# Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou

***Abstract*—Audio-Visual Scene-Aware Dialog (AVSD) is a task to generate responses when chatting about a given video, which is organized as a track of the 8th Dialog System Technology Challenge (DSTC8). There are two challenges in this task: 1) making effective interaction among different modalities; 2) better understanding dialogues and generating informative responses. To tackle the challenges, we propose a universal multimodal transformer and introduce the multi-task learning method to learn joint representations among different modalities as well as generate informative and fluent responses by leveraging the pre-trained language model. Our method extends the natural language generation pre-trained model to multimodal dialogue generation task, which allows fine-tuning language models to capture information across both visual and textual modalities. Our system achieves the best performance in the objective evaluation in both DSTC7-AVSD and DSTC8-AVSD dataset and achieves an impressive 98.4% of the human performance based on human ratings in the DSTC8-AVSD challenge.**

***Index Terms*—Dialogue System, Multimodal, Natural Language Processing, Video Understanding.**

## I. Introduction

RECENTLY, scene-aware dialogue generation has attracted increasing attention in both industry and academia due to its broad application. Zhou et al. [1] propose a dataset for text-based conversations grounded in documents about movies. Urbanek et al. [2] build a large-scale text adventure game platform, in which agents can act and speak grounded on the scenes described in the text. Inspired by human inherent multimodal understanding ability, Alamri et al. [3] integrate multimodality to scene-aware dialogue and propose the Audio-Visual Scene-Aware Dialog (AVSD) task. These works aim to generate informative and fluent dialogue responses grounding on the given scenes. The goal of the Audio-Visual Scene-Aware Dialog task is to generate correct and fluent responses by understanding all modalities (e.g., text, video and audio), which is a more challenging task than image-based or text-grounded dialog tasks. Figure 1 shows an example dialogue in DSTC8-AVSD dataset [3].

**Caption**: A woman standing in a hallway takes off her slippers. She then climbs on a chair and starts doing something with the ceiling light.

**Summary**: A woman about 30 years old wearing a jean skirt and top is standing on a stool and fixing something in the hallway next to a door. The hallway has linoleum floors.

**Q1**: where is the video happening ?

**A1**: it is happening inside in the hallway

**Q2**: are there any people in the video ?

**A2**: yes there is one person in the video.

…

**Q10**: what is the person doing ?

**A10**: she is standing on a stool doing something with the ceiling light.

Fig. 1. A dialogue sampled from the DSTC8-AVSD dataset. For each dialogue, there are video, audio, video caption, dialogue summary and 10 turns of conversations about the video.

There are two challenges in this task: 1) acquiring the accurate representation of the video and making effective interaction among different modalities; 2) better understanding dialogues and generating responses. Some recent works focus on the first challenge and explore a lot on multimodal representation. Hori et al. [4] introduce an LSTM-based encoder and decoder with multimodal attention. Dat Tien Nguyen and Asri [5] proposes a hierarchical recurrent encoder-decoder framework based on a FiLM-based audio-visual feature extractor. Pasunuru and Bansal [6] adopt a dual attention mechanism to encode and align multiple modalities. The winning team of the DSTC7-AVSD task [7] focuses on using hierarchical attention to combine textual and visual modalities and employ the How2 dataset for pre-training. Moreover, MTN [8] proposes multimodal transformer networks to encode video and incorporate information from different modalities. These existing methods mainly use independent encoders to separately encode different modalities and then exploit the attention mechanism to fuse the representations of different modalities, in which the lower-level representation of a single modality can not benefit from the information from the other modalities. For example, some text information like video caption is useful for the understanding of the video.

Joint work with Pattern Recognition Center, WeChat AI, Tencent Inc, China. Yang Feng is the corresponding author. This work was done when Zongjia Li was interning at Pattern Recognition Center, WeChat AI, Tencent.

Zekang Li and Yang Feng are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lizekang19g@ict.ac.cn; fengyang@ict.ac.cn).

Zongjia Li is with School of EECS, Peking University, Beijing 100080, China (e-mail: zongjiali@pku.edu.cn).

Jinchao Zhang and Jie Zhou are with Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing 100080, China (e-mail: dayerzhang@tencent.com; withtomzhou@tencent.com).

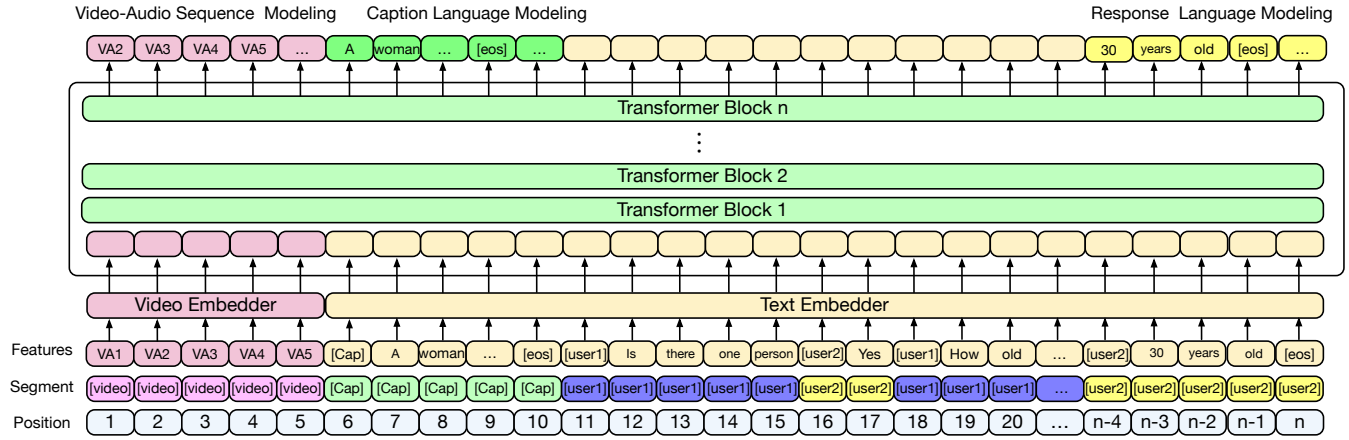


Fig. 2. Our universal multimodal transformer architecture. We concatenate video-audio, caption, dialogue history, and response features to a long sequence. For different types of input, we adopt different segments tokens ("[video]", "[caption]", "[user1]", "[user2]"). We initialize our model with pre-trained GPT2 and introduce three tasks to fine-tune our model: *Response Language Modeling (RLM)*, *Video-Audio Sequence Modeling (VASM)*, *Caption Language Modeling (CLM)*.

The second challenge of this task, scene-aware dialogue response generation, is also quite difficult. A dialogue agent needs to fully understand the dialogue history given the scene and capture relevant dependencies across dialogue turns to generate informative and correct responses. Besides, it is costly to build a large-scale scene-aware dialogue dataset, and the generation model only trained on the dataset of this task has limited performance. Adopting pre-trained language models could improve the limited dialogue datasets by leveraging rich linguistic dependencies learned from other available text data.

To tackle the aforementioned challenges, in this paper, we design a universal multimodal transformer to encode different modalities jointly and generate responses at the same time. Inspired by Bert [9], GPT2 [10], and other pre-training works, we use the self-supervised learning method and adopt the multi-task learning (response language modeling, video-audio sequence modeling, and caption language modeling) approach to learn joint representations and generate informative and fluent responses. Following the great success in many downstream dialogue generation tasks by leveraging large-scale pre-trained language models, we extend the pre-trained GPT2 [10] model to tackle the challenges by combining both visual and textual representations into a structured sequence and fine-tune it to capture cross-modal dependencies and generate informative responses.

Our contributions are as follows:

- We are the first to use pre-trained natural language generation models in multimodal dialogue generation.
- We integrate multimodal features in one encoder and introduce a multi-task learning method to learn better joint representations and generate more informative responses.
- We achieve a state-of-the-art result on Audio-Visual Scene-Aware Dialog (AVSD) Dataset with an impressive 98.4% of human performance, outperforming existing methods and other teams in DSTC8-AVSD challenge by a large margin.

## II. Related Work

Most work on the dialogue systems focuses on open-domain dialogues or task-oriented dialogues. As in human-to-human conversations, there is always background knowledge. Some recent efforts develop dialogue systems that can generate responses grounding on a document or structured knowledge graph [11, 1, 12, 13, 14]. These systems can generate responses that are either more relevant to background knowledge or make more correct interactions. There are also some works incorporating multimodal information in question answering and dialogues. In Visual QA [15, 16], the system's goal is to answer a given question about the content of an image. Visual dialog [17] is a task to generate natural responses in a dialogue based on the given image and the dialogue context. These works consider text or images as the background knowledge, whereas in Audio-Visual Scene-Aware Dialog the knowledge is text, video, and audio.

It has been shown that pre-trained language models play an important role in improving the performance of language generation tasks, such as dialogue systems and text summarization. Zhang et al. [18] propose a natural language generation model based on BERT to make good use of the pre-trained language model in the encoding and decoding process. Wolf et al. [19] introduce transfer learning to generative data-driven dialogue systems using Generative Pretrained Transformer [10]. In our work, we extend this transfer learning method to multimodal language generation tasks and propose a self-supervised learning method for better video representation.

## III. Methodology

In this section, we will describe our approaches to build the multimodal dialogue system. We will first introduce the Audio Visual Scene-Aware Dialog (AVSD) task. Then we will present our multimodal dialogue system and the training methods.

### A. Task Formulation

Our goal is to generate informative and fluent responses integrating multimodal information, which consists of video,

audio, video caption, and dialog context. Formally, let $\mathbf{V}$ and $\mathbf{A}$ represent video and audio respectively. Considering the similarity between the summary and the video caption, we concatenate summary and caption as a whole caption $\mathbf{C} = \{c_1, c_2, \ldots, c_I\}$, which typically provides a linguistic summary of the video and the whole dialogue. We use $\mathbf{U} = \{\mathbf{Q}_1, \mathbf{R}_1, \mathbf{Q}_2, \mathbf{R}_2, \ldots \mathbf{Q}_N, \mathbf{R}_N, \}$ to denote the $N$ turns of dialogue, where $\mathbf{Q}_n$ represent the question $n$ and $\mathbf{R}_n = \{r_n^1, r_n^2, \ldots, r_n^m\}$ represent the response $n$ containing $m$ words. Therefore, the probability to generate the response $\mathbf{R}_n$ for the given question $\mathbf{Q}_n$ considering video $\mathbf{V}$, audio $\mathbf{A}$, dialogue history $\mathbf{U}_{<n}$, and caption $\mathbf{C}$ can be computed as:

$$P(\mathbf{R}_n|\mathbf{V}, \mathbf{A}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n; \theta) = \prod_{j=1}^{m} P(r_n^j|\mathbf{V}, \mathbf{A}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n, r_n^{<j}; \theta) \quad (1)$$

where $r_n^{<j}$ represents the first $j-1$ words of the response $\mathbf{R}_n$.

### B. Model Overview

Our model architecture is illustrated in Figure 2, which is a multilayer Transformer model based on the GPT2 architecture [10]. More specifically, we employed a 12-layer decoder-only transformer with multi-head self-attention.

### C. Input Features

*1) Text Input:* For text features, we follow GPT2 [10] and tokenize the input sentence into WordPieces [20].

*2) Video and Audio Input:* For the given video $V_k$, we split the video to $T_k$ segments with a sliding window of $l$ video frames. As shown in Figure 3, for each segment $S_t = \{f_1, f_2, \ldots, f_l\}$, where $f_i$ represents one frame, we use a pre-trained I3D-rgb and I3D-flow model [21] to extract $d_v$-dimensional video features $\mathbf{V}_{rgb}$ and $\mathbf{V}_{flow}$. Considering audio is synchronous with video, we select the audio from the same segment and use a pretrained VGGish model [22] to extract $d_a$-dimensional audio features as $\mathbf{A}_{vggish}$. We then concatenate video I3D-rgb features, I3D-flow features, and VGGish features:

$$\mathbf{VA}_t = [\mathbf{V}_{rgb}, \mathbf{V}_{flow}, \mathbf{A}_{vggish}], \mathbf{VA}_t \in \mathbb{R}^{2d_v+d_a} \quad (2)$$

Then video-audio features $\mathbf{VA}$ are fed into a fully-connected layer (Video Embedder), as shown in Figure 2, and projected to the same embedding space as text embedding.

As shown in Figure 2, to make our model have the ability to distinguish among the different part of the input (video, caption, speaker1, and speaker2) and make use of the order of the sequence, the final representation for each word token is obtained via summing up its word embedding (WE), positional encoding (PE) and segment embedding (SE). Note that "[video]", "[cap]", "[user1]", and "[user2]" are used to represent the segment of video, captions and summary, speaker1, and speaker2 respectively.

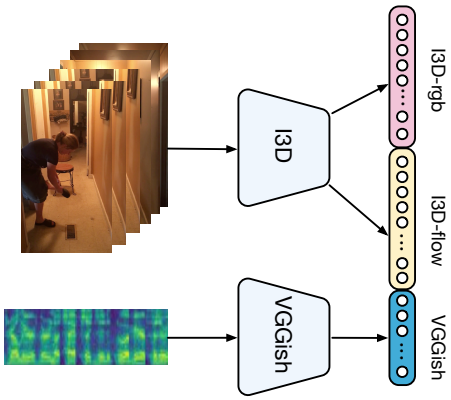


Fig. 3. Video and audio feature extractors. For video, we adopt pre-trained I3D-rgb and I3D-flow to extract rgb features and optical flow features. For audio, we use pre-trained VGGish model.

### D. Multi-task Learning

We introduce three tasks to fine-tune our model: *Response Language Modeling* conditioned on video, audio, caption and dialogue history, *Video-Audio Sequence Modeling* conditioned on caption and dialogue, and *Caption Language Modeling* conditioned on video and audio.

*1) Response Language Modeling (RLM):* The goal of this task is to generate responses $\mathbf{R}_n = \{r_n^1, r_n^2, \ldots, r_n^m\}$ based on the video-audio features $\mathbf{VA}$, caption $\mathbf{C}$, dialogue history $\mathbf{U}_{<n}$, and question $\mathbf{Q}_n$, by minimizing the negative log-likelihood loss function:

$$\mathcal{L}_{RLM}(\theta) = -E_{(\mathbf{VA},\mathbf{C},\mathbf{U},\mathbf{Q},\mathbf{R})\sim D} \log \prod_{j=0}^{m} P(r_n^j|\mathbf{VA}, \mathbf{C}, \mathbf{U}_{<n}, \mathbf{Q}_n, r_n^{<j}) \quad (3)$$

where $r_n^{<j}$ represents the first $j-1$ words of the response $\mathbf{R}_n$, $\theta$ represents the trainable parameters, and $(\mathbf{VA}, \mathbf{C}, \mathbf{U}, \mathbf{Q})$ sets are sampled from the whole training set $D$.

*2) Video-Audio Sequence Modeling (VASM):* This task is to predict video-audio features given caption and dialogue history. Unlike textual tokens which are represented as discrete labels, video-audio features are high-dimensional and continuous. Instead of clustering video-audio features to discrete labels as Sun et al. [23] do, we adopt the video-audio feature regression method following [24]. This task regresses the Transformer output of video-audio feature $\mathbf{o}_t$ to the next video-audio feature $\mathbf{VA}_{t+1}$. In particular, we apply a fully-connected layer to transform the output to a vector $g_\theta(\mathbf{o}_t)$ of the same dimensional as $\mathbf{VA}_{t+1}$. We train this task by minimizing L2 loss:

$$\mathcal{L}_{VASM}(\theta) = E_{(\mathbf{VA},\mathbf{C},\mathbf{U})\sim D} \frac{1}{T}\sum_{t=1}^{T} \|g_\theta(\mathbf{o}_t) - \mathbf{VA}_{t+1}\|_2^2 \quad (4)$$

where $\mathbf{o}_t = f_\theta(\mathbf{VA}_{<t+1}, \mathbf{C}, \mathbf{U})$ and $f_\theta$ represents the function of our GPT2 model.

TABLE I
OBJECTIVE EVALUATION RESULTS ON THE TEST SET PROVIDED BY THE ORGANIZERS IN DSTC7-AVSD CHALLENGE (6 GROUNDTRUTH RESPONSES ARE AVAILABLE PER VIDEO).

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|
| *Input: text-only* | | | | | | | |
| JMAN | 0.644 | 0.488 | 0.383 | 0.302 | 0.238 | 0.518 | 0.891 |
| Hierarchical Attention | - | - | - | 0.376 | 0.264 | 0.554 | 1.076 |
| **Our model (RLM)** | **0.747** | **0.627** | **0.527** | **0.445** | **0.287** | **0.594** | **1.261** |
| *Input: text + video* | | | | | | | |
| JMAN | 0.667 | 0.521 | 0.413 | 0.334 | 0.239 | 0.533 | 0.941 |
| Hierarchical Attention | - | - | - | 0.394 | 0.267 | 0.563 | 1.094 |
| MSTN | - | - | - | 0.377 | 0.275 | 0.566 | 1.115 |
| MTN | - | - | - | 0.392 | 0.269 | 0.559 | 1.066 |
| **Our model (RLM)** | 0.759 | 0.635 | 0.533 | 0.448 | 0.293 | 0.602 | 1.282 |
| **+ VASM** | **0.765** | **0.643** | **0.543** | **0.459** | **0.294** | **0.606** | **1.308** |
| *Input: text + video w/o caption / summary* | | | | | | | |
| Baseline | - | - | - | 0.309 | 0.215 | 0.487 | 0.746 |
| DSTC7-AVSD Team 9 | - | - | - | 0.315 | 0.239 | 0.481 | 0.773 |
| MSTN | - | - | - | 0.379 | **0.261** | **0.548** | 1.028 |
| **Our model (RLM)** | **0.694** | **0.570** | **0.476** | **0.402** | 0.254 | 0.544 | **1.052** |
| **+ VASM** | 0.677 | 0.556 | 0.462 | 0.389 | 0.250 | 0.533 | 1.004 |
| **+ CLM** | 0.670 | 0.537 | 0.438 | 0.362 | 0.254 | 0.535 | 1.022 |

TABLE II
OBJECTIVE AND SUBJECTIVE EVALUATION RESULTS ON THE TEST SET PROVIDED BY THE ORGANIZERS IN DSTC8-AVSD CHALLENGE (6 GROUNDTRUTH RESPONSES ARE AVAILABLE PER VIDEO). NOTE THAT OUR MODEL (+VASM) IS ADDITIONAL EXPERIMENT AFTER THE CHALLENGE, SO THERE IS NO SUBJECTIVE EVALUATION FOR IT. HUMAN RATING FOR THE REFERENCE IS 4.000.

| Models | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | Human rating |
|---|---|---|---|---|---|---|---|---|
| *Input: text-only* | | | | | | | | |
| **Our model (RLM)** | **0.744** | **0.626** | **0.525** | **0.442** | **0.287** | **0.595** | **1.231** | **3.934** |
| *Input: text + video* | | | | | | | | |
| JMAN | 0.645 | 0.504 | 0.402 | 0.324 | 0.232 | 0.521 | 0.875 | 3.123 |
| STSGR | - | - | - | 0.357 | 0.267 | 0.553 | 1.004 | 3.433 |
| MSTN | - | - | - | 0.385 | 0.270 | 0.564 | 1.073 | - |
| **Our model (RLM)** | 0.739 | 0.624 | 0.528 | **0.447** | 0.284 | 0.592 | 1.226 | 3.895 |
| **+ VASM** | **0.746** | **0.626** | **0.528** | 0.445 | **0.286** | **0.598** | **1.240** | - |
| *Input: text + video w/o caption / summary* | | | | | | | | |
| Baseline | - | - | - | 0.289 | 0.210 | 0.480 | 0.651 | 2.885 |
| MSTN | - | - | - | 0.375 | **0.251** | 0.544 | 0.975 | - |
| **Our model (RLM)** | **0.677** | **0.556** | **0.462** | **0.387** | 0.249 | **0.544** | **1.022** | - |
| **+ VASM** | 0.669 | 0.550 | 0.457 | 0.385 | 0.246 | 0.540 | 0.988 | - |
| **+ CLM** | 0.661 | 0.533 | 0.437 | 0.364 | 0.242 | 0.533 | 0.991 | - |

*3) Caption Language Modeling (CLM):* Similar to Response Language Modeling task, we train the model to generate caption $\mathbf{C} = \{c_1, c_2, \ldots, c_I\}$ based on the video-audio feature $\mathbf{VA}$ by minimizing the negative loglikelihood loss function:

$$\mathcal{L}_{CLM}(\theta) = -E_{(\mathbf{VA},\mathbf{C})\sim D} \log \prod_{i=0}^{I} P(c_i|\mathbf{VA}, c_{<i}) \quad (5)$$

where $c_{<i}$ represents the first $i-1$ words of the caption $\mathbf{C}$.

## IV. EXPERIMENTS

### A. Datasets

We use the Audio-Visual Scene-Aware Dialog (AVSD) dataset [3] from DSTC7 and DSTC8. In this dataset, each dialog has two participants, a questioner and an answerer. Each dialogue consists of a sequence of questions and answers about a given video. There is a video caption and a dialogue summary for each video. The video caption is a description of the given video. The Dialogue summary is a summarization of the dialogue. We use the state-of-the-art video feature extractor I3D model [21] pre-trained on YouTube videos and the Kinetics dataset [25]. Specifically, we use the output from the "Mixed 5c" layer of the I3D network, which is a 2048-dimensional vector. For audio features, we adopt the famous VGGish model [22] which outputs a 128-dimensional embedding. There are 7,659 dialogues for training, 1787 dialogues for validation, and 1710 dialogues for testing. In DSTC7 and DSTC8, the training set and the validation set are the same, while the testing sets are different. We evaluate our model on both two datasets.

### B. Baselines

We compare our model with several related baseline methods: the official baseline model, the DSTC7-AVSD winning system, and some of the other DSTC8-AVSD submitted systems:

*1) Baseline:* The multimodal baseline provided by the organizers, which combines all modalities with a projection matrix [4].

TABLE III
OBJECTIVE EVALUATION RESULTS ON THE TEST SET OF DSTC7-AVSD (6 GROUNDTRUTH RESPONSES ARE AVAILABLE PER VIDEO) IN WHICH MAXIMUM HISTORY LENGTH (NUMBER OF DIALOGUE TURNS USED BY THE NETWORK) RANGES FROM 0 TO 9. BEST RESULT IN EACH METRIC IS HIGHLIGHTED IN BOLD.

| History Length | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|
| **0** | 0.729 | 0.599 | 0.496 | 0.413 | 0.275 | 0.573 | 1.182 |
| **1** | 0.760 | 0.638 | 0.536 | 0.452 | **0.296** | 0.605 | 1.305 |
| **2** | 0.755 | 0.632 | 0.532 | 0.450 | **0.296** | 0.601 | 1.297 |
| **3** | **0.765** | **0.643** | **0.543** | **0.459** | 0.294 | **0.606** | **1.308** |
| **5** | 0.758 | 0.634 | 0.533 | 0.451 | 0.292 | 0.601 | 1.293 |
| **9** | 0.759 | 0.631 | 0.526 | 0.441 | **0.296** | 0.603 | 1.294 |

TABLE IV
OBJECTIVE EVALUATION RESULTS ON THE TEST SET OF DSTC7-AVSD (6 GROUNDTRUTH RESPONSES ARE AVAILABLE PER VIDEO) COMPARED BETWEEN DIFFERENT DECODING METHODS. BEST RESULT IN EACH METRIC IS HIGHLIGHTED IN BOLD.

| Decoding Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|
| **Greedy Search** | 0.743 | 0.610 | 0.503 | 0.416 | 0.284 | 0.587 | 1.217 |
| **Nucleus Sampling** | 0.680 | 0.525 | 0.410 | 0.321 | 0.252 | 0.527 | 0.955 |
| **Beam Search** | **0.765** | **0.643** | **0.543** | **0.459** | **0.294** | **0.606** | **1.308** |

*2) Hierarchical Attention:* The hierarchical attention approach to combine textual and visual modalities. This is the method that was used by the team ranked $1^{st}$ in the DSTC7-AVSD task.

*3) MTN:* The state-of-the-art system before the DSTC8-AVSD Challenge, which proposes Multimodal Transformer Networks (MTN) to encode videos and incorporate information from different modalities [8].

*4) JMAN:* Multi-step joint-modality attention network based on RNN, which performs a multi-step attention mechanism and jointly considers both visual and textual representations [26].

*5) MSTN:* MSTN employs a transformer-based architecture with an attention-based word embedding layer considering the meaning of words at the generation stage [27].

*6) STSGR:* This work represents a video as two spatio-temporal scene graphs, which encode graphs via graph attention and perform high-level reasoning using multi-modal transformers [28].

### C. Metrics

*1) Objective evaluation:* We report the metrics that are commonly used in the natural language generation tasks, such as BLEU [29], METEOR [30], ROUGE-L [31], and CIDEr [32]. These metrics are formulated to compute the word overlap between the predicted responses and the ground-truth responses. We evaluate our models using the toolkit provided by the DSTC8-AVSD challenge organizers.

*2) Subjective Evaluation:* Subjective evaluations are essential for dialogue generation. The organizers evaluated some systems based on crowd-sourced human ratings. The annotators were asked to consider the correctness, naturalness, informativeness, and appropriateness of the generated responses and gave a score at five levels, from 1 to 5. The human-generated reference responses' rating is 4.000.

### D. Experimental Settings

In our experiment, we initialize our model with the pre-trained weights from the GPT2 base model [10, 33]. In the training process, we use up to 3 turns of dialogue history. The hidden size of the transformer blocks is 768, and the batch size is 32. We use Adam optimizer with a learning rate of 6.25e-5. In the decoding process, we use beam search with a beam size of 5, max length of 20, and a length penalty of 0.3.

For the Audio-Visual Scene-Aware Dialog task, there are three different settings: text-only, text+video, and text+video without caption. Text-only setting is equivalent to text knowledge grounded dialogue generation task. Text+video setting is a complete scene-aware dialogue generation task considering both textual and visual information. Text+video without caption mainly focus on the situation where there is only video but no video caption, which is typical in the real world.

### E. Experimental Results

In this section, we report the experimental results under three different settings: text-only, text + video, and text + video without caption/summary.

*1) Text-only:* We use only text input, which comprises dialogue history, video caption, and dialogue summary. We train the model using only the Response Language Modeling (RLM) task. These results are reported in the "Our model (RLM)" row of "text-only" setting in Table I and Table II.

As shown in Table I, compared to Hierarchical Attention (which was used in the winning system of the DSTC7-AVSD challenge) and JMAN, our model gets better performance on all metrics on the DSTC7-AVSD dataset. In detail, our model improves BLEU-4 by 0.069 and CIDEr by 0.185. Additionally, Table II also shows the human evaluation rating in the DSTC8-AVSD track. During the human evaluation, the evaluators are asked to rate even the groundtruth references, which are scored 4.000. Our model for this task scores 3.934, which is the highest human rating among all DSTC8 submissions. From the human rating perspective, the results of our model are very close to human dialogue.

*2) Text + video:* We use text input and video-audio input, and we train the model with the Response Language Modeling (RLM) task and the Video-Audio Sequence Modeling (VASM)

TABLE V
CASE STUDY OF RESPONSE GENERATION. TEXT-ONLY INFORMATION IS ENOUGH TO PROVIDE A GOOD RESPONSE FOR THE FIRST CASE, BUT THAT INCLUDING VIDEO IMPROVES THE PERFORMANCE FOR THE OTHER TWO CASES.

| | |
|---|---|
| Video caption | A man with glasses wearing a green t-shirt is playing with a device on a counter top in a kitchen. |
| Dialogue summary | Male is standing at an oven, flipping and tapping a spatula. They then go and close the fridge door. |
| Dialogue history | [User1]: So, this person is standing in that spot for the entire length of time, fiddling with something?<br>[User2]: He stands there and flips the item around, and taps it, then at the very end, he goes and closes the fridge door.<br>[User1]: Is this person wearing glasses? |
| Baseline | [User2]: No, they are not wearing glasses. |
| Our model (text-only) | [User2]: Yes, he is wearing glasses. |
| Our model (text + video) | [User2]: Yes, he is wearing glasses. |
| Ground-truth | [User2]: Yes, he is wearing glasses. |
| Video caption | A person stands in a bedroom holding a folded sheet. They are at a panel of switches. They turn on the light. |
| Dialogue summary | A man stands with a folded sheet in one hand, and a panel of switches on the wall. He flips one light switch on. |
| Dialogue history | [User1]: Does the person hold it during the whole video?<br>[User2]: Yes, while standing at the panel of switches, he is holding the folded sheet the entire time.<br>[User1]: What does the man touch with his left hand on the wall? |
| Baseline | [User2]: He doesn't touch anything. |
| Our model (text-only) | [User2]: He touches with his right hand on the wall. |
| Our model (text + video) | [User2]: He flips a light switch on. |
| Ground-truth | [User2]: He is touching a panel of switches and he turns on the light. |
| Video caption | A man is laying in bed watching tv. He mumbles something and then gets up and begins vacuuming the room. |
| Dialogue summary | A man is lying on a bed. He gets up and uses a vacuum to sweep the room. |
| Dialogue history | [User1]: In the door way to the next room, there is an object. What is that ?<br>[User2]: He went to the doorway for a vacuum.<br>[User1]: When he gets up, does he have anything in his hands? |
| Baseline | [User2]: No he is not talking. |
| Our model (text-only) | [User2]: He has a vacuum in his hand. |
| Our model (text + video) | [User2]: No, he has nothing in his hands. |
| Ground-truth | [User2]: He had nothing in his hands when he got up. |

task described in Section III.D. These results are reported in the "Our model (RLM) + VASM" row of "text + video" setting in Table I and Table II.

As we can see in Table I, compared to MTN (the former state-of-the-art model for this task), our model also achieves a huge improvement. In particular, our model improved the BLEU-4 score by 0.056, and CIDEr by 0.216. Compared to the text-only task, our models achieve better results on the objective evaluations, which indicates that our method for video understanding is effective. We adopt multi-task learning as we described before. Video-Audio sequence modeling (VASM) task improves the score of BLEU-4 by 0.011 and CIDEr by 0.026. In the DSTC8 results shown in Table II, this method improves upon CIDEr score by 0.014, which shows the method is effective. Note that our model (+VASM) is an additional experiment after the challenge, so there is no subjective evaluation for it. The human performance of "text + video" is theoretically better than that of "text only" through many case studies.

*3) Text + video w/o caption/summary:* In this setting, there are two methods: 1) In both training and testing, use neither the caption nor the dialogue summary. Train the model with Response Language Modeling (RLM) and Video-Audio Sequence Modeling (VASM). The results are reported in the "Our model (RLM) + VASM" row of "text + video w/o caption and summary" setting in Table I and Table II. 2) In training, use captions and summary and train the model with the three tasks described in Section III.D. When testing, first generate video captions and summary based on the given video-audio input (recaption), and then generate responses using video-audio input, generated caption, and dialogue history. The results are reported in the "Our model (RLM) + VASM + CLM" row of "text + video w/o caption and summary" setting in Table I and Table II.

This setting is most similar to the real world scene-aware dialogue: we only have video-audio information and dialogue history. Therefore, this task is more challenging. As shown in Table I, we see lower performance than the text + video task, as expected, but it is gratifying that our model still performed relatively well. We outperform the DSTC7-AVSD Team 9, who got the highest performance in this task, by a large margin. In this task, we also tried using the multi-task learning method including video-audio sequence modeling (VASM) and caption language modeling (CLM), but this resulted in lower performance on almost all metrics. We will discuss this phenomenon in the next section.

### *F. Analysis and Discussion*

*1) Training Method Analysis:* As we introduced in the experimental results, after we adopt the visual feature regression, our model gets better performance in the text + video task, but gets lower performance in the text + video w/o caption/summary task. We consider the reason is that it is difficult to rebuild the masked video feature only using the dialogue history without the captions and summary. Compared to rebuilding text from adjacent context, we think our model doesn't have a very strong ability in extracting information from videos. Therefore, the method doesn't perform very well in the text + video w/o caption/summary setting.

For the poor performance of CLM, we think the reason may be similar: the limited ability in extracting video information of our model limited the performance for inferring caption from the video. So we think future work can focus more on video comprehension to get better basic video features.

*2) History Length:* We experiment with our model in text + video settings with Video-Audio Sequence Modeling (VASM) loss to explore the influence of dialogue history length. As shown in Table III, our model generally performs best when the maximum dialogue history length is 3.

*3) Decoding Methods:* To find an effective decoding method for multimodal dialogue generation, we try various decoding methods, including greedy search, beam search, and nucleus sampling [34], which samples text from the dynamic nucleus of the probability distribution and is often used to generate diverse text. As shown in Figure IV, decoding with beam search gets the best results on all objective metrics among these three decoding methods. We consider that in Audio-Visual Scene-Aware Dialog, grounding on video and caption, responses are relatively more definite than that in open-domain dialogues. Therefore, it is better to use beam search when decoding in this task.

### G. Case Study

Table V compares the responses generated by the baseline model, our model (text-only), and our model (text + video). Compared to the baseline model, our model can generate more informative responses. As shown in case 1, for our text-only model, it performs well when the information can be found in the caption/summary. However, as shown in case 2 and case 3, when it comes to asking about specific information from the video that is not present in the captions or summary, our text-only model does not perform well. In these cases, our text + video model can refer to the video, find related information, and generate correct responses.

## V. Conclusion and Future Work

In this paper, we propose a universal multimodal dialogue generation model based on a pre-trained language model for Audio-Visual Scene-Aware Dialog. We also introduce three tasks to fine-tune our model: Response Language Modeling, Video-Audio Sequence Modeling, Caption Language Modeling. Through these tasks, the model can learn more accurate joint representation across multiple modalities and generate more informative responses. Our system achieves the best performance in the objective evaluation in both DSTC7-AVSD and DSTC8-AVSD dataset and achieves an impressive 98.4% of the human performance based on human ratings in the DSTC8-AVSD challenge. In the future, we plan to use more video features, such as ResNet features, and explore more training tasks to improve the joint understanding of video and text. Also, we hope to extend these methods to other tasks, such as video captioning, image captioning, and visual dialog.

## VI. Acknowledgement

We sincerely thank the anonymous reviewers and editors for their thorough reviewing and valuable suggestions. This work is supported by National Key R&D Program of China (NO. 2017YFE0192900 and NO. 2018YFC0825201).

## References

[1] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," *arXiv preprint arXiv:1809.07358*, 2018.

[2] J. Urbanek, A. Fan, S. Karamcheti, S. Jain, S. Humeau, E. Dinan, T. Rocktschel, D. Kiela, A. Szlam, and J. Weston, "Learning to speak and act in a fantasy text adventure game," *arXiv preprint arXiv:1903.03094*, 2019.

[3] H. Alamri, C. Hori, T. K. Marks, D. Batra, and D. Parikh, "Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7," in *DSTC7 at AAAI2019 Workshop*, vol. 2, 2018.

[4] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. 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)*. IEEE, 2019, pp. 2352–2356.

[5] H. S. Dat Tien Nguyen, Shikhar Sharma and L. E. Asri, "From film to video: Multi-turn question answering with multi-modal context," in *DSTC7 at AAAI2019 Workshop*, 2019.

[6] R. Pasunuru and M. Bansal, "Dstc7-avsd: Scene-aware video-dialogue systems with dual attention," in *DSTC7 at AAAI2019 Workshop*, 2019.

[7] R. Sanabria, S. Palaskar, and F. Metze, "Cmu sinbad's submission for the dstc7 avsd challenge," in *DSTC7 at AAAI2019 Workshop*, 2019.

[8] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5612–5623.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[11] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental transformer with deliberation decoder for document grounded conversations," *arXiv preprint arXiv:1907.08854*, 2019.

[12] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[13] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[14] A. Madotto, C.-S. Wu, and P. Fung, "Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," *arXiv preprint arXiv:1804.08217*, 2018.

[15] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[16] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.

[17] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 326–335.

[18] H. Zhang, Y. Gong, Y. Yan, N. Duan, J. Xu, J. Wang, M. Gong, and M. Zhou, "Pretraining-based natural language genera-

tion for text summarization," *arXiv preprint arXiv:1902.09243*, 2019.
[19] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *arXiv preprint arXiv:1901.08149*, 2019.
[20] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.
[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.
[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
[23] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," *arXiv preprint arXiv:1904.01766*, 2019.
[24] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv preprint arXiv:1909.11740*, 2019.
[25] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[26] Y.-W. Chu, K.-Y. Lin, C.-C. Hsu, and L.-W. Ku, "Multi-step joint-modality attention network for scene-aware dialogue system," *arXiv preprint arXiv:2001.06206*, 2020.
[27] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator," *arXiv preprint arXiv:2004.08299*, 2020.
[28] S. Geng, P. Gao, C. Hori, J. L. Roux, and A. Cherian, "Spatio-temporal scene graphs for video dialog," *arXiv preprint arXiv:2007.03848*, 2020.
[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
[30] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.
[31] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[32] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.
[34] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

**Zekang Li** received the Bachelor's degree in electronic information engineering from Huazhong University of Science and Technology, China, in 2019. He worked as the AI Group Leader of Dian Group (2017-2019) (Dian Group ID: D610).

He is currently working toward the Master degree in NLP Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, advised by Prof. Yang Feng. Now he is in a research intern at Pattern Recognition Center, WeChat AI, Tencent, advised by Jinchao Zhang, Fandong Meng and Cheng Niu. His research interests lie within deep learning for Natural Language Processing, particularly in dialogue systems and multimodal representation.

**Zongjia Li** Zongjia Li is currently an undergraduate in the School of EECS of Peking University, Beijing, China. His research interests include natural language processing and machine learning.

**Jinchao Zhang** Jinchao Zhang received his Ph.D. degree in computer software theory from Chinese Academy of Science (CAS), China, in 2018. He is currently working at Pattern Recognition Center, WeChat AI, Tencent Ltd. as a senior research scientist. His research interests lie within deep learning for Natural Language Processing, particularly in Dialogue systems and Machine Translaton.

**Yang Feng** Yang Feng is an Associate Professor in Institute of Computing Technology, Chinese Academy of Sciences where she got her phD degree in 2011. She worked in University of Sheffield and Information Sciences Institute, University of Southern California from 2011 to 2014. Now she leads the natural language process group in ICT/CAS and her research interest is natural language processing, mainly focusing on machine translation and dialogue. She was the recipient of the Best Long Paper Award of ACL 2019.

**Jie Zhou** Jie Zhou received his bachelor degree from USTC in 2004 and his Ph.D. degree from Chinese Academy of Sciences in 2009, and is now a senior director of Pattern Recognition Center, WeChat AI, Tencent Inc. His research interests include natural language processing and machine learning.