

# Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems

Hung Le<sup>1,2</sup>, Doyen Sahoo<sup>1</sup>, Nancy F. Chen<sup>2</sup>, Steven C.H. Hoi<sup>1,3</sup>

<sup>1</sup>Singapore Management University

<sup>2</sup>Institute of Inforcomm Research (I2R), Singapore

<sup>3</sup>Salesforce Research Asia

{hungle.2018, doyens}@smu.edu.sg

nfychen@i2r.a-star.edu.sg, chhoi@smu.edu.sg

## Abstract

Developing Video-Grounded Dialogue Systems (VGDS), where a dialogue is conducted based on visual and audio aspects of a given video, is significantly more challenging than traditional image or text-grounded dialogue systems because (1) feature space of videos span across multiple picture frames, making it difficult to obtain semantic information; and (2) a dialogue agent must perceive and process information from different modalities (audio, video, caption, etc.) to obtain a comprehensive understanding. Most existing work is based on RNNs and sequence-to-sequence architectures, which are not very effective for capturing complex long-term dependencies (like in videos). To overcome this, we propose Multimodal Transformer Networks (MTN) to encode videos and incorporate information from different modalities. We also propose query-aware attention through an auto-encoder to extract query-aware features from non-text modalities. We develop a training procedure to simulate token-level decoding to improve the quality of generated responses during inference. We get state of the art performance on Dialogue System Technology Challenge 7 (DSTC7). Our model also generalizes to another multimodal visual-grounded dialogue task, and obtains promising performance.

## 1 Introduction

A video-grounded dialogue system (VGDS) generates appropriate conversational response to queries of humans, by not only keeping track of the relevant dialogue context, but also understanding the relevance of the query in the context of a given video (knowledge grounded in a video) (Hori et al., 2018). An example dialogue exchange can be seen in Figure 1. Developing such systems has recently received interest from the research community (e.g. DSTC7 challenge (Yoshino et al., 2018)). This task is much



C: a man is standing in a kitchen putting groceries away. He closes the cabinet when finished, walks over to a table and pulls out a chair and sits down.

S: a man puts away his groceries and then sits at a kitchen table and stares out the window.

Q1: how many people are in the video?

A1: there is just one person

Q2: is there sound to the video?

A2: yes there is audio but no one is talking

...

Q10: is he happy or sad?

A10: he appears to be neutral in expression

Figure 1: A sample dialogue from the DSTC7 Video Scene-aware Dialogue training set with 4 example video scenes. *C*: Video Caption, *S*: Video Summary, *Q<sub>i</sub>*: *i<sup>th</sup>*-turn question, *A<sub>i</sub>*: *i<sup>th</sup>*-turn answer

more challenging than traditional text-grounded or image-grounded dialogue systems because: (1) feature space of videos is larger and more complex than text-based or image-based features because of diverse information, such as background noise, human speech, flow of actions, etc. across multiple video frames; and (2) a conversational agent must have the ability to perceive and comprehend information from different modalities (text from dialogue history and human queries, visual and audio features from the video) and semantically shape a meaningful response to humans.

Most existing approaches for multi-modal dialogue systems are based on RNNs as the sequence processing unit and sequence-to-sequence network as the overall architecture to model the sequential information in text (Das et al., 2017a,b; Hori et al., 2018; Kottur et al., 2018). Some efforts adopted query-aware attention to allow the models

to focus on specific parts of the features most relevant to the dialogue context (Hori et al., 2018; Kottur et al., 2018). Despite promising results, these methods are not very effective or efficient for processing video-frames, due to the complexity of long term sequential information from multiple modalities. We propose Multimodal Transformer Networks (MTN) which model the complex sequential information from video frames, and also incorporate information from different modalities. MTNs allow for complex reasoning over multimodal data such as in videos, by jointly attending to information in different representation subspaces, and making it easier (than RNNs) to fuse information from different modalities. Inspired by the success of Transformers (Vaswani et al., 2017) for text, we propose novel neural architectures for VGDS: (1) We propose to capture complex sequential information from video frames using multi-head attention layers. Multi-head attention is applied across several modalities (visual, audio, captions) repeatedly. This works like a memory network to allow the models to comprehensively reason over the video to answer human queries; (2) We propose an auto-encoder component, designed as query-aware attention layer, to further improve the reasoning capability of the models on the non-text features of the input videos; and (3) We employ a training approach to improve the generated responses by simulating token-level decoding during training.

We evaluated MTN on a video-grounded dialogue dataset (released through DSTC7 (Yoshino et al., 2018)). In each dialogue, video features such as audio, visual, and video caption, are available, which have to be processed and understood to hold a conversation. We conduct comprehensive experiments to validate our approach, including automatic evaluations, ablations, and qualitative analysis of our results. We also validate our approach on the visual-grounded dialogue task (Das et al., 2017a), and show that MTN can generalize to other multimodal dialog systems.

## 2 Related Work

The majority of work in dialogues is formulated as either open-domain dialogues (Shang et al., 2015; Vinyals and Le, 2015; Yao et al., 2015; Li et al., 2016a,b; Serban et al., 2017, 2016) or task-oriented dialogues (Henderson et al., 2014; Borde and Weston, 2016; Fatemi et al., 2016; Liu

and Lane, 2017; Lei et al., 2018; Madotto et al., 2018). Some recent efforts develop conversational agents that ground their responses on external knowledge, e.g. online encyclopedias (Dinan et al., 2018), social networks, or user recommendation sites (Ghazvininejad et al., 2018). The agent generates a response that can relate to the current dialogue context as well as exploit the information source. Recent dialogue systems use Transformer principles (Vaswani et al., 2017) for incorporating attention and focus on different dialogue settings, e.g. text-only or response selection settings (Zhu et al., 2018; Mazaré et al., 2018; Dinan et al., 2018). These approaches consider the knowledge to be grounded in text, whereas in VGDS, the knowledge is grounded in videos (with multimodal sources of information).

There are a few efforts in NLP domain, where multimodal information needs to be incorporated for the task. Popular research areas include image captioning (Vinyals et al., 2015; Xu et al., 2015), video captioning (Hori et al., 2017; Li et al., 2018) and visual question-answering (QA) (Antol et al., 2015; Goyal et al., 2017). Image captioning and video captioning tasks require to output a description sentence about the content of an image or video respectively. This requires the models to be able to process certain visual features (and audio features in video captioning) and generate a reasonable description sentence. Visual QA involves generating a correct response to answer a factual question about a given image. The recently proposed movie QA (Tapaswi et al., 2016) task is similar to visual QA but the answers are grounded in movie videos. However, all of these methods are restricted to answering specific queries, and do not maintain a dialogue context, unlike what we aim to achieve in VGDS. We focus on generating dialogue responses rather than selecting from a set of candidates. This requires the dialogue agents to model the semantics of the visual and/or audio contents to output appropriate responses.

Another related task is visual dialogues (Das et al., 2017a,b; Kottur et al., 2018). This is similar to visual QA but the conversational agent needs to track the dialogue context to generate a response. However, the knowledge is grounded in images. In contrast, we focus on knowledge grounded in videos, which is more complex, considering the large feature space spanning across multiple video frames and modalities that need to be understood.

### 3 Multimodal Transformer Networks

Given an input video  $V$ , its caption  $C$ , a dialogue context of  $(t - 1)$  turns, each including a pair of (question, answer)  $(Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})$ , and a factual query  $Q_t$  on the video content, the goal of a VGDS is to generate an appropriate dialogue response  $A_t$ . We follow the attention-based principle of Transformer network (Vaswani et al., 2017) and propose a novel architecture: *Multimodal Transformer Networks* to elegantly fuse feature representations from different modalities. MTN enables complex reasoning over long video sequences by attending to important feature representations in different modalities.

MTN comprises 3 major components: encoder, decoder, and auto-encoder layers. (i) *Encoder layers* encode text sequences and input video into continuous representations. Positional encoding is used to inject the sequential characteristics of input text and video features at token and video-frame level respectively; (ii) *Decoder layers* project the target sequences and perform reasoning over multiple encoded features through a multi-head attention mechanism. Attention layers coupled with feed-forward and residual connections process the projected target sequence over  $N$  attention steps before passing to a generative component to generate a response; (iii) *Auto-encoder layers* enhance video features with a query-aware attentions on the visual and audio aspects of the input video. A network of multi-head attentions layers are employed as a query auto-encoder to learn the attention in an unsupervised manner. We combine these modules as a Multimodal Transformer Network (MTN) model and jointly train the model end-to-end. An overview of the MTN architecture is shown in Figure 2. Next, we will discuss the details of each of these components.

#### 3.1 Encoder Layers

**Text Sequence Encoders.** The encoder layers map each sequence of tokens  $(x_1, \dots, x_n)$  to a sequence of continuous representation  $z = (z_1, \dots, z_n) \in \mathbb{R}^d$ . An overview of text sequence encoder can be seen in Figure 3. The encoder is composed of a token-level learned embedding, a fixed positional encoding layer, and layer normalization. We use the positional encoding to incorporate sequential information of the source sequences. The token-level positional embedding is added on top of the embedding layer by us-

ing element-wise summation. Both learned embedding and positional encoding has the same dimension  $d$ . We used the sine and cosine functions for the positional encoding as similarly adopted in (Vaswani et al., 2017). Compared to a Transformer encoder, we do not use stack of encoder layers with self-attention to encode source sequences. Instead, we only use layer normalization (Ba et al., 2016) on top of the embedding. We also experimented with using stacked Transformer encoder blocks, consisting of self-attention and feed-forward layers, and compare with our approach (see Table 4 Row A and B-1). The target sequence  $A_t = (y_1, \dots, y_m)$  is offset by one position to ensure that the prediction in the decoding step  $i$  is auto-regressive only on the previously positions  $1, \dots, (i - 1)$ . Here we share the embedding weights of encoders for source sequences i.e. query, video caption, and dialogue history.

**Video Encoders.** For a given video  $V$ , its features are extracted with a sliding window of  $n$ -video-frame length. This results in modality feature vector  $f_m \in \mathbb{R}^{num.Seqs \times d_m}$  for a modality  $m$ . Each  $f_m$  represents the features for a sequence of  $n$  video frames. Here we consider both visual and audio features  $M = (v, a)$ . We use pre-trained feature extractors and keep the weights of the extractors fixed during training. For a set of scene sequences  $s_1, \dots, s_v$ , the extracted features for modality  $m$  is  $f_m = (f_1, \dots, f_v)$ . We apply a linear network with ReLU activation to transform the feature vectors from  $d_m$ - to  $d$ -dimensional space. We then also employ the same positional encoding as before to inject sequential information into  $f_m$ . Refer to Figure 3 for an overview of video encoder.

#### 3.2 Decoder Layers

Given the continuous representation  $z_s$  for each source sequence  $x_s$  and  $z_t$  for the offset target sequence, the decoder generates an output sequence  $(y_2, \dots, y_m)$  (The first token is always an  $\langle sos \rangle$  token). The decoder is composed of a stack of  $N$  identical layers. Each layer has  $4 + \|M\|$  sub-layers, each of which performs attention on an individual encoded input: the offset target sequence  $z_t$ , dialogue history  $z_{his}$ , video caption  $z_{cap}$ , user query  $z_{que}$ , and video non-text features  $\{f_a, f_v\}$ . Each sub-layer consists of a multi-head attention mechanism and a position-wise feed-forward layer. Each feed-forward network con-

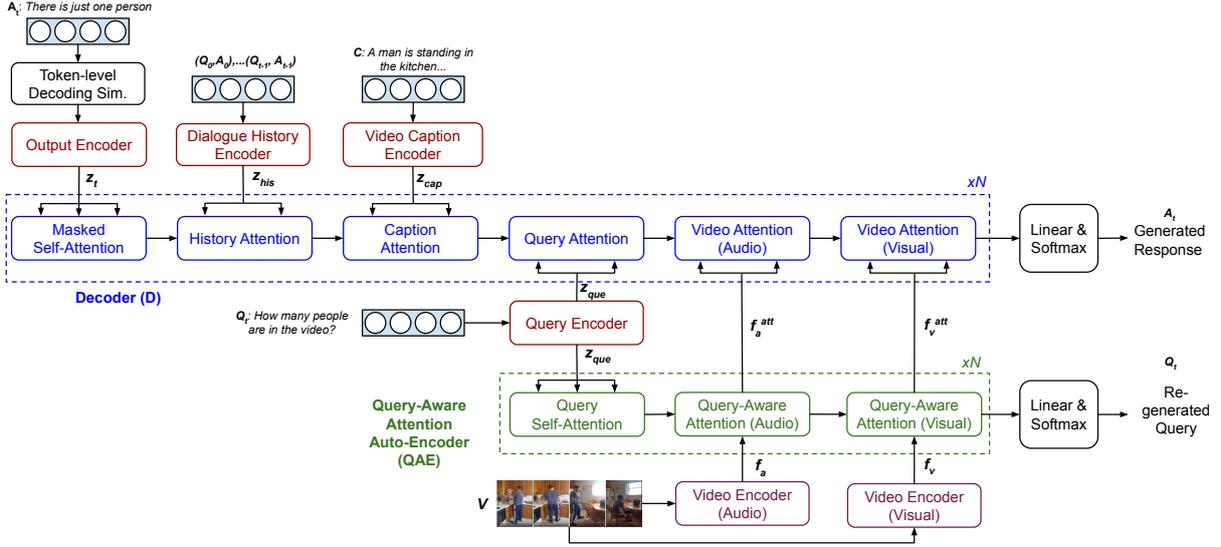


Figure 2: Our MTN architecture includes 3 major components: (i) encoder layers encode text sequences and video features; (ii) decoder layers (D) project target sequence and attend on multiple inputs; and (iii) Query-Aware Auto-Encoder layers (QAE) attend on non-text modalities from query features. For simplicity, Feed Forward, Residual Connection and Layer Normalization layers are not presented. Best viewed in color.

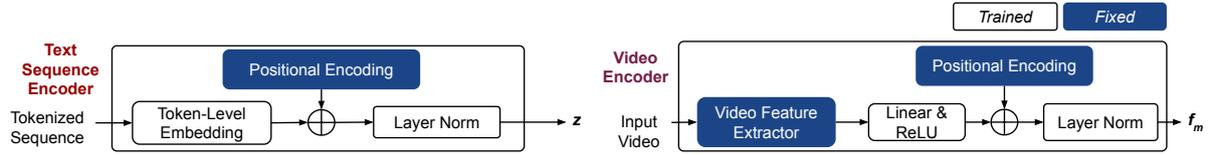


Figure 3: 2 types of encoders are used: text-sequence encoders (left) and video encoders (right). Text-sequence encoders are used on text input, i.e. dialogue history, video caption, query, and output sequence. Video encoders are used on visual and audio features of input video.

sists of 2 linear transformation with ReLU activation in between. We employed residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) around each attention block. The multi-head attention on  $z_s$  is defined as:

$$m_s = \text{Concat}(h_1, \dots, h_h)W^O \quad (1)$$

$$h_i = \text{Attn}(z_{out}^{dec}W_i^Q, z_sW_i^K, z_sW_i^V) \quad (2)$$

$$\text{Attn}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v \quad (3)$$

where  $W_i^Q \in \mathbb{R}^{d \times d_k}$ ,  $W_i^K \in \mathbb{R}^{d \times d_k}$ ,  $W_i^V \in \mathbb{R}^{d \times d_k}$ ,  $W_i^O \in \mathbb{R}^{hd_v \times d}$  (the superscripts of  $s$  and  $t$  are not presented for each  $W$  for simplicity).  $z_{out}^{dec}$  is the output of the previous sub-layer.

The multi-head attention allows the model to attend on text sequence features at different positions of the sequences. By using multi-head attention on visual and audio features, the model can attend on frame sequences to project and extract information from different parts of the video. Using multiple attentions for different input components

also allows the model attend differently on inputs rather than using the same attention network for all. We also experimented with concatenating the input sequences and only use one attention block in each decoding layer, similarly to a Transformer decoder ( See the appendix Section B).

### 3.3 Auto-Encoder Layers

As the multi-head attentions allow dynamic attentions on different input components, the essential interaction between the input query and non-text features of the input video is not fully implemented. While a residual connection is employed and the video attention block is placed at the end of the decoder layer, the attention on video features might not be optimal. We consider adding query-aware attention on video features as a separate component. We design it as a query auto-encoder to allow the model to focus on query-related features of the video in an unsupervised manner. The auto-encoder is composed of a stack of  $N$  layers, each of which includes an query self-attention and

query-aware attention on video features. Hence, the number of sub-layers is  $1 + \|M\|$ . For self-attention, the output of the previous sub-layer  $z_{out}^{ae}$  (or  $z_{que}$  in case of the first auto-encoder stack) is used identically as  $q$ ,  $k$  and  $v$  in Equation 3, while for query-aware attention,  $z_{out}^{ae}$  is used as  $q$  and  $f_m$  is used as  $k$  and  $v$ . For an  $n^{th}$  auto-encoder layer, each output of the query-aware attention on video features  $f_{m,n}^{att}$  is passed to video attention module of the corresponding  $n^{th}$  decoder layer. Each video attention head  $i$  for a given modality  $m$  at decoding layer  $n^{th}$  is defined as:

$$h_i = \text{Attn}(z_{out,n}^{dec} W_i^Q, f_{m,n}^{att} W_i^K, f_{m,n}^{att} W_i^V)$$

The decoder and auto-encoder create a network similar to the One-to-Many setting in (Luong et al., 2015) as the encoded query features are shared between the two modules. We also consider using the auto-encoder as stacked query-aware encoder layers i.e. use query self-attention and query-based attention on video features and extract the output of final layer at  $N^{th}$  block to the decoder. Comparison of the performance (See Table 4 Row C-5 and D) shows that adopting an auto-encoder architecture is more effective in capturing relevant video features.

### 3.4 Generative Network

Similar to sequence generative models (Sutskever et al., 2014; Manning and Eric, 2017), we use a Linear transformation layer with softmax function on the decoder output to predict probabilities of the next token. In the auto-encoder, the same architecture is used to re-generate the query sequence. We separate the weight matrix between the source sequence embedding, output embedding, and the pre-softmax linear transformation.

**Simulated Token-level Decoding.** Different from training, during test time, decoding is still an auto-regressive process where the decoder generates the sentence token-by-token. We aim to simulate this process during training by performing the following procedures:

- Rather than always using the full target sequence of length  $L$ , the token-level decoding simulation will do the following:
- With a probability  $p$ , e.g.  $p = 0.5$  i.e. for 50% of time, crop the target sequence at a uniform-randomly selected position  $i$  where

$i = 2, \dots, (L - 1)$  and keep the left sequence as the target sequence e.g.  $\langle sos \rangle$  there is just one person  $\langle eos \rangle \rightarrow \langle sos \rangle$  there is just one

- As before, the target sequence is offset by one position as input to the decoder

We employ this approach to reduce the mismatch of input to the decoder during training and test time and hence, improve the quality of the generated responses. We only apply this procedure for the target sequences to the decoder but not the query auto-encoder.

## 4 Experiments

### 4.1 Data

We used the dataset from DSTC7 (Yoshino et al., 2018) which consists of multi-modal dialogues grounded on the Charades videos (Sigurdsson et al., 2016). Table 1 summarizes the dataset and Figure 1 shows a training example. We used the audio and visual feature extractors pre-trained on YouTube videos and the Kinetics dataset (Kay et al., 2017) (Refer to (Hori et al., 2018) for the detail video features). Specifically we used the 2048-dimensional I3D\_flow features from the ‘‘Mixed\_5c’’ layer of the I3D network (Carreira and Zisserman, 2017) for visual features and 128-dimensional Audio Set VGGish (Hershey et al., 2017) for audio features. We concatenated the provided caption and summary for each video from the DSTC7 dataset as the default video caption *Cap+Sum*. Other data pre-processing procedures are described in the appendix Section A.1.

	Train	Validation	Test
# of Dialogs	7,659	1,787	1,710
# of Turns	153,180	35,740	13,490
# of Words	1,450,754	339,006	110,252

Table 1: DSTC7 Video Scene-aware Dialogue Dataset

### 4.2 Training

We use the standard objective function log-likelihood of the target sequence  $T$  given the dialogue history  $H$ , user query  $Q$ , video features  $V$ , and video caption  $C$ . The log-likelihood of re-

generated query is also added when QAE is used:

$$\begin{aligned}
 L &= L(T) + L(Q) \\
 &= \sum_m \log P(y_m | y_{m-1}, \dots, y_1, H, Q, V, C) + \\
 &= \sum_n \log P(x_n^q | x_{n-1}^q, \dots, x_1^q, Q, V)
 \end{aligned}$$

We train MTN models in two settings: Base and Large. The Base parameters are  $N = 6, h = 8, d = 512, d_k = d_v = d/h = 64$ , and the Large parameters are  $N = 10, h = 16, d = 1024, d_k = d_v = d/h = 64$ . The probability  $p$  for simulating token-level decoding is 0.5. We trained each model up to 17 epochs. We used the Adam optimizer (Kingma and Ba, 2014). The learning rate is varied over the course of training with strategy adopted similarly in (Vaswani et al., 2017). We used `warmup_steps` as 9660. We employed dropout (Srivastava et al., 2014) of 0.1 at all sub-layers and embeddings. Label Smoothing (Szegedy et al., 2016) is also applied during training. For all models, we select the latest checkpoints that achieve the lowest perplexity on the validation set. We used beam search with beam size 5 and a length penalty 1.0. The maximum output length during inference is 30 tokens. All models were implemented using PyTorch (Paszke et al., 2017)<sup>1</sup>.

### 4.3 Video-Grounded Dialogues

We compared MTN models with the baseline (Hori et al., 2018) and other submission entries to the DSTC7 Track 3. The evaluation includes 4 word-overlapping-based objective measures: BLEU (1 to 4) (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). The results were computed based on one reference ground-truth response per test dialogue in the test set. As can be seen in Table 3, both Base- and Large-MTN models outperform the baseline (Hori et al., 2018) in all metrics. Our Large model outperforms the best previously reported models in the challenge across all the metrics. Even our Base model with smaller parameters outperforms most of the previous results, except for *entry1*, which we outperform in BLEU1-3 and METEOR measures. While some of the submitted models to the

<sup>1</sup>The code is released at <https://github.com/henryhungle/MTN>

challenge utilized external data or ensemble techniques (Alamri et al., 2018), we only use the given training data from the DSTC7 dataset similarly as the baseline (Hori et al., 2018).

#### Impact of Token-level Decoding Simulation.

We consider text-only dialogues (no visual or audio features) to study the impact of the token-level decoding simulation component. We also remove the auto-encoder module i.e. *MTN w/o QAE*. We study the differences of performance when the simulation probability  $p = 0, 0.1, \dots, 1$ . 0 is equivalent to always keeping the target sequences as a whole and 1 is cropping all target sequences at random points during training. As shown in Figure 4, adding the simulation helps to improve the performance in most cases of  $p > 0$  and  $< 1$ . At  $p = 1$ , the performance is suffered as the decoder receives only fragmented sequences during training.

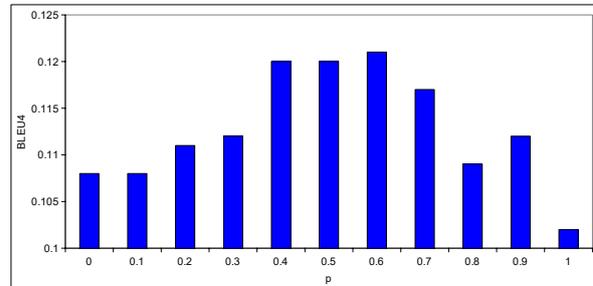


Figure 4: Impact of simulation probability  $p$  in BLEU4 measure on the test data. At  $p = 0.4$  to  $0.6$ , the improvement in BLEU4 scores is more significant.

**Ablation Study.** We tested variants of our models with different combinations of data input in Table 4. With text-only input, compared to our approach (Row B-1), using encoder layers with self-attention blocks (Row A) does not perform well. The self-attention encoders also make it hard to optimize the model as noted by (Liu et al., 2018). When we remove the video caption from the input (hence, no caption attention layers) and use either visual or audio video features, we observe that the proposed auto-encoder with query-aware attention results in better responses. For example, with audio feature, adding the auto-encoder component (Row C-1) increases BLEU4 and CIDEr measures as compared to the case where no auto-encoder is used (Row B-2). When using both caption and video features, the proposed auto-encoder (Row C-5) improves all metrics from the decoder-only model (Row B-4). We also consider using the auto-encoder structure as an encoder (i.e. without the generative component to re-generate query)

and decouple from the decoder stacks (i.e. output of the  $N^{th}$  encoder layer is used as input to the  $1^{st}$  decoder layer) (Row D). The results show that an auto-encoder structure is superior to stacked encoder layers. Our architecture is also better in terms of computation speed as both decoder and auto-encoder are processed in parallel, layer by layer. Results of other model variants are available in the appendix Section B.

#### 4.4 Visual Dialogues

We also test if MTN could generalize to other multi-modal dialogue settings. We experiment on the visually grounded dialogue task with the VisDial dataset (Das et al., 2017a). The training dataset is much larger than DSTC7 dataset with more than 1.2 million training dialogue turns grounded on images from the COCO dataset (Lin et al., 2014). This task aims to select a response from a set of 100 candidates rather than generating a new complete response. Here we still keep the generative component and maximize the log-likelihood of the ground-truth responses during training. During testing, we use the log-likelihood scores to rank the candidates. We also remove the positional encoding component from the encoder to encode image features as these features do not have sequential characteristics. All other components and parameters remain unchanged.

We trained MTN with the Base parameters on the Visual Dialogue v1.0<sup>2</sup> training data and evaluate on the *test-std* v1.0 set. The image features are extracted by a pre-trained object detection model (Refer to the appendix Section A.2 for data preprocessing). We evaluate our model with Normalized Discounted Cumulative Gain (NDCG) score by submitting the predicted ranks of the response candidates to the evaluation server (as the ground-truth for the *test-std* v1.0 split is not published). We keep all the training procedures unchanged from the video-grounded dialogue task. Table 2 shows that our proposed MTN is able to generalize to the visually grounded dialogue setting. It is interesting that our generative model outperforms other retrieval-based approaches in NDCG without any task-specific fine-tuning. There are other submissions with higher NDCG scores from the leaderboard<sup>3</sup> but the approaches of these submis-

<sup>2</sup><https://visualdialog.org/data>

<sup>3</sup><https://evalai.cloudcv.org/web/challenges/challenge-page/103/leaderboard/298>

sions are not clearly detailed to compare with.

Model	NDCG
MTN (Base)	<b>55.33</b>
CorefNMN (Kottur et al., 2018)	54.70
MN (Das et al., 2017a)	47.50
HRE (Das et al., 2017a)	45.46
LF (Das et al., 2017a)	45.31

Table 2: Comparison of MTN (Base) to state-of-the-art visual dialogue models on the *test-std* v1.0. The best measure is highlighted in bold.

## 5 Qualitative Analysis

Figure 6 shows some samples of the predicted test dialogue responses of our model as compared to the baseline (Hori et al., 2018). Our generated responses are more accurate than the baseline to answer human queries. Some of our generated responses are more elaborate e.g. “with a cloth in her hand”. Our responses can correctly describe single actions (e.g. “cleaning the table”, “stays in the same place”) or a series of actions (e.g. “walks over to a closet and takes off her jacket”). This shows that our MTN approach can reason over complex features came from multiple modalities. Figure 5 summarizes the CIDEr measures of the responses generated by our Base model and the baseline (Hori et al., 2018) by their position in dialogue e.g.  $1^{st}$ ... $10^{th}$  turn. It shows that our responses are better across all dialogue turns, from  $1^{st}$  to  $10^{th}$ . Figure 5 also shows that MTN perform better at shorter dialogue lengths e.g. 1-turn, 2-turn and 3-turn, in general and the performance could be further improved for longer dialogues.

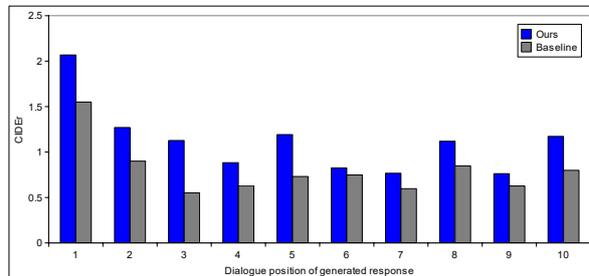


Figure 5: Comparison of CIDEr measures on the test data between MTN (Base) and the baseline (Hori et al., 2018) across different turn position of the generated responses. Our model outperforms the baselines at all dialogue turn positions.

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
<b>MTN</b>							
MTN (Base)	<b>0.357</b>	0.241	0.173	0.128	0.162	0.355	1.249
MTN (Large)	0.356	<b>0.242</b>	<b>0.174</b>	<b>0.135</b>	<b>0.165</b>	<b>0.365</b>	<b>1.366</b>
<b>DSTC7 submissions</b>							
Entry-top1	0.331	0.231	0.171	0.131	0.157	0.363	1.360
Entry-top2	0.329	0.228	0.167	0.126	0.154	0.357	1.306
Entry-top3	0.327	0.225	0.164	0.123	0.155	0.350	1.269
Entry-top4	0.312	0.210	0.152	0.115	0.148	0.357	1.271
Entry-top5	0.329	0.216	0.153	0.114	0.140	0.331	1.103
(Hori et al., 2018)	0.279	0.183	0.13	0.095	0.122	0.303	0.905

Table 3: Evaluated on the test data, the proposed approach achieves better objective measures than the baselines and the submissions to the challenge. The best result in each metric is highlighted in bold.

	CapFea	VidFea	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
<b>MTN w/o QAE + Stacked Self-Attention in Encoder</b>									
A	Cap+Sum	N/A	0.327	0.216	0.154	0.114	0.147	0.332	1.106
<b>MTN w/o QAE</b>									
B-1	Cap+Sum	N/A	0.346	0.231	0.164	0.120	0.158	0.344	1.176
B-2	N/A	A	0.316	0.207	0.145	0.105	0.138	0.315	0.963
B-3	N/A	V	0.328	0.222	0.158	0.118	0.147	0.331	1.102
B-4	Cap+Sum	A+V	0.347	0.234	0.168	0.124	0.158	0.344	1.197
<b>MTN</b>									
C-1	N/A	A	0.324	0.214	0.152	0.113	0.142	0.326	1.031
C-2	N/A	V	0.328	0.223	0.155	0.119	0.147	0.330	1.115
C-3	Cap+Sum	A	0.344	0.236	0.170	0.127	0.159	0.354	1.220
C-4	Cap+Sum	V	0.343	0.229	0.161	0.118	0.160	0.348	1.151
C-5	Cap+Sum	A+V	<b>0.357</b>	<b>0.241</b>	<b>0.173</b>	<b>0.128</b>	<b>0.162</b>	<b>0.355</b>	<b>1.249</b>
<b>MTN (replacing QAE with QE - Query-Aware Encoder)</b>									
D	Cap+Sum	A+V	0.334	0.227	0.164	0.123	0.153	0.344	1.200

Table 4: Ablation analysis of MTN evaluated on the test data. The video features being used is either VGGish for audio features (A) or I3D-Flow for visual features (V). All models are trained with the Base parameters. Best result in each metric is highlighted in bold.

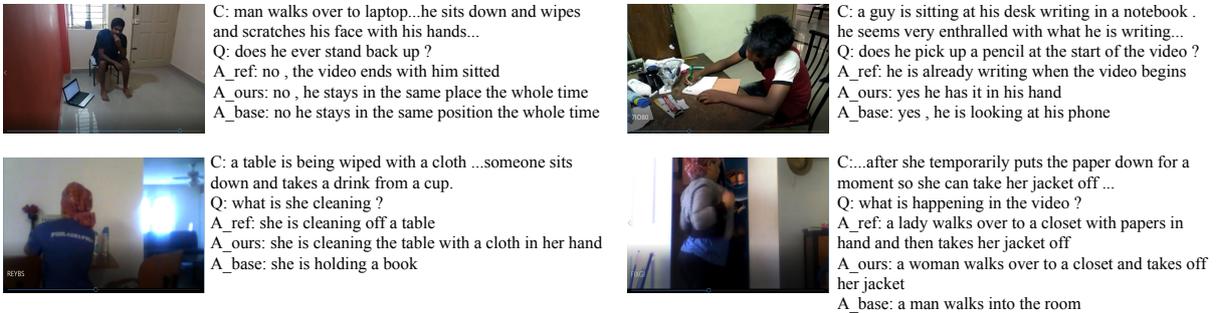


Figure 6: Example test dialogue responses extracted from the ground-truth  $A_{ref}$  and generated by MTN (Base)  $A_{ours}$  and the baseline (Hori et al., 2018)  $A_{base}$ . For simplicity, the dialogue history is not presented and only parts of the video caption  $C$  are shown. Our model provides answers that are more accurate than the baseline, capturing single human action or a series of actions in the videos.

## 6 Conclusion

In this paper, we showed that MTN, a multi-head attention-based neural network, can generate good conversational responses in multimodal settings. Our MTN models outperform the reported baseline and other submission entries to the DSTC7. We also adapted our approach to a visual dialogue task and achieved excellent performance. A possible improvement to our work is adding pre-trained embedding such as BERT (Devlin et al., 2018) or image-grounded word embedding (Kiros et al., 2018) to improve the semantic understanding capability of the models.

## Acknowledgements

The first author is supported by A\*STAR Computing and Information Science scholarship (formerly A\*STAR Graduate scholarship). The third author is supported by the Agency for Science, Technology and Research (A\*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

## References

- Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batra, and Devi Parikh. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7.-. In *DSTC7 at AAAI2019 Workshop*.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Satanjeev Banerjee and Alon Lavie. 2005. 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.
- Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.
- Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.
- Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.
- Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics.
- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Scott Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, page 3.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.
- 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. 2017. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE.

- Chiori Hori, Huda Alamri, Jue Wang, Gordon Winch-ern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv preprint arXiv:1806.08409*.
- Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4203–4212. IEEE.
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natshev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 922–933.
- Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1437–1447.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. [A persona-based neural conversation model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Bing Liu and Ian Lane. 2017. An end-to-end trainable neural network model with belief tracking for task-oriented dialog. *arXiv preprint arXiv:1708.05956*.
- Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.
- Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.
- Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. [Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478. Association for Computational Linguistics.
- Christopher D. Manning and Mihail Eric. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *EACL*.
- Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

- Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. [A hierarchical latent variable encoder-decoder model for generating dialogues](#).
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. [Building end-to-end dialogue systems using generative hierarchical neural network models](#). In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.
- Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. [Neural responding machine for short-text conversation](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Oriol Vinyals and Quoc V. Le. 2015. [A neural conversational model](#). *CoRR*, abs/1506.05869.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.
- Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. [Attention with intention for a neural network conversation model](#). *CoRR*, abs/1510.08565.
- Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan Kummerfeld, Michael Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Sean Gao, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2018. The 7th dialog system technology challenge. *arXiv preprint*.
- Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

## A Data Pre-processing

### A.1 Video-Grounded Dialogues

We split all sequences into (case-insensitive) tokens and selected those in the training data with the frequency more than 1 to build the vocabulary for embeddings. This results in 6175 unique tokens, including the  $\langle eos \rangle$ ,  $\langle sos \rangle$ ,  $\langle pad \rangle$ , and  $\langle unk \rangle$  tokens. Sentences are batched together by approximate sequence lengths, in order of dialogue history length, video caption length, question length, and target sequence length. We use batch size of 32 during training.

### A.2 Visual-Grounded Dialogues

The *test-std* v1.0 set include about 4000 dialogues grounded on COCO-like images collected from Flickr. We only selected tokens that have frequency at least 3 in the training data to build the vocabulary. This results in 13832 unique tokens. We use bottom-up attention features (Anderson et al., 2018) extracted from Faster R-CNN (Ren et al., 2015) which is pre-trained on the Visual Genome data (Krishna et al., 2017). This results in 36 2048-dimensional feature vectors per image.

## B Additional Experiment Results

We experimented our models with text-only input e.g. no video audio or visual features and hence, no auto-encoder layers involved (*MTN w/o QAE*). We tested cases where the maximum dialogue history length  $L_{his}^{max}$  is limited to 1, 2, or 3 turns only. For each case, we also tried to concatenate all the source sequences, including dialogue history, video caption, and query, into a single sequence and use only one multi-head attention block on this concatenated sequence in each decoding layer (Similar to a Transformer decoder). Table 5 summarizes the results. The results show that concatenating the sequences into one affects the quality of the generated responses significantly. When the input sequences are separated and attended differently by different attention modules, the results improve. This could be explained as different sequences contain different signals to generate responses e.g. dialogue history contains information of references or ellipses in the user queries, user queries include direct signals for feature attention in input videos. Another observation is using all possible dialogue turns in the dialogue history i.e.  $L_{his}^{max} = 10$  achieves the best results. We did not conduct experiments of concatenating source sequences with  $L_{his}^{max} = 10$  due to memory issues with large input sequences.

Max. HisLen	Concat. Source Sequence?	BLEU4	ROUGE-L	CIDEr
10	No	<b>0.120</b>	<b>0.344</b>	<b>1.176</b>
3	No	0.116	0.343	1.141
3	Yes	0.097	0.308	0.924
2	No	0.115	0.343	1.150
2	Yes	0.090	0.304	0.900
1	No	0.119	0.343	1.163
1	Yes	0.095	0.301	0.894

Table 5: Evaluation results on the test set for *MTN w/o QAE* models in which maximum history length is range from 1 to 3 or 10 (i.e. all dialogue turns possible). We also experiments when all the source sequences are concatenated into one and the decoder only has one attention block on the concatenated sequence. The auto-encoder components are also removed. Best result in each metric is highlighted in bold.