of BERT is used as semantic information for the entire sentence. Then, in terms of audio processing, for each audio signal, we first segment it into 16 non-overlapping clips. The feature representation of each clip can be obtained by average-pooling the audio features for the frames included in the clip, which are extracted from the vggish model provided by the organizer. Then, similar to (Zhang et al., 2020), the audio signal is encoded in the form of a two-dimensional temporal feature map designed to represent key features appearing across a specific time span by max-pooling features for consecutive clips. Now, the auditory and language information can be fused by Hadamard product, and the relevance score between auditory and natural language sentence can be calculated by a temporal adjacent network, which is proposed by (Zhang et al., 2020), with the multiple convolution operations on the fused 2D temporal feature map. Finally, the semantic similarity score between the given query and system-generated utterances and each temporal moment can be obtained in the form of a two-dimensional score matrix. We exploit the moment with the highest value in the score map as the final output.

This process is equally applied to the video-based 2D-TAN. The difference is that video features are used, which are I3D features provided by the organizer. In order to train the two networks independently, in the query analysis step, we split the samples that can be answered only by auditory modality from the training data and use them only to train the audio-based 2D-TAN. All the other samples are used to train the video-based 2D-TAN. The training process is based on the (Zhang et al., 2020), and we train two networks from scratch with the DSTC10 reasoning data provided by the organizer.

Experiment

Experimental Setup

In our experiment, we use Audio-Visual Scene-Aware Dialog (AVSD) dataset provided by the organizers in DSTC10. AVSD dataset consists of 10 consecutive question-answer pairs between two speakers for each video in Charades dataset(Sigurdsson et al., 2016). It also provides a timestamp in video corresponding to each question-answer pair, a visual feature acquired with I3D, and an audio feature acquired with Vggish. 7,659 dialogues are given for training, 1,787 dialogues for validation, and 1,804 dialogues for testing.

For the quantitative evaluation metric, BLEU, METEOR, ROUGE-L, and CIDER, which are commonly used in natural language generation tasks, are used for answer generation (task1). In addition, IoU-1 and IoU-2 score are used as for answer reasoning (task2). Moreover, in natural language generation tasks, human evaluation is used as a significant evaluation metric. For human evaluation, the organizers of the competition form an evaluator. Human evaluation is conducted in a way that the evaluator rates the system from (1-5)points based on the answers generated by each system.

Experimental Results

In this section, we show the experimental results for the two following tasks such as answer generation(task1) and answer reasoning(task2).

Answer Generation(Task1)

We describe the experiment results in three settings: text + visual, text + visual + audio, text + visual + audio + caption, as shown in Table 1. Based on the BLEU-4 value, the performance in text + visual + audio + caption setting was the highest.

We first conducted the experiment with **only text (question-answer pair) and visual information**. As can be seen in Table 1, our model has shown higher performance in all metrics compared to the baseline model. It has shown an improvement of 0.0477 in BLEU-4, 0.2234 in CIDEr. Next, in a **Text+Visual+Audio** setting, the model uses audio information as well as visual information in the video to generate an answer. Our model in this setting also has shown higher performance in all metrics than the baseline model. It shows an improvement of 0.0515 in BLEU-4 and 0.2382 in CIDER. Compared to text + visual task without audio information, it has shown an improvement of 0.0038 in BLEU-4 and a decrease of 0.0148 in CIDER. Lastly, in **Text+Visual+Audio+Caption** setting, we have observed the effectiveness of the multi-task learning method. Specifically, our model has shown performance improvement by 0.0517 in BLEU-4 and 0.2211 in CIDER compared to the baseline model. The CIDEr value has been slightly lower than the text + visual + audio task without captions, but the BLEU-4 value has been higher.

Answer Reasoning(Task2)

We report the experimental results in four settings for the reasoning task as shown in Table 2. In both IoU-1 and IoU-2, performance in text + visual + audio + caption setting has shown the highest, and in particular, we have achieved the state-of-the-art performance among the reported results.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Human Rating
baseline	0.5716	0.4223	0.3196	0.2469	0.1909	0.4386	0.5657	2.851
Our model								
+T+V	0.6509	**0.4897**	0.3764	0.2946	**0.2274**	**0.5022**	0.7891	-
+T+V+A	0.6406	0.4885	0.3786	0.2984	0.2251	0.5016	**0.8039**	-
+T+V+A+C	**0.6455**	0.4889	**0.3796**	**0.2986**	0.2253	0.4991	0.7868	3.3

Table 1. Experimental results for answer generation task on the test set provided by the organizers in DSTC10-AVSD challenge(T: text; V: visual; A: audio; C: caption)

Models	IoU-1	IoU-2
baseline	0.3614	0.3798
T+V+A+C	**0.5157**	**0.5443**
-C	0.5061	0.5338
-(C+A)	0.5048	0.5329
-Query Analyzer	0.5023	0.5304

Table 2. Experimental results for answer reasoning task on the test set provided by the organizers in DSTC10-AVSD challenge(T: text; V: visual; A: audio; C: caption)

We first conducted the experiment of the reasoning task using the answer generated by the **T+V+A+C** answer generation model. As shown in Table 2, compared to the baseline model, our model has shown better scores with a large margin of 0.1543 and 0.1645 at IoU-1 and IoU-2 both, respectively. This is the highest IoU-1 and IoU-2 results in all submissions of DSTC10. In addition, there is a significant difference of 0.0312, 0.0341, respectively, with the IoU-1,2 score of the highest model among submissions excluding our proposed model. In addition, we conducted an experiment on what results would be shown if a caption was excluded from the above T+V+A+C setting. As a result, IoU-1, 2 scores were higher than the baseline model, with a difference of 01447 and 0.1540, respectively. Compared to tasks containing captions, IoU-1, 2 scores decrease 0.0096, 0.0105, respectively, when captions were not included. We also conducted an experiment on settings where both caption and audio information were excluded. Compared with the baseline model, IoU-1, 2 scores were higher at 0.1434 and 0.1531, respectively. Compared to tasks that included audio information and did not include caption, IoU-1, 2 scores fell 0.0013, 0.0009, respectively, when both audio information and caption were excluded.

We conducted an ablation experiment to analyze the effect of Query Analyzer. Answer generation task were performed based on the T+V+A+C setting. In the Reasoning task, the Query Analyzer was removed from the existing setting and used. In other words, in this setting, the reasoning model was trained and evaluated using only visual information without audio information. As a result, it was found that the IoU-1,2 scores decreased by 0.0134 and 0.0139, respectively, compared to the task of determining whether the query requires visual information or audio information at the reasoning task.

Conclusions and Future Work

In this paper, we introduce an interpretable multimodal dialogue system consisting of a multimodal-based answer generation and a modality-specific moment localization. Our proposed method has shown considerable results on the DSTC10-AVSD challenge, and in particular, we have achieved the state-of-the-art performance in the answer reasoning task. In the future, we plan to expand the study of multimodal dialogue systems with symbolic knowledge graphs such as common sense.

Acknowledgement

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT under Grant NRF-2021R1F1A1061558.

References

Chu, Y.-W.; Lin, K.-Y.; Hsu, C.-C. and Ku, L.-W. 2020. Multi-step Joint-Modality Attention Network for Scene-Aware Dialogue System. *CoRR* abs/2001.0, https://arxiv.org/abs/2001.06206.

Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M. and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Gong, Y.; Chung, Y.-A. and Glass, J. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pp. 571–575, https://doi.org/10.21437/Interspeech.2021-698.

Hori, C.; Alamri, H.; Wang, J.; Winchern, G.; Hori, T.; Cherian, A.; Marks, T.K.; Cartillier, V.; Lopes, R.G.; Das, A. and others. 2018. End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features. *ArXiv preprint arXiv:1806.08409*.

Hori, C.; Marks, T.K.; Parikh, D. and Batra, D. 2017. Video Scene-Aware Dialog Track in DSTC7. In *Video Scene-Aware Dialog Track in DSTC7.*, http://workshop.colips.org/dstc7/proposals/DSTC7_Scene_Aware_Dialog.pdf.

Huang, Y.; Xue, H.; Liu, B. and Lu, Y. 2021. Unifying Multimodal Transformer for Bi-Directional Image and Text Generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1138–1147. New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/3474085.3481540.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M. and Zisserman, A. 2017. The Kinetics Human Action Video Dataset2017.

Le, H.; Sahoo, D.; Chen, N. and Hoi, S. 2019. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5612–5623. Florence, Italy: Association for Computational Linguistics, https://doi.org/10.18653/v1/P19-1564.

Li, Z.; Li, Z.; Zhang, J.; Feng, Y.; Niu, C. and Zhou, J. 2021. Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 2476–2483.

Liao, L.; Ma, Y.; He, X.; Hong, R. and Chua, T.-S. 2018. Knowledge-Aware Multimodal Dialogue Systems. In *Proceedings of the 26th ACM International Conference on MultimediaMM '18*, pp. 801–809. New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/3240508.3240605.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S. and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S. and Hu, H. 2021. Video Swin Transformer2021.

Nie, L.; Wang, W.; Hong, R.; Wang, M. and Tian, Q. 2019. Multimodal Dialog System: Generating Responses via Adaptive Decoders. In *Proceedings of the 27th ACM International Conference on MultimediaMM '19*, pp. 1098–1106. New York, NY, USA: Association for Computing Machinery, https://doi.org/10.1145/3343031.3350923.

Papineni, K.; Roukos, S.; Ward, T. and Zhu, W.W. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pp. 311–318, https://doi.org/10.3115/1073083.1073135.

Pasunuru, R. and Bansal, M. 2019. DSTC 7-AVSD : Scene-Aware Video-Dialogue Systems with Dual Attention. In *DSTC7 at AAAI2019 workshop*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D. and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Shah, A.P.; Geng, S.; Gao, P.; Cherian, A.; Hori, T.; Marks, T.K.; Roux, J. Le and Hori, C. 2021. Audio-Visual Scene-Aware Dialog and Reasoning using Audio-Visual Transformers with Joint Student-Teacher Learning. *ArXiv: 2110.06894*.

Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I. and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Leibe, B.; Matas, J.; Sebe, N. and Welling, M. (Eds), *Computer Vision - {ECCV} 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part {I}Lecture Notes in Computer Science*, pp. 510–526. Springer, https://doi.org/10.1007/978-3-319-46448-0_31.

Vedantam, R.; Zitnick, C. and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, https://doi.org/10.1109/CVPR.2015.7299087.

Zhang, S.; Peng, H.; Fu, J. and Luo, J. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(07): 12870–12877.