QUALIFIER: Question-Guided Self-Attentive Multimodal Fusion Network for Audio Visual Scene-Aware Dialog

Muchao Ye
The Pennsylvania State University
University Park, PA
muchao@psu.edu

Quanzeng You
Microsoft Azure Computer Vision
Bellevue, WA
quyou@microsoft.com

Fenglong Ma*
The Pennsylvania State University
University Park, PA
fenglong@psu.edu

Abstract

Audio video scene-aware dialog (AVSD) is a new but more challenging visual question answering (VQA) task because of the higher complexity of feature extraction and fusion brought by the additional modalities. Although recent methods have achieved early success in improving feature extraction technique for AVSD, the technique of feature fusion still needs further investigation. In this paper, inspired by the success of self-attention mechanism and the importance of understanding questions for VQA answering, we propose a question-guided self-attentive multi-modal fusion network (QUALIFIER) to improve the AVSD practice in the stage of feature fusion and answer generation. Specifically, after extracting features and learning a comprehensive feature for each modality, we first use the designed self-attentive multi-modal fusion (SMF) module to aggregate each feature with the correlated information learned from others. Later, by prioritizing the question feature, we concatenate it with each fused feature to guide the generation of a natural language response to the question. As for experimental results, QUALIFIER shows better performance than other baseline methods in the large-scale AVSD dataset named DSTC7. Additionally, the human evaluation and ablation study results also demonstrate the effectiveness of our network architecture.

1. Introduction

The development of intelligent agents has attracted increasing interests in the academia and the industry recently.

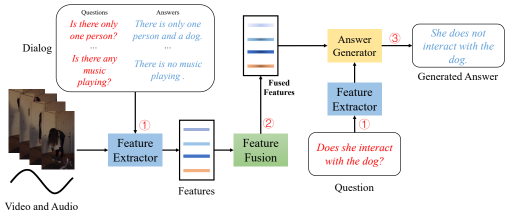

Dialog
Questions
Answers
Is there only one person?
...
Is there any music playing?
There is only one person and a dog.
...
There is no music playing .
Video and Audio
Feature Extractor
①
Features
Feature Fusion
②
Fused Features
Answer Generator
③
She does not interact with the dog.
Generated Answer
Feature Extractor
①
Does she interact with the dog?
Question

Figure 1. Framework of the proposed method. It has three modules for AVSD including Feature Extractor, Feature Fusion and Answer Generator. The details of each module are designed with two principles: prioritizing the role of question feature and fusing content features for better fused feature.

The key to realizing intelligent agents lies in empowering machines to observe different scenes, understand human dialogs and generate reasonable responses. Such a perspective has shifted the focus of computer vision community from traditional visual question answering (VQA) [4, 5, 8, 18] tasks to more advanced ones, such as the audio visual scene-aware dialog (AVSD) [3, 16] task, which requires the designed agent to generate a natural language response to a question after observing a short video and listening to the corresponding audio with the prior knowledge provided by data collectors from their dialog history. With such abilities, compared to previous VQA tasks in images, the designed agents in AVSD are more close to the ideal intelligent agents that humans dream about. However, in the AVSD task it is challenging to design better mechanisms in understanding the temporal developments of scenes and conversions as well as fusing inputs of different modalities like audio and video simultaneously.

To solve this challenge, existing methods [16, 20, 36]

* indicates corresponding author.

first concatenate features of different modalities together and then generate a natural language response. When extracting the features of each modality, they also use attention mechanisms [32]. However, they ignore the significance of the question input in this task: in the practice of incorporating different features together, they treat the question feature as the same as other features and fuse them together to get a new one. Nonetheless, since the aim of the task is generating a reasonable response to the asked question, the fusion of features and answer generation should center on the question. With the observations mentioned before, we need to design a new multi-modal fusion mechanism and prioritize the role of VQA question in the AVSD task.

In this paper, we propose an effective solution by designing a question-guided self-attentive multi-modal fusion network (QUALIFIER for short) for the AVSD task, with an emphasis on the role of question features in answer generation after the self-attentive multi-modal fusion. In particular, we divide the processing of multi-modal inputs into three stages, as shown in Figure 1. Firstly, we handle the temporal characteristics of video, audio and dialog separately in order to get a comprehensive representation for each modality by the feature extractor. After that, we use the designed self-attentive [30] multi-modal fusion (SMF) module to fuse the comprehensive representations of video, audio and dialog. Finally, given the mutually fused features, we can generate natural language responses in the answer generator module with the guidance of the question feature.

More specifically, the improvement of the designed QUALIFIER model for the AVSD task comes from the hierarchies of dealing with multi-modal data. In order to generate a more reasonable response, the designed agent should have a good understanding both on what happens in observed scenes and ongoing conversations, which are the "contents", and the question that it is required to answer. If the question is not handled separately from content features, it is likely that the question will be disregarded somehow because it is not synchronous with the contents when models mainly focus on contents in fusion. Thus, by fusing the content features first and separating the question feature later, the proposed method reshapes the role of question feature as the guide of response generation and retains the details of contents at the same time.

To summarize, our contributions are listed as follows:

- We propose a question-guided mechanism for generating a natural language response to questions in the AVSD task. By prioritizing the role of questions, the proposed method avoids the conflicts of finding out important details in content features and generating question-related responses, which is helpful for the improvement of the AVSD task.

- In order to learn good representations for multi-modal features, we design a self-attentive multi-modal fusion module for the feature fusion. It ensures finding highly related features between representations of different modalities and aggregating the salient ones. In this way, we can attain more semantically correlated fused features.

- We conduct our experiments on a widely-used AVSD dataset, *i.e.* the challenging DSTC7 dataset. The comparison results and ablation studies both show the effectiveness of the proposed method.

2. Related Work

2.1. VQA and AVSD

VQA is a task of generating a natural language response to a question based on the given visual input, which includes images or videos. The early works are grounded in image question answering. The common frameworks they use include image encoders, question encoders, multi-modal fusion modules and answer generators. With the advancement of deep learning, they mostly use convolutional neural networks (CNNs) such as faster R-CNN [26] for image feature extractions and recurrent neural networks (RNNs) like long short-term memory [15] for question feature extraction and answer generation. As for multi-modal fusion module, simple operations like concatenations, summations and multiplication are preferable. When it comes to video question answering, more complicated CNNs like C3D [29] and I3D [9] are needed to process the video temporality.

We refer to image question answering [4, 5, 8, 18] and video question answering [12, 21, 35] tasks as traditional VQAs. As for AVSD, it is a newer form of VQA proposed by Alamri *et al.* [3]. Similar to VQA agents, an AVSD agent normally includes content and question encoders, multi-modal fusion module and response generator. The difference is that not only does AVSD need to handle an additional audio modality, but it also requires the understanding of dialog information. Let us take the used dataset DSTC7 as an example to illustrate the challenges, which uses the videos of Charades as benchmark videos with a dialog history collected by the collectors themselves. Compared to existing image question answering datasets such as VQA [5] and VQA-CP [2], the objects occur in DSTC7 are not static but dynamical, and DSTC7 has a multi-turn dialog, so it is much harder to be dealt with than existing image question answering datasets. And compared to the setting of multiple choice answering in video question answering datasets like MovieQA [28] and TVQA [21], DSTC7 requires the agents to generate answers to questions themselves, which is more realistic but far more challenging.

2.2. Deep Audio-Visual Learning

Tasks related to audio-visual learning have received great attention from the deep learning community because of the superior performance that DNNs have in both audio and visual modalities, and the multi-modal fusion mechanism is the keys to audio-visual learning. For example, [23] uses early-fusion and self-supervision to improve the audio-visual feature learning, and [1] uses attention mechanisms to aggregate these two types of features. The coherence of audio-visual signals is also important for the tasks such as Audio-Visual Source Separation and Localization, and [33, 34, 11] provide successful solutions for these problems by introducing new fusion mechanisms, e.g., the self-attention fusion for audio and video input in [11]. AVSD is also a task relevant to deep audio-visual learning. Existing work such as [16, 20, 36] also adopts the attention mechanism for feature fusion, but they may ignore the importance of question input. As for the proposed work, it fuses features with self-attention and prioritizes the question input and uses it to guide the answer generation.

3. Method

In this section, we would like to introduce the notations of the task and then detail the design of QUALIFIER.

3.1. Basic Notations and Problem Setting

In AVSD task, the inputs that we are given include video V, audio A, dialog history H and question q_t, and the designed agents are required to generate a response to the question based on the inputs. Specifically, the input video V is divided into m segments, which is represented by $V = [v_1, ..., v_m]$, where v_i is the feature of the i-th segment. Similarly, the audio A is divided into n segments, and A is represented by $[a_1, ..., a_n]$, where a_j is the feature of the j-th segment of the audio. (Note that m may not equal to n.) We are also given a dialog H of data collects with $t-1$ turns of question and answer pair, which is $H = [(q_1, a_1), ..., (q_{t-1}, a_{t-1})]$, where q_k and a_k is the question and answer in the k-th turn, respectively. Given a question q_t in the t-th turn, we need to generate an answer $\hat{a}_t$ by,

$$\hat{a}_t = G(V, A, H, q_t), \tag{1}$$

where G is the designed agent. The goal of this task is generating an $\hat{a}_t$ that is similar to the ground truth a_t.

The generator G used in this paper is the proposed QUALIFIER. Given the training data, it learns to represent the temporal dynamics for each modality, fuse them together with SMF, and generate answer with a question guide. In the following paragraphs, we would like to discuss the setting of AVSD, and then introduce how QUALIFIER processes the multi-modal information by the following feature extractors, self-attentive multi-modal feature fusion (SMF) module and question-guided answer generator.

3.2. Feature Extractors

The first challenge for the AVSD task is how to exploit the temporal structure of each modality to learn representative features. In the proposed QUALIFIER model, we adopt attention mechanism for video V and audio A while using long short-term memory (LSTM) [15] network to encode the textual inputs including dialog history H and question q_t.

Video. Recent years have witnessed the improvement in video feature extractions. For the video segments $[v_1, ..., v_m]$, we would like to use the state-of-the-art video recognition architecture I3D [9] network to extract the feature of each segment. Given the video input V, for frame segments $[v_1, ..., v_m]$, we have features

$$f_1^v, ..., f_m^v = \text{I3D}(v_1, ..., v_m), \tag{2}$$

where $f_i^v \in \mathbb{R}^{d_v}$ is the feature extracted from segment v_i.

To prepare for the fusion of all inputs later, we need to learn comprehensive features for all modalities. Inspired by the recent success that attention mechanism [30, 32] has achieved in temporal-data-related tasks, we use the following attention mechanism to aggregate temporal feature $[f_1^v, ..., f_m^v]$. First, we adopt a 2-layer feedforward network (FFN) to learn an attention score from each feature f_i^v,

$$\alpha_i = \text{FFN}(f_i^v). \tag{3}$$

After obtaining α_i for each feature f_i^v, we need to normalize the attention scores to attention weights in the range of $[0, 1]$ by softmax function, which is

$$a_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{m} \exp(\alpha_j)}. \tag{4}$$

With the attention weights that we have, we finally get the representation for the whole video V by the weighted sum

$$f_V = \sum_{i=1}^{m} a_i \cdot f_i^v. \tag{5}$$

Recalling that in video classification task there is a common practical skill using both video frames and optical flows [10] to obtain a better motion representation of video, we also utilize the optical flow features $[o_1, ..., o_m]$ extracted from $[v_1, ..., v_m]$ for better feature learning. Likewise, given the optical flow features, we use I3D network as the feature extractor,

$$f_1^o, .., f_m^o = \text{I3D}(o_1, .., o_m), \tag{6}$$

where $f_i^o \in \mathbb{R}^{d_o}$ is the feature extracted from optical flow segment o_i.

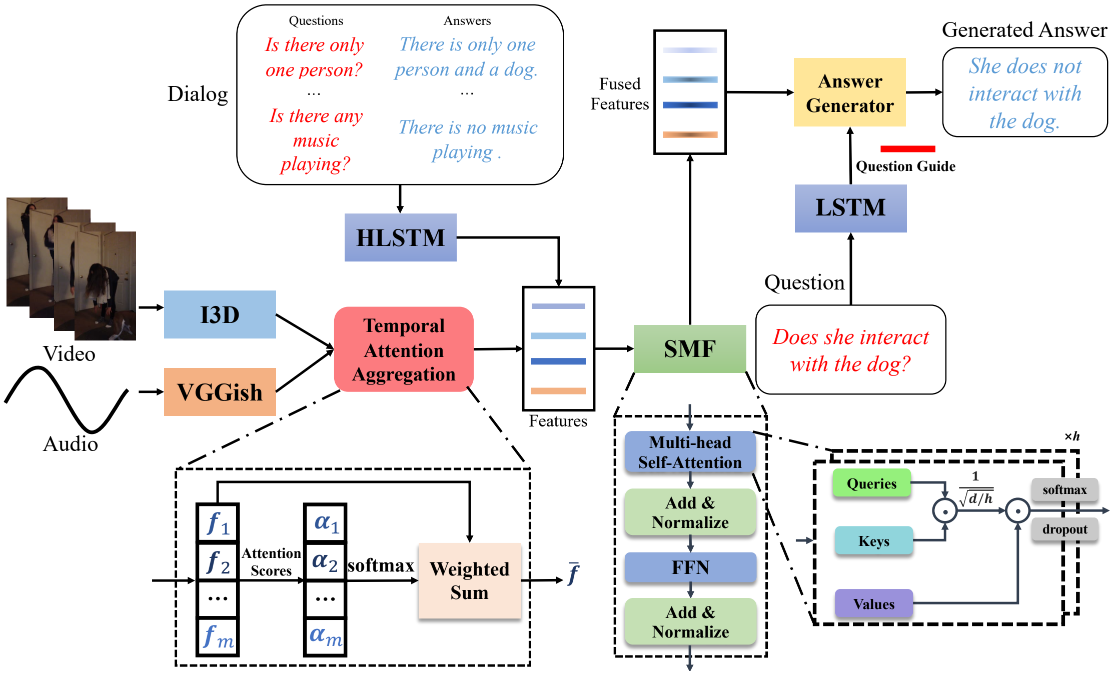


Figure 2. Framework of the proposed method. We first extract the features from all modalities and aggregate them by either temporal attention or LSTM to get the comprehensive features. Then the content features are fused by the SMF module to learn the correlation between one another. Finally, the question feature feature is used as the guide for generating answer.

Similarly, we can get a comprehensive representation f_O for optical flow when the same process from 3 to 5 is conducted in optical flow inputs $[f^o_1, .., f^o_m]$.

Audio. Audio is in the form of 2-dimensional temporal signal. With the n audio segments $[a_1, ..., a_n]$, their features are extracted by the VGGish [14] network, *i.e.*

$$f^a_1, ..., f^a_n = \mathrm{VGGish}(a_1, ..., a_n), \tag{7}$$

where $f^a_i \in \mathbb{R}^{d_a}$ is the feature extracted from audio segment a_i by VGGish.

Similar to Equations from 3 to 5, we can also get the comprehensive representation f_A for audio inputs by another weighted sum $f_A = \sum_{i=1}^{n} b_i \cdot f^a_i$, where b_i is the normalized attention weight for the i-th attention score β_i learned from f^a_i.

Dialog History and Question. During the collection of data, one of the data collectors sees the whole video and then use their understanding about the video to answer the questions raised by another data collector who has only watched the first, middle and final frames of the video. For this dialog history input H of data collectors, we use hierarchical LSTMs (HLSTM) to extract the features. For i-th turn of dialog in H, which has both question q_i and answer a_i, we concatenate them together and put the pair (q_i, a_i) into the bottom LSTM for feature learning, and then the last hidden states of all question-answer pairs are used as the input of the top LSTM in HLSTM for final feature learning. Thus, for dialog history H, we have features

$$f_H = \mathrm{HLSTM}((q_1, a_1), ..., (q_{t-1}, a_{t-1})), \tag{8}$$

where $f_H \in \mathbb{R}^{d_h}$ is the feature of the whole dialog derived from the last hidden state output by the top LSTM in the hierarchical LSTMs.

As for the question q_t, it is processed by a 2-layer LSTM, *i.e.* $\mathrm{LSTM_q}$, and we get its feature f_q by

$$f_q = \mathrm{LSTM_q}(q_t), \tag{9}$$

where $f_q \in \mathbb{R}^{d_h}$ is the last hidden state output by LSTM.

3.3. Self-attentive Multi-modal Feature Fusion

After obtaining the features of content modalities, the next step is to fuse them together so as to use the complementary information between different modalities to get better representations. This step is necessary because the related information between the different modalities are helpful for generation: by fusing different modalities together, the designed agent can learn to find out the salient information that appears in different modalities and further strengthen their representations. Previous works [17] have explored the fusion mechanisms such as early fusion, late fusion and slow fusion. In this work, we utilize the self-attention [30] mechanism to fuse the comprehensive multimodal inputs. Notably, we do not fuse question feature f_q

in our SMF module because we would like to prioritize f_q and use it to guide the answer generation process. Thus, the inputs of SMF include comprehensive features of video segment f_V, optical flow f_O, audio f_A, and dialog history f_H.

In the SMF module, we design a one-layer multi-headed self-attention mechanism, which has η head. For the j-th head, the multi-modal features $\{f_V, f_O, f_A, f_H\}$ are all mapped into a common space $\mathbb{R}^d$ and packed into a matrix $M \in \mathbb{R}^{d \times 4}$. Then we multiply M by projection matrices $W_j^Q \in \mathbb{R}^{d/\eta \times d}$, $W_j^K \in \mathbb{R}^{d/\eta \times d}$ and $W_j^V \in \mathbb{R}^{d/\eta \times d}$ to generate queries $Q_j \in \mathbb{R}^{d/\eta \times 4}$, keys $K_j \in \mathbb{R}^{d/\eta \times 4}$ and values vectors $V_j \in \mathbb{R}^{d/\eta \times 4}$ as follows,

$$Q_j = W_j^Q M, K_j = W_j^K M, V_j = W_j^V M. \qquad (10)$$

By the utilization of self-attention mechanism, we aggregate each comprehensive feature with the correlated content information learned from other modalities,

$$\mathrm{head}_j = \mathrm{softmax}(\frac{Q_j K_j^{\mathrm{T}}}{\sqrt{d/\eta}} V_j), \qquad (11)$$

where $\mathrm{head}_j \in \mathbb{R}^{d/\eta \times 4}$ is the updated features with correlation information.

Later, we concatenate the η heads together and project the concatenated feature back to the original feature space,

$$O = W_o \cdot (\mathrm{head}_1 || ... || \mathrm{head}_\eta), \qquad (12)$$

where $||$ is concatenation operation and $W_o \in \mathbb{R}^{d \times d}$ is the transformation matrix. Next, we put O into a dropout [27] layer, a layer normalization [6] layer and add the output to the original input with a residual connection [13],

$$M_r = M + \mathrm{dropout}(\mathrm{norm}(O)), \qquad (13)$$

where $\mathrm{dropout}(\cdot)$ is the dropout layer and $\mathrm{norm}(\cdot)$ is the normalization layer. Finally, we put M_r into a 2-layer feed-forward network to get the final output:

$$M_\omega = W_2(\mathrm{ReLU}(W_1 M_r + b_1)) + b_2, \qquad (14)$$

where W_1, W_2, b_1, b_2 are the parameters of the feedforward network. The output $M_\omega = [f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H]$ are the fused features aggregated with the correlated information between different modalities.

3.4. Question-Guided Answer Generator

If the designed agent wants to generate a natural response to the question, not only does it need to understand the scenes from different modalities, but it is also required to well apprehend the question and correctly use the information it gets from different modalities to answer the question. So we argue that the question feature f_q should be separated from the content features and used as a guide of generating the answer to the question. Thus, in the SMF module, we do not include f_q as input because we think that question feature should be the guide of answer generation instead of multi-modal feature fusion. Now with the output $[f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H]$ from the SMF module, we describe the process of using the question feature f_q to guide the generation of answer as follows.

Since q_t is the key to question answering, we should retain the details of question feature f_q for generation. Therefore, we concatenate f_q with $f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H$, and we get a question guided fused feature,

$$c_\omega^q = f_\omega^V || f_\omega^O || f_\omega^A || f_\omega^H || f_q, \qquad (15)$$

where c_ω^q is referred as to the context vector for answer generation.

With the question-guided feature c_ω^q, we now use it to generate the answer $\hat{a}_t$ to the question q_t. In this step, we use a 2-layer LSTM ($\mathrm{LSTM}_{\mathrm{de}}$) to generate the answer. Specifically, given the previous generated word y_{i-1} and hidden state h_{i-1}, we generate the hidden state of the i-th word in the answer by

$$h_i = \mathrm{LSTM}_{\mathrm{de}}(w_{i-1} || c_\omega^q, h_{i-1}), \qquad (16)$$

where $w_{i-1} \in \mathbb{R}^{d_e}$ is the embedding of the word y_{i-1}. A $\mathrm{softmax}$ function is then used to output the probability distribution of the next word y_i over the words in the dictionary,

$$P(y_i | w_1, ..., w_{i-1}) = \mathrm{softmax}(W_{\mathrm{de}} h_i + b_{\mathrm{de}}), \qquad (17)$$

where W_{de} and b_{de} are parameters of the mapping function. The initial input y_0 is set to $\langle \mathrm{SOS} \rangle$, which is a special character denoting the start of the sentence. Suppose the answer generation stops in step s, we can have our answer organized in $\hat{a}_t = [y_1, ..., y_s]$, where y_i is the word with highest probability in step i.

4. Experiments

In this section, we will first discuss the dataset and experimental settings. Then we will show the experimental results of the proposed model. Finally, we will give some human evaluation and analysis results of our method.

	Training	Validation	Test
Number of Videos	7,659	1,787	1,710
Number of Turns	153,180	35,740	13,490
Number of Words	1,450,754	339,006	110,252

Table 1. Statistics of DSTC7.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline [16]	0.256	0.161	0.109	0.078	0.113	0.277	0.727
GuidedLDA-all [20]	0.272	0.173	0.118	0.085	**0.119**	0.293	**0.793**
LSTM+AF [36]	**0.276**	0.176	0.119	0.084	0.117	0.293	0.766
QUALIFIER	**0.276**	**0.177**	**0.121**	**0.086**	**0.119**	**0.294**	0.789

Table 2. Comparison results on DSTC7.

4.1. Experimental Setup

Our experiment is conducted on the DSTC7 dataset, which is the widely used large-scale dataset for the task of AVSD. The video contents in DSTC7 are provided by the Charades dataset, and the dialog contents in DSTC7 are collected from a questioner and an answerer during the data collection process. The statistics on the DSTC7 dataset are shown in Table 1.

Baselines. The baselines we use for comparison are as follows.

- **Baseline [16].** The Baseline method for DSTC7 is introduced in [16] when DSTC7 is released. After extracting the features of video, audio, dialog history and question, the baseline method concatenates them together as the context feature to generate the natural language response.

- **GuidedLDA-all [20].** GuidedLDA-all explores how the topic information of different scenes and the attention mechanism can improve the performance of answer generation. It uses a GuidedLDA to model the topic information and adds an attention module in the decoder to generate better response.

- **LSTM+AF [36].** The main technique that LSTM+AF uses is employing a LSTM in the multi-modal encoder module and the attention mechanism for multi-modal fusion.

Implementation Details. Our method is implemented in PyTorch [25] framework in an NVIDIA Tesla P100 GPU and Intel Xeon E5-2680 CPUs. The number of training epochs is 15, and the mini-batch size is 64. The parameters are trained by Adam [19]. The I3D feature sizes d_v and d_o for video frames and optical flows are both 512, and the VGGish feature size d_a for audio is 64. The hidden state sizes d_h of history dialog and question embeddings are both 128, and the embedding size d_e of each word in sentence generation is 100. For SMF, the output feature sizes of the keys, values and queries all are 128, the number of self-attention heads is 8, the dropout rate is 0.1, and we use layer normalization [6] as the normalization layer.

Evaluation Metrics. The results of sentence generation are evaluated by four types of metrics: BLEU [24], METEOR [7], ROUGE-L [22] and CIDEr [31]. BLEU is related to the precision and recall, and it is calculated by a modified n-gram precision and best match length, respectively. METEOR is a metric similar to BLEU, but it improves BLEU using a weighted F-score to calculate precision and recall and a penalty function to correct the scores with human judgement. ROUGE-L uses the longest common subsequence to calculate the precision and recall, and its final score is an F-score based on the calculated precision and recall. CIDEr also accounts both recall and precision using the average of scores from the varying length n-grams, which is computed by the average cosine similarity between the candidate sentence and the reference sentences. The higher these scores are, the better performance the designed agent has.

4.2. Performance Evaluation

In this part, we first compare the performance of our model with baseline methods, and then we use an ablation study to evaluate the reasonableness of the model design.

Comparison with baselines. The comparison results with the baselines are illustrated in Table 2. Compared with other methods, our method has the best performance in BLEU, METEOR and ROUGE-L, and it has the second best performance in CIDEr with a minor gap between the first one. These methods only use attention mechanisms in the stage of feature extraction, but in the stage of feature fusion and answer generation, they basically treat the question feature and content features equally and use mere concatenation for fusion when they generate answers. One of the important differences between these methods and ours is that our method emphasizes the role that the question feature plays in generating the answer and uses it to guide the answer generation. Another difference is that we use a modified multi-headed self-attention to fuse the features. From these results, we can conclude that our method has a more reasonable fusion mechanism for multi-modal inputs.

Human evaluation and analysis. Although the improvements that our model has are not huge on the quantitative metrics, we need to note that they may not well reflect whether generated answers are natural to human beings. Meanwhile, we also find that a manual evaluation on the generated answers is necessary because the DSTC7 dataset contains quite a few annotation noise, some of which are

ID	Description
41	*Question*: is it a male or female ? *Reference Answer*: hi this is an easy task send me 10 ques and we are done in about 5 mins ... the person is male *Analysis*: The referenced answer mistakenly includes some annotation instructions.
99	*Question*: - 1 *Reference Answer*: yes , the woman walks on and the man keeps drinking . *Analysis*: The question is missing in this test sample.
118	*Question*: what color is the bag ? *Reference Answer*: the bag is a clear plastic bag . *Analysis*: The referenced answer does not provide the information of color, which is not what the question want.

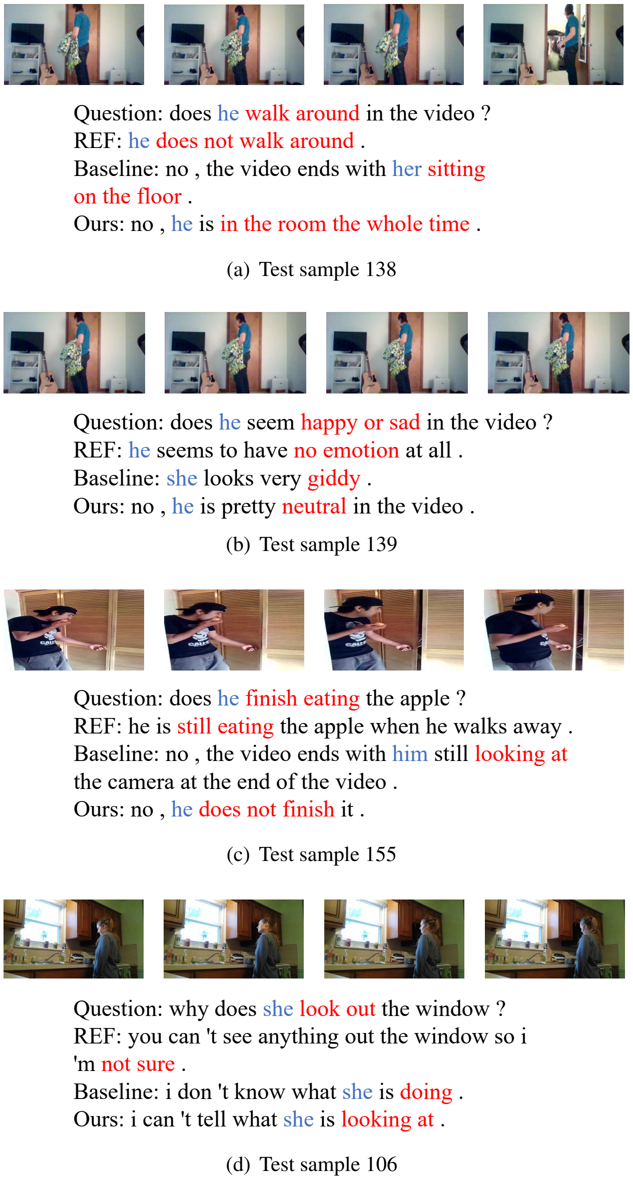
Table 3. Annotation examples in test set and analysis.

listed in Table 3. Considering that the used metrics are not able to show the naturalness of answers and the test set has some noisy reference questions and answers, it is necessary to manually evaluate the quality of the generated sentences.

Thus, we need to manually compare the quality of generated answers, and we have shown some examples in Figure 3. From the comparison between the Baseline method and the proposed method, we can see that our self-attention-based fusion module SMF and question-guided generator are helpful for the response generation. For one thing, SMF allows the generated answers to describe the actions and objects with higher accuracy. To specify, our answer accurately depicts the action of not walking around in test sample 138 (Figure 3(a)), expresses the state of “no emotion” by using the word “neutral” in test example 139 (Figure 3(b)), and describes the action of not finishing eating the apple in test sample in example 155 (Figure 3(c)). For another, the question-guided generator also helps the proposed method generate answers that are highly related to questions. Take test sample 138 and test sample 139 as examples. Unlike the Baseline method which sometimes makes the mistake of using wrong pronouns, our method avoids doing it by using the question as the guide of generating answers. Furthermore, in the test sample 106 (Figure 3(d)), our answer reiterates the action of “looking at” by using the question feature as answer generation guide while the Baseline method uses an ambiguous word “doing” to answer the question. Thus, our method can learn to use important objects and subjects for answer generation, which is helpful for getting a more semantic answer.

Ablation study. Finally, after manually checking the quality of generated response, we need to break down our model into modules and see the performance improvement brought by each of them. We firstly use the features of video, audio, dialog and question and utilize Baseline method to fuse them together for answer generation. The results are shown in the first row of Table 4 which are used as a benchmark for the ablation study.

The first question that we aim to investigate is how much performance improvement SMF can bring to our method. Thus, with the features f^V_ω, f^O_ω, f^A_ω, f^H_ω, f_q, we use the sim-



Question: does he walk around in the video ?
REF: he does not walk around .
Baseline: no , the video ends with her sitting on the floor .
Ours: no , he is in the room the whole time .

(a) Test sample 138

Question: does he seem happy or sad in the video ?
REF: he seems to have no emotion at all .
Baseline: she looks very giddy .
Ours: no , he is pretty neutral in the video .

(b) Test sample 139

Question: does he finish eating the apple ?
REF: he is still eating the apple when he walks away .
Baseline: no , the video ends with him still looking at the camera at the end of the video .
Ours: no , he does not finish it .

(c) Test sample 155

Question: why does she look out the window ?
REF: you can 't see anything out the window so i 'm not sure .
Baseline: i don 't know what she is doing .
Ours: i can 't tell what she is looking at .

(d) Test sample 106

Figure 3. Visualization examples of manually comparisons between Baseline method and ours.

Input	Fusion	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
$f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H, f_q$	Baseline method	0.271	0.173	0.117	0.083	0.117	0.291	0.761
$f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H, f_q$	Concatenation	**0.279**	0.175	0.118	0.083	0.117	0.290	0.732
$f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H$	SMF (1 layer)	0.206	0.119	0.076	0.051	0.087	0.217	0.434
$f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H, f_q$	SMF (3 layer)	0.271	0.174	0.119	0.085	0.118	0.292	0.778
$f_\omega^V, f_\omega^O, f_\omega^A, f_\omega^H, f_q$	SMF (1 layer)	0.276	**0.177**	**0.121**	**0.086**	**0.119**	**0.294**	**0.789**

Table 4. Ablation study results on DSTC7.

ple concatenation and the designed SMF to fuse them respectively. The results are shown in the second and the final row in Table 4. From the results, we can see that SMF outperforms concatenation in all metrics except BLEU-1. In the metric of CIDEr especially, we can see the SMF achieves a far more satisfying result than concatenation. These results show that SMF is a more effective fusion method than concatenation and the the fusion method used in the Baseline method as well.

The next question that we want to investigate is how important the question guide is for answer generation. Thus, we try to remove f_q for answer generation and use the content features and SMF only. The results are shown in the third row of Table 4. From the results, we can see that the performance decreases sharply once the question guide q_t is removed. The comparison just demonstrates that we should attach great importance to the question feature.

Finally, we also explore the number of layers in SMF, and we try to increase the self-attention layer number of SMF to 3 in order to see if it will bring performance improvement. However, as the results in the fourth row show, the performance evaluated by each metric drops a little bit. Thus, the number of self-attention layer in SMF should be set to 1.

5. Conclusions

In this paper, we propose a novel network named QUALIFIER to improve the feature fusion and answer generation technique for the Audio Video Scene-Aware Dialog problem in order to generate more semantically natural answers for questions. By separating question feature and other content features, we first use a self-attention-based multi-modal fusion module called SMF to extract the relevant information between content features of different modalities and use it for feature aggregations. In the answer generation stage, we utilize the prioritized question feature as an important guide and concatenate it with the fused features as the content feature for answer generator. From the experiments on the benchmark dataset DSTC7, we find our treatments of these two aspects are helpful for improving the generated responses with higher performance measured by commonly used metrics as well as more natural semantics judged by human evaluation.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[3] Huda Alamri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K Marks, et al. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*, 2018.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] 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.

[8] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2758–2766, 2015.

[11] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[12] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356. IEEE, 2019.

[17] 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.

[18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Shachi H Kumar, Eda Okur, Saurav Sahay, Jonathan Huang, and Lama Nachman. Exploring context, attention and audio features for audio visual scene-aware dialog. *arXiv preprint arXiv:1912.10132*, 2019.

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

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[23] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[24] 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 of the Association for Computational Linguistics*, pages 311–318, 2002.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[28] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[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.

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[33] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019.

[34] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[35] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017.

[36] Bairong Zhuang, Wenbo Wang, and Takahiro Shinozaki. Investigation of attention-based multimodal fusion and maximum mutual information objective for dstc7 track3. In *DSTC7 at AAAI 2019 workshop*, 2019.