# Video Dialog via Progressive Inference and Cross-Transformer

**Weike Jin**[1], **Zhou Zhao**[1]*, **Mao Gu**[1], **Jun Xiao**[1], **Furu Wei**[2], **Yueting Zhuang**[1]
[1]Zhejiang University, Hangzhou
[2]Microsoft Research Asia, Beijing
{weikejin,zhaozhou,21821134,junx,yzhuang}@zju.edu.cn
fuwei@microsoft.com

## Abstract

Video dialog is a new and challenging task, which requires the agent to answer questions combining video information with dialog history. And different from single-turn video question answering, the additional dialog history is important for video dialog, which often includes contextual information for the question. Existing visual dialog methods mainly use RNN to encode the dialog history as a single vector representation, which might be rough and straightforward. Some more advanced methods utilize hierarchical structure, attention and memory mechanisms, which still lack an explicit reasoning process. In this paper, we introduce a novel progressive inference mechanism for video dialog, which progressively updates query information based on dialog history and video content until the agent think the information is sufficient and unambiguous. In order to tackle the multi-modal fusion problem, we propose a cross-transformer module, which could learn more fine-grained and comprehensive interactions both inside and between the modalities. And besides answer generation, we also consider question generation, which is more challenging but significant for a complete video dialog system. We evaluate our method on two large-scale datasets, and the extensive experiments show the effectiveness of our method.

## 1 Introduction

Visual dialog can be seen as an extension of the visual question answering (VQA) (Antol et al., 2015; Yu et al., 2017). Unlike visual question answering, in which each question is asked independently, visual dialog requires the agent to answer multi-round questions and the previous round of question-answer pairs form the dialog history. Currently, most of the existing visual dialog approaches mainly focus on image dialog (Das et al., 2017a; Seo et al., 2017; Kottur et al., 2018; Lee et al., 2018). Video dialog is still less explored. For video dialog, given a video, a dialog history and a question about the video content, the agent has to combine video information with context information from dialog history to infer the answer accurately. And due to the complexity of video information, it's more challenging than image dialog.

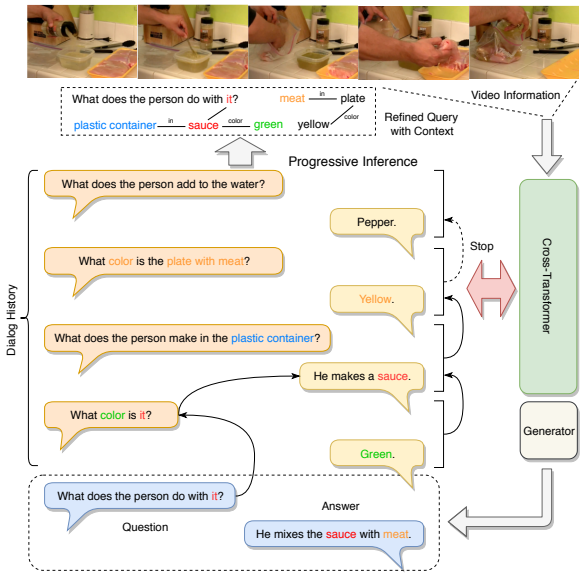



Figure 1: An illustration of our video dialog model with progressive inference and cross-transformer.

Obviously, there are two major problems for video dialog: (1) How to obtain valuable information from the dialog history. We find that the content of the dialogue usually has a certain continuity, and the subsequent dialogue is likely to continue what was discussed before. Thus, it's important to infer valuable information from the dialog history to help answer the question. Sometimes, the question can be ambiguous without such information. For instance, as illustrated in Figure 1, the question "What does the person do with

*Zhou Zhao is the corresponding author.



*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2109–2118, Hong Kong, China, November 3–7, 2019. 

it?" contains a pronoun 'it', which brings ambiguity to the question. We can't figure out what 'it' really refers to, if only rely on the question information. There is no more key information in the question, thus it's also hard to get help from video content. In such a situation, the dialog history is necessary. The existing visual dialog methods (Das et al., 2017a) mainly use recurrent neural network (like LSTM) to encode the dialog history as a single vector representation, which we think might be a bit rough and straightforward. Some more advanced methods (Seo et al., 2017; Zhao et al., 2018) utilize hierarchical structure, attention and memory mechanisms to refine the dialog history representation, which still lacks an explicit reasoning process. Recently, Kottur *et al*. (2018) propose a neural module network architecture including two novel modules, which perform coreference resolution at a word level. However, their work is based on the image, which means they only take static visual characters into consideration, like objects and attributes. As for video dialog, there are additional dynamic characters in the video, such as action and state transition. Thus, we employ multi-stream video information in our model. And due to the continuity of the dialog history, we attempt to design a reverse-order progressive reasoning process to extract useful information and eliminate ambiguity. (2) How to answer the question related to the video content. This is also the key problem in video question answering. For instance, a mainstream method is to utilize recurrent neural network to learn the hidden states of video sequences, due to the temporal structure. Then, the encoded query information is used to select relevant information from the hidden states, through attention mechanism or memory network. Finally, the query information is combined with the refined video information by multi-modal fusion to generate the answer.

In this paper, in order to make better use of the dialog history and video information, we propose a progressive inference mechanism for video dialog, which progressively updates query information based on the dialog history and video content, as shown in Figure 1. And the progressive inference will stop when the agent thinks the query information is sufficient and unambiguous, or it has arrived at the beginning of the dialog history. For the query information is a sequence of vector representations which contains key information of the question, dialog history and video content, we propose a cross-transformer module to learn a stronger joint representation of them. And in addition to the generation of answers, we also attempt to generate questions according to the dialog history, which is more challenging but significant for a complete video dialog system. The generator we use is an extension of the Transformer (Vaswani et al., 2017) decoder, in order to process multi-stream inputs. And the main contributions of this paper are as follows:

- Unlike previous studies, we propose a novel progressive inference mechanism for video dialog, which progressively updates query information based on the dialog history and video content.
- We design a cross-transformer module to tackle the multi-modal fusion problem, which could learn more fine-grained and comprehensive interactions both inside and between the visual and textual information.
- Besides the answer generation, we also realize question generation under the same framework to construct a complete video dialog system.
- Our method achieves the state-of-the-art performance on two large-scale datasets.

## 2 Related Work

In this section, we briefly review some related work of visual dialog, including image dialog and video dialog.

As first proposed in (Das et al., 2017a), visual dialog requires the agent to predict the answer of a given question based on an image and dialog history. While dialog system (Serban et al., 2016, 2017) has been widely explored, visual dialog is still a young task. Until recently, some approaches are proposed. Das *et al.* (2017a) propose three models based on late fusion, attentional hierarchical LSTM and memory networks respectively. They also propose the VisDial dataset by pairing two subjects on Amazon Mechanical Turk to chat about an image. Lu *et al.* (2017) propose a generator-discriminator architecture where the outputs of the generator are updated using a perceptual loss from a pre-trained discriminator. De Vries *et al.* (2017) propose a GuessWhat game style dataset, on which one person

asks questions about an image to guess which object has been selected and the second person answers questions. Das *et al.* (Das et al., 2017b) utilize deep reinforcement learning to learn the policies of a 'Questioner-Bot' and an 'Answerer-Bot', based on the goal of selecting the right images that the two agents are talking. Seo *et al.* (2017) resolve visual references in the question based on a new attention mechanism with an attention memory, and the model indirectly resolves coreferences of expressions through the attention retrieval process. Jain *et al.* (2018) propose a simple symmetric discriminative baseline, which can be applied to both predicting an answer as well as predicting a question. Kottur *et al.* (2018) propose an introspective reasoning about visual coreferences, which links coreferences and grounds them in the image at a word-level, rather than at a sentence-level as in prior visual dialog work. Massiceti *et al.* (2018) propose FLIPDIAL model, a generative convolutional model for visual dialogue which is able to generate answers as well as generate questions based on a visual context. Lee *et al.* (2018) propose a practical goal-oriented dialog framework using information-theoretic approach, in which the questioner figures out the answerer's intention via selecting a plausible question by calculating the information gain of the candidate intentions and possible answers of each question. Wu *et al.* (2018) propose a sequential co-attention generative model that can jointly learn the image, dialog history information with question, and a discriminator which can dynamically access to the attention memories with an intermediate reward.

The aforementioned work mainly focuses on the image dialog. As for the task of video dialog, it's still less explored. One similar work is proposed by Zhao *et al.* (2018). They study the problem of multi-turn video question answering by employing a hierarchical attention context learning method with recurrent neural networks for context-aware question understanding and a multi-stream attention network that learns the joint video representation. They also propose two large-scale multi-turn video question answering datasets, which are employed in our experiments. And recently, Hori *et al.* (2018) propose a model that incorporates technologies for multimodal attention-based video description into an end-to-end dialog system. Pasunuru *et al.* (2018) propose a new game-chat based video-context, many-speaker dialogue task. In this work, we utilize a more explicit progressive inference mechanism and a novel cross-transformer module to generate both answers and questions for video dialog.

## 3 Our Approach

### 3.1 Problem Formulation

Before introducing our method, we first introduce some basic notions. We denote the video by $\mathbf{v} \in V$, the dialog history by $\mathbf{c} \in C$, the new question by $\mathbf{q} \in Q$ and the corresponding answer by $\mathbf{a} \in A$, respectively. For video is a sequence of frames, the frame-level representation for video $\mathbf{v}$ is denoted by $\mathbf{v}^f = (v_1^f, v_2^f, \ldots, v_{T_1}^f)$, where $T_1$ is the number of frames in video $\mathbf{v}$. And besides the frame-level representation, we also employ the segment-level representation to bring more information of video, which is given by $\mathbf{v}^s = (v_1^s, v_2^s, \ldots, v_{T_2}^s)$, where $T_2$ is the number of segments and $v_j^s$ is the vector representation of the $j$-th segment. The dialog history $\mathbf{c} \in C$ is given by $\mathbf{c} = (c_1, c_2, \ldots, c_N)$, where $c_i$ is the $i$-th round of dialog, which consists of a question $q_i$ and an answer $a_i$. Using these notations, the task of video dialog could be formulated as follows: given a set of video $V$ and the associated dialog history $C$, the goal of video dialog is to train a model that learns to generate human-like answers when a new question about the visual content is asked. Similar to the video question answering task, there are two kinds of methods to produce the answer, generative and discriminative. For generative type, a word sequence generator (normally a RNN) is employed to fit the ground truth answer sequences. As for discriminative type, an additional candidate answer vocabulary is provided and the problem is reformulated as a multi-class classification problem. In this paper, we try to tackle the generative version of video dialog, for it is more challenging than discriminative version and is more valuable in the practical system.

### 3.2 Progressive Inference

We first introduce our dialog encoder with progressive inference mechanism. Figure 2 shows the overview of the dialog encoder, in which the inference process has been expanded for intuitive understanding. As shown in the figure, the inputs of this module are a sequence of dialog history $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ and the query $x$. Specifically, the query $x$ could be different ac-

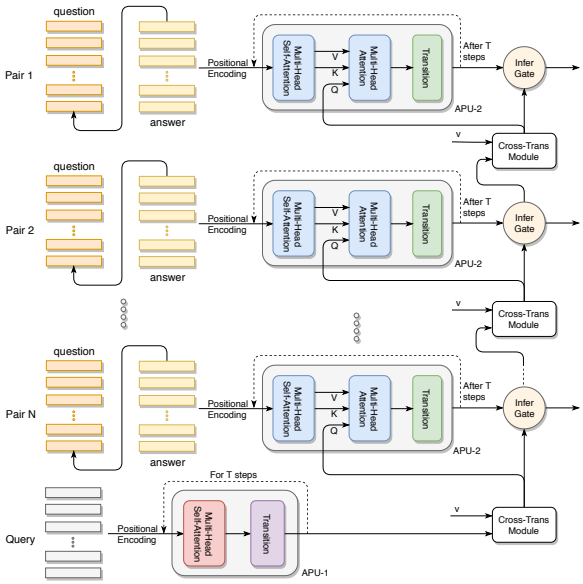



Figure 2: The overview of the dialog encoder with progressive inference.

cording to different purposes. For answer generation, the query $x$ is the question $q$ waiting to be answered. As for question generation, we combine the latest round of dialog $c_N$ in dialog history with ground-truth answer $a$ of the question as the query $x$, considering the topic continuity between the dialog. And it's intuitive that the new round of dialog is more likely to be relevant to the latest dialog history, especially in video dialog, in which the dialog focuses on a specific scenario not chit-chat. Each round of dialog consists of a question $q_i = (w^q_{i1}, w^q_{i2}, \ldots, w^q_{il})$ and an answer $a_i = (w^a_{i1}, w^a_{i2}, \ldots, w^a_{il'})$, where $w^q, w^a$ are word embeddings and $l, l'$ are corresponding sentence length. And for the i-th round dialog $c_i$, instead of using two LSTM-like neural networks to learn sentence-level representations individually, we concatenate the question $q_i$ and the answer $a_i$ end to end, given by $c_i = (w^q_{i1}, w^q_{i2}, \ldots, w^q_{il}, w^a_{i1}, w^a_{i2}, \ldots, w^a_{il'})$, and then employ self-attention mechanism to extract more fine-grained word-level interactions between the question-answer pair. And the query $x$ is also encoded by a similar self-attention mechanism. Specifically, the attention processing units (APU) we employ is based on the Transformer model (Vaswani et al., 2017), which has achieved great success in natural language processing. The first type of attention processing unit (APU-1) is as same as the encoder module of the Transformer model. It consists of a multi-head self-attention layer and a transition layer. And it has to be noted that we do not show residual connections and layer normalization in the figure for conciseness. Here, the transition layer is a fully-connected neural network that consists of two linear transformations with a rectified-linear activation in between. The second type of attention processing unit (APU-2) is similar to the decoder module of the Transformer model. However, there is a difference in the middle multi-head attention layer. We assume the outside input is $I_o$ and the output of the multi-head self-attention layer is $O_a$, then the normal operation in the Transformer decoder is given by

$$\text{Attention}(O_a, I_o, I_o) = \text{softmax}(\frac{O_a I_o^\top}{\sqrt{d}})I_o \quad (1)$$

where $d$ is the dimension of sequence elements. In our method, the order of the inputs are replaced, which is given by

$$\text{Attention}(I_o, O_a, O_a) = \text{softmax}(\frac{I_o O_a^\top}{\sqrt{d}})O_a \quad (2)$$

The reason of this replacement is that we want the outside input information to guide the inner attention operation, not the inside information. Under the video dialog scenario, we expect to use the query information to filter related and helpful information from the dialog history.

After introducing the basic modules of our dialog encoder, now we describe the progressive inference mechanism, as shown in Algorithm 1. Firstly, we add its position encoding to the input query $x$ to bring order information, which is similar to (Vaswani et al., 2017). Then, we employ the first type of attention processing unit (APU-1) to encode the initial query information and the encoded query is denoted as $q$. After that, we progressively update query information from each round of dialog. Due to the continuity of the dialog history, the order of the inference is from the latest round to the beginning round. For the i-th round, firstly, Cross-Transformer module is used to capture both query-aware video information and video-aware query information, and they are merged into the updated queries $\{O_{fq}, O_{sq}\}$, due to different visual features. Then, the information of current round of dialog history $c_i$ is encoded and filtered by the second type of attention processing unit (APU-2), using the latest updated query. Through this way, we can obtain the query related information $\hat{c}_i$ from current dialog, which

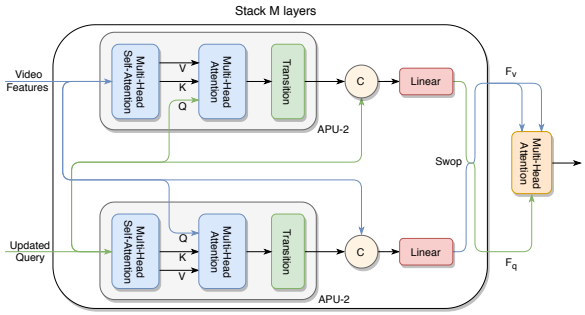



Figure 3: The overview of the cross-transformer module for multi-modal fusion.

could complete the query information and eliminate ambiguity. Finally, we utilize an inference gate to update the query information again and determine whether it is needed to stop reasoning, given by

$$S = \sigma([\hat{c}_i^1; O_{fq}]W_1 + b_1), \quad (3)$$
$$S' = \sigma([\hat{c}_i^1; O_{fq}]W_2 + b_2), \quad (4)$$
$$O_{fq} = S' \odot \hat{c}_i^1 + S \odot O_{fq}, \quad (5)$$
$$\mu_1 = \sigma(W_3(O_{fq}W_4 + b_3) + b_4) \quad (6)$$

where $\sigma$ is the sigmoid function, $[;]$ is concatenation operation, $\odot$ is element-wise multiplication, $W_1, W_2, W_3, W_4$ are weight matrixes, $b_1, b_2, b_3, b_4$ are biases, $S, S'$ are score vectors of the gate and $\mu_1$ (or $\mu_2$) is a scalar representing the information score. If $\mu = (\mu_1 + \mu_2)/2$ is greater than a threshold $\tau$, the inference gate will output current $\{O_{fq}, O_{sq}\}$ as the final output of the encoder and stop the inference, which means the agent think that the current query information is sufficient and unambiguous for answer or question generation. And the inference will also stop when it has arrived at the beginning of the dialog history.

### 3.3 Cross-Transformer Module

In this section, we introduce the cross-transformer (CT) module for multi-modal fusion, which could learn more fine-grained and comprehensive interactions both inside and between the input modalities. As shown in Figure 3, there are two input channels, one for video information and the other for query information. In our method, we utilize two kinds of video information, the frame-level information $v^f$ and the segment-level information $v^s$. Here, we take $v^f$ as an example, and the process of $v^s$ is the same. There are two attention process units (APU-2) in our CT module. And instead of encoding these two channels individually, we design a cross connection between the APU-2 to learn the interactions both inside and between different modalities. Specifically, we use the updated query $O_{fq}$ to guide the middle attention layer of the video channel APU-2, in order to learn the query-aware video information $v_q^f$. And the frame-level information $v^f$ is also utilized to guide the same attention layer of the query channel APU-2, for the purpose of learning frame-aware query information $q_f^u$. Then, we further fuse the information of each output of the APU-2 with its input guidance by a concatenation and a linear layer, given by

$$F_v = \text{Linear}([v^f : q_f^u]) \quad (7)$$
$$F_q = \text{Linear}([v_q^f : O_{fq}]) \quad (8)$$

Before outputting, we swop the output streams again to restore the original input order for next layer process. By stacking M layers, we expect to enhance the fusion effect. Finally, we employ another multi-head attention layer to fuse $F_v$ and $F_q$, and the output is denoted as $O_{fq}$. For segment-level video information $v^s$, we also could get the output $O_{sq}$.

### 3.4 Answer & Question Decoder

The decoder we use is shown in Figure 4. In this decoder, the third type of attention process unit (APU-3) is employed, which is as same as the Transformer decoder. And because there is

**Algorithm 1** Progressive Inference

**Input**: $c, x, v^f, v^s, N$
**Parameter**: $\hat{c}, q, i, \mu, \mu_1, \mu_2$
**Output**: $O_{fq}, O_{sq}$

Let $i = N, \mu = 0$
$q \leftarrow \text{APU\_1}(x + \text{position\_encoding})$
$O_{fq} \leftarrow O_{sq} \leftarrow q$
**while** $\mu < \tau$ and $i > 0$ **do**
  $O_{fq} \leftarrow \text{Cross\_Trans}(O_{fq}, v^f)$
  $O_{sq} \leftarrow \text{Cross\_Trans}(O_{sq}, v^s)$
  $\hat{c}_i^1 \leftarrow \text{APU\_2}(c_i, O_{fq})$
  $\hat{c}_i^2 \leftarrow \text{APU\_2}(c_i, O_{sq})$
  $\mu_1, O_{fq} \leftarrow \text{Infer\_Gate}(\hat{c}_i^1, O_{fq})$
  $\mu_2, O_{sq} \leftarrow \text{Infer\_Gate}(\hat{c}_i^2, O_{sq})$
  $\mu = (\mu_1 + \mu_2)/2$
  $i = i - 1.$
**end while**
**return** $O_{fq}, O_{sq}$

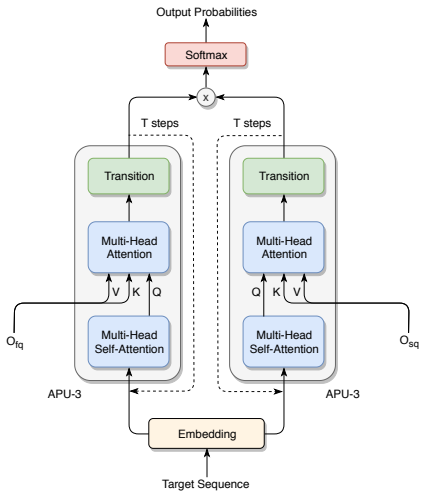



Figure 4: The overview of the answer and question decoder.

two outputs of the dialog encoder, $O_{fq}$ and $O_{sq}$, we utilize two individual attention process units (APU-3) to decode their information. Thus, we could obtain two different answer (or question) distributions based on the appearance and motion information correspondingly. Then, this two distributions are multiplied elemently and a softmax layer is employed to generate the final output probabilities. We use teacher-forcing strategy during training stage, and at generation time the answer (or question) is generated autoregressively.

## 4 Experiments

### 4.1 Setup

**Dataset.** The datasets that we use are proposed in (Zhao et al., 2018), which are based on YouTube-Clips (Chen and Dolan, 2011) dataset and TACoS-MultiLevel (Rohrbach et al., 2014) dataset. These two datasets consist of 1,987 and 1,303 videos individually. Each video in YouTubeClips is composed of 60 frames, and as for TACoS-MultiLevel, the number of frames is 80. Each video has five different dialogs. There are 6515 video dialogs in YouTubeClips dataset and 9935 video dialogs in TACoS-MultiLevel dataset. The numbers of question-answer pairs are 66,806 and 37,228 correspondingly. Statistically, there are five rounds of conversation in most of the video dialogs for TACoS-MultiLevel dataset and it is mostly between three and twelve rounds for YouTubeClips dataset. The percentages of training data, validation data and testing data for both datasets are 90%, 5%, and 5% respectively, according to the number of constructed video dialogs.

**Implementation.** Firstly, we use the pre-trained Glove (Pennington et al., 2014) model to obtain the word embeddings of the dialog. The dimension of the word vector is 512. Then, we employ the pre-trained VGGNet to learn appearance feature of each frame. The motion features of video are extracted by the pre-trained 3D-ConvNet (Tran et al., 2015). And we utilize transition layers (Section 3.2) to transform both appearance and motion features into the same dimension as word vectors to ease later process. We set the circulation steps T to 5 and the stack layers M to 4 after doing a lot of attempts. The threshold $\tau$ in the progressive inference is also important. It will influence the effect of the progressive inference process. We find that if the threshold is too small, the inference process will stop early and miss some important information. On the contrary, if the threshold is too large, it will stop too late leading to more interference information and time consuming. We did a lot of experiments and adjustments to get a balanced threshold of 0.85, and this value might be a bit different in different scenarios. For the training stage, we use Adam optimizer with the initial learning rate of 0.0005, and we also adopt a warm-up strategy to improve the effectiveness of network learning.

**Evaluation Metrics.** In this paper, we employ several metrics to evaluate the quality of the generated answers and questions. They are BLEU-N (N=1,2), ROUGE-L and METEOR, which have been widely used in natural language generation tasks.

### 4.2 Comparisons and Analysis

Because video dialog is still less explored especially for generative results, we extend some existing image dialog and video question answering models as baseline models for video dialog, which are introduced in the following:

- **ESA+, STAN+ and CDMN+** methods extend three video QA models respectively, which are ESA model (Zeng et al., 2017), STAN (Zhao et al., 2017) and CDMN model (Gao et al., 2018). A hierarchical LSTM network is added to model the dialog history.

- **LF+, HRE+ and MN+** methods extend three image dialog models (Das et al., 2017a) by

Table 1: Experimental results of answer generation on TACoS-MultiLevel and YoutubeClip datasets.

| Method | TACoS-MultiLevel | | | | YoutubeClip | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | ROUGE | METEOR | BLEU-1 | BLEU-2 | ROUGE | METEOR |
| ESA+ | 0.356 | 0.244 | 0.422 | 0.109 | 0.268 | 0.151 | 0.276 | 0.082 |
| STAN+ | 0.408 | 0.312 | 0.449 | 0.133 | 0.315 | 0.185 | 0.306 | 0.090 |
| CDMN+ | 0.429 | 0.341 | 0.460 | 0.142 | 0.293 | 0.161 | 0.311 | 0.094 |
| LF+ | 0.404 | 0.290 | 0.465 | 0.135 | 0.284 | 0.183 | 0.307 | 0.083 |
| HRE+ | 0.438 | 0.320 | 0.502 | 0.153 | 0.293 | 0.172 | 0.308 | 0.094 |
| MN+ | 0.430 | 0.326 | 0.472 | 0.149 | 0.306 | 0.185 | 0.290 | 0.086 |
| SFQIH+ | 0.438 | 0.334 | 0.481 | 0.153 | 0.326 | **0.202** | 0.319 | 0.085 |
| HACRN | 0.451 | 0.346 | 0.499 | 0.161 | 0.307 | 0.174 | 0.331 | 0.104 |
| RICT (ours) | **0.464** | **0.361** | **0.527** | **0.178** | **0.333** | 0.194 | **0.332** | **0.104** |

Table 2: Experimental results of question generation on TACoS-MultiLevel and YoutubeClip datasets.

| Method | TACoS-MultiLevel | | | | YoutubeClip | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU-1 | BLEU-2 | ROUGE | METEOR | BLEU-1 | BLEU-2 | ROUGE | METEOR |
| ESA+ | 0.693 | 0.582 | 0.718 | 0.341 | 0.497 | 0.333 | 0.565 | 0.212 |
| STAN+ | 0.706 | 0.599 | 0.730 | 0.354 | 0.483 | 0.322 | 0.559 | 0.208 |
| CDMN+ | 0.707 | 0.603 | 0.740 | 0.357 | 0.507 | 0.341 | 0.567 | 0.219 |
| LF+ | 0.704 | 0.598 | 0.728 | 0.349 | 0.512 | 0.346 | 0.574 | 0.218 |
| HRE+ | 0.694 | 0.592 | 0.729 | 0.348 | 0.515 | 0.350 | 0.571 | 0.223 |
| MN+ | 0.698 | 0.589 | 0.718 | 0.345 | 0.488 | 0.324 | 0.556 | 0.204 |
| SFQIH+ | 0.694 | 0.592 | 0.729 | 0.349 | 0.503 | 0.339 | 0.563 | 0.217 |
| HACRN | 0.715 | 0.616 | 0.741 | 0.358 | 0.524 | 0.352 | 0.577 | 0.229 |
| RICT (ours) | **0.733** | **0.625** | **0.748** | **0.367** | **0.536** | **0.375** | **0.593** | **0.234** |

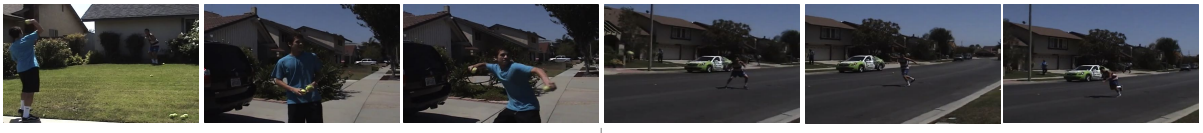
utilizing a LSTM network to encode the video information, which are based on late fusion, attention based hierarchical LSTM, and memory networks respectively.

- **SFQIH+** method extends SF-QIH-se model (Jain et al., 2018) by employing a LSTM network to encode the video information, which concatenates all of the input embeddings for each of the possible answer options.

Besides the baseline models, we also compare our method with the HACRN model (Zhao et al., 2018). They propose a similar work called multi-turn video question answering. It uses LSTM and attention mechanism to encode the dialog history and question to get a joint question representation. They combine this joint representation with video features by a multi-stream attention network. And a multi-step reasoning strategy is applied to enhance the reasoning ability. For the above models, a normal LSTM recurrent sentence generator is employed to generate the answer and question.

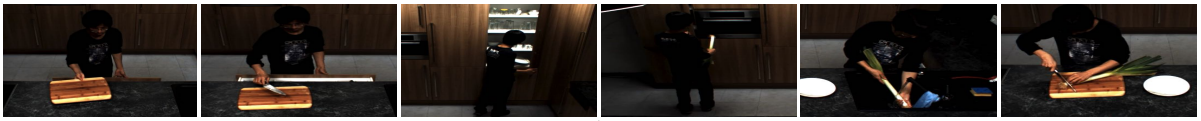
Table 1 shows the experimental results of answer generation on TACoS-MultiLevel and YoutubeClip datasets, and Table 2 shows the question generation results on same datasets. Our method (RICT) outperforms all above models in almost all metrics. This fact shows the effectiveness of our overall network architecture. And we find that the image dialog models perform better than video QA models in answer generation, but worse in question generation on both datasets. This might indicate that for these two datasets, the answer generation is more dependent on dialog, and question generation is more dependent on video content.

We perform an ablation study to evaluate the impact of each component in our model. The experimental results are listed in Table 3, where $\text{RICT}_{(lstm)}$, $\text{RICT}_{(wo.ct)}$, $\text{RICT}_{(wo.pi)}$ represent our model with a LSTM sentence generator, replacing cross-transformer module and replacing progressive inference with baseline modules correspondingly. The results on YoutubeClip dataset is similar, which it's not shown in the paper. The fact that the variants perform worse than full RICT model but still better than baseline models that



| questions | answers | μ-scores | Input: What does the boy in striped T-shirt see before running away? |
|---|---|---|---|
| Where are the two boys? | On the lawn. | - | Ground truth: Tennis ball. |
| What is the boy in blue holding? | Tennis ball. | 0.89 | HACRN: Automobile. |
| Who is the boy in blue throwing balls at? | A boy in striped T-shirt. | 0.73 | Ours: Tennis ball. |

(a)



| questions | answers | μ-scores | Input: Where did the person wash it briefly? In the sink. Ans: On the cutting board. |
|---|---|---|---|
| What did the person take from the drawer? | A cutting board. | 0.90 | Ground truth: Where did the person cut the roots of it? |
| What did the person also take out? | A knife. | 0.67 | HACRN: What did the person do? |
| What did the person retrieve from the cabinet? | One long leek. | 0.55 | Ours: Where did the person cut it? |

(b)

Figure 5: Visualization examples of experimental results. (a) is an answer generation result on YoutubeClip dataset, and (b) is a question generation result on TACoS-MultiLevel dataset.

Table 3: Ablation study results on TACoS-MultiLevel dataset. The upper part is the results of answer generation, and the lower part is the results of question generation.

| Method | BLEU-1 | ROUGE | METEOR |
|---|---|---|---|
| RICT$_{(lstm)}$ | 0.456 | 0.508 | 0.167 |
| RICT$_{(wo.ct)}$ | 0.449 | 0.497 | 0.159 |
| RICT$_{(wo.pi)}$ | 0.433 | 0.475 | 0.151 |
| RICT (ours) | **0.464** | **0.527** | **0.178** |
| RICT$_{(lstm)}$ | 0.724 | 0.745 | 0.362 |
| RICT$_{(wo.ct)}$ | 0.705 | 0.733 | 0.352 |
| RICT$_{(wo.pi)}$ | 0.711 | 0.740 | 0.357 |
| RICT (ours) | **0.733** | **0.748** | **0.367** |

have the similar modules proves the effectiveness of each part of our model.

We also show some visualization examples of experimental results in Figure 5. For the example (a), the progressive inference successfully stops at the second round of dialog, for the current $\mu$-score has exceeded the threshold and the former round of dialog can't bring more valuable information. The example (b) also shows a similar result on question generation, however, this time the inference stops at the first round for it still contains the related thing 'cutting board'. Part of the words with high attention by our model are shown in different colors. And both of the generated answer and question of our model in the examples are better than the compared model, which are much closer to the ground-truth.

## 5 Conclusion

In this paper, in order to have a better understanding of both dialog history and video contents, we propose a novel progressive inference mechanism for video dialog, which progressively updates query information until it is sufficient and unambiguous, or it has arrived at the beginning of the dialog history. We also design a cross-transformer module to tackle multi-modal fusion problem, which could learn more fine-grained interactions between the visual and textual information. And in addition to answer generation, we also consider question generation based on the dialog history and video content under the same framework, which is more challenging but significant for a complete video dialog system. The qualitative and quantitative experimental results on two large-scale video dialog datasets show the effectiveness of our method.

## Acknowledgments

This work was supported by Zhejiang Natural Science Foundation (LR19F020002, LZ17F020001), National Natural Science Foundation of China (61976185, 61572431), the Fundamental Research Funds for the Central Universities, Chinese Knowledge Center for Engineering Sciences and Technology and Microsoft Research Asia.

## References

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.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 190–200.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1080–1089.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision*, pages 2970–2979.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475.

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

Chiori Hori, Huda Alamri, Jue Wang, Gordon Winchern, 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*.

Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. 2018. Two can play this game: visual dialog with discriminative question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763.

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.

Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2018. Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. In *Advances in Neural Information Processing Systems*, pages 2580–2590.

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324.

Daniela Massiceti, N Siddharth, Puneet K Dokania, and Philip HS Torr. 2018. Flipdial: A generative model for two-way visual dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6097–6105.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

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

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6106–6115.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 8.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *AAAI*, pages 4334–4340.

Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu. 2018. Multi-turn video question answering via multi-stream hierarchical attention context network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3690–3696.

Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video question answering via hierarchical spatio-temporal attention networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*.