BiST: Bi-directional Spatio-Temporal Reasoning for Video-Grounded Dialogues

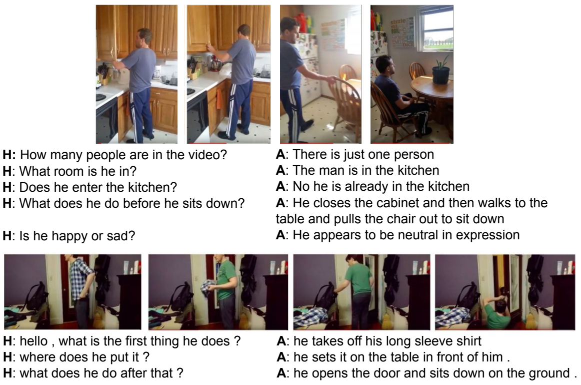
Hung Le[†§*], **Doyen Sahoo**[‡], **Nancy F. Chen**[§], **Steven C.H. Hoi**[†‡]
[†] Singapore Management University
hungle.2018@smu.edu.sg
[‡] Salesforce Research Asia
{dsahoo,shoi}@salesforce.com
[§]Institute for Infocomm Research, A*STAR
nfychen@i2r.a-star.edu.sg

Abstract

Video-grounded dialogues are very challenging due to (i) the complexity of videos which contain both spatial and temporal variations, and (ii) the complexity of user utterances which query different segments and/or different objects in videos over multiple dialogue turns. However, existing approaches to video-grounded dialogues often focus on superficial temporal-level visual cues, but neglect more fine-grained spatial signals from videos. To address this drawback, we propose Bi-directional Spatio-Temporal Learning (BiST), a vision-language neural framework for high-resolution queries in videos based on textual cues. Specifically, our approach not only exploits both spatial and temporal-level information, but also learns dynamic information diffusion between the two feature spaces through spatial-to-temporal and temporal-to-spatial reasoning. The bidirectional strategy aims to tackle the evolving semantics of user queries in the dialogue setting. The retrieved visual cues are used as contextual information to construct relevant responses to the users. Our empirical results and comprehensive qualitative analysis show that BiST achieves competitive performance and generates reasonable responses on a large-scale AVSD benchmark. We also adapt our BiST models to the Video QA setting, and substantially outperform prior approaches on the TGIF-QA benchmark.

1 Introduction

A video-grounded dialogue agent aims to converse with humans not only based on signals from natural language but also from other modalities such as sound and vision of the input video. Recent efforts (Alamri et al., 2018; Sanabria et al., 2019; Alamri et al., 2019) consider video-grounded dialogues as an extension of video Question-Answering (QA) (Tapaswi et al., 2016; Jang et al., 2017; Lei et al., 2018) whereby the agent answers questions from humans over multiple turns rather than a single turn (See Figure 1). This is a very complex task as the dialogue agent needs to possess not only strong language understanding to generate natural responses but also sophisticated reasoning over video information, including the related objects, their positions and motions, etc. Compared to image-based NLP tasks such as image QA and captioning (Antol et al., 2015; Xu et al., 2015; Goyal et al., 2017), video-grounded dialogues are more challenging as the feature representation of a video involves both spatial and temporal dimensions. Ideally, a dialogue agent has to process information of both dimensions to address the two major questions: "where to look" (spatial reasoning) and "when to look" (temporal reasoning) in the video.



H: How many people are in the video? **A**: There is just one person
H: What room is he in? **A**: The man is in the kitchen
H: Does he enter the kitchen? **A**: No he is already in the kitchen
H: What does he do before he sits down? **A**: He closes the cabinet and then walks to the table and pulls the chair out to sit down
H: Is he happy or sad? **A**: He appears to be neutral in expression

H: hello , what is the first thing he does ? **A**: he takes off his long sleeve shirt
H: where does he put it ? **A**: he sets it on the table in front of him .
H: what does he do after that ? **A**: he opens the door and sits down on the ground .

Figure 1: Examples of video-grounded dialogues from the benchmark datasets of Audio-Visual Scene Aware Dialogues (AVSD) challenge (Alamri et al., 2018, 2019). **H**: human, **A**: the dialogue agent.

However, current approaches in video-grounded dialogues (Hori et al., 2019; Le et al., 2019b; Sanabria et al., 2019) often overlook spatial features and assume each spatial region is equally important to the current task (each spatial region is

[*] This work was mostly done when Hung Le was an intern at Salesforce Research Asia, Singapore.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1846–1859,
November 16–20, 2020. ©2020 Association for Computational Linguistics

assigned with a uniform weight). Such approach is appropriate for cases where the video involves just few objects and spatial positions can be treated similarly. However, in many scenarios (e.g. examples in Figure 1), each video frame often contains multiple distinct objects and not all of them are relevant to the given question.

Related tasks to video-grounded dialogues are video QA and video captioning. Previous efforts in these research areas such as (Jang et al., 2017; Aafaq et al., 2019) explicitly consider both spatial and temporal features of input video. These models learn to summarize spatial features based on their importance to question rather than considering each region equally. We are motivated by these approaches and propose to extend spatio-temporal reasoning to dialogues. However, rather than fixing on processing spatial inputs then learning temporal inputs, we note that in some cases, e.g. extended videos over a long period, it is more practical to first identify the relevant video segments before pinpointing the specific subjects of interest. Considering questions in a dialogue setting, it is appropriate to assume the questions are relevant to varying temporal locations of the video rather than just a small fixed segment. We, thus, propose to explore a bidirectional vision-language reasoning approach to fully exploit both spatial and temporal-level features through two reasoning directions.

Our approach includes two parallel networks to learn relevant visual signals from the input video based on the language signals from user utterances. Each network projects the language-based features to a three-dimensional tensor which is then used to independently learn video signals following a reasoning direction either as *spatial→temporal* or *temporal→spatial*. The output from each network is dynamically combined by importance scores computed based on language and visual features. The weighted output is recurrently used as input to the reasoning modules to allow the models to progressively derive relevant video signals over multiple steps. Intuitively, *spatial→temporal* reasoning is more appropriate for human queries related to specific entities or for input video involving many objects. *temporal→spatial* reasoning is more suitable for human queries about a particular video segment or for videos of extensive lengths.

We name our proposed approach Bidirectional Spatio-Temporal Learning (*BiST*), with the following contributions: (1) Rather than exploiting temporal-level information only, our approach equally emphasizes both spatial and temporal features of videos for higher-resolution queries of visual cues. (2) To tackle the diverse queried information from conversational queries, we propose a bidirectional strategy, denoted *spatial↔temporal*, to enable comprehensive information diffusion between the two visual feature spaces. (3) Our models achieve competitive performance on the “AVSD” (Audio-Visual Scene Aware Dialogues) benchmark from the 7^{th} Dialogue System Technology Challenge (DSTC7) (Alamri et al., 2018, 2019). We adapt our models to a video QA task “TGIF-QA” (Jang et al., 2017) and achieve significant performance gains. (4) We conduct a comprehensive ablation and qualitative analysis and demonstrate the efficacy of our bidirectional reasoning approach.

2 Related Work

Our work is related to two research topics: video-grounded dialogues and spatio-temporal learning.

Video-grounded Dialogues. Following recent efforts that combine NLP and Computer Vision research (Antol et al., 2015; Xu et al., 2015; Goyal et al., 2017), video-grounded dialogues are extended from the two major research fields: video action recognition and detection (Simonyan and Zisserman, 2014; Yang et al., 2016; Carreira and Zisserman, 2017) and dialogues/QA (Rajpurkar et al., 2016; Budzianowski et al., 2018; Gao et al., 2019a). Approaches to video-grounded dialogues (Sanabria et al., 2019; Hori et al., 2019; Le et al., 2019b) typically use pretrained video models, such as 2D CNN models on video frames (Donahue et al., 2015; Feichtenhofer et al., 2016), and 3D CNN models on video clips (Tran et al., 2015; Carreira and Zisserman, 2017), to extract visual features. However, these approaches mostly exploit the superficial information from the temporal dimension and neglect spatial-level signals. These approaches integrate spatial-level features simply through sum pooling with equal weights to obtain a global representation at the temporal level. They are, thus, not ideal for complex questions that investigate entity-level or spatial-level information (Jang et al., 2017; Alamri et al., 2019). The dialogue setting exacerbates this limitation as it allows users to explore various aspects of the video contents, including both low-level (spatial) and high-level (temporal) information, over multiple dialogue turns. Our approach aims to address this challenge in video-grounded dialogues by retrieving fine-grained information from video through a bidirectional reasoning framework.

Spatio-temporal Learning. Most efforts in spatio-temporal learning focus on action recognition or detection tasks. (Yang et al., 2019) proposes to progressively refine coarse-scale information through temporal extension and spatial displacement for action detection. (Li et al., 2019a) uses a shared network of 2D CNNs over three orthogonal views of video to obtain spatial and temporal signals for action recognition. (Qiu et al., 2019) adopts a two-path network architecture that integrates global and local information of both temporal and spatial dimensions for video classification. Other research areas that investigate spatio-temporal learning include video captioning (Aafaq et al., 2019), video super-resolution (Li et al., 2019b), and video object segmentation (Xu et al., 2019). In general, spatio-temporal learning approaches aim to process higher-resolution information from complex videos that involve multiple objects in each video frame or motions over video segments (Yang et al., 2019). We are motivated by a similar reason observed in video-grounded dialogues and explore a vision-language bidirectional reasoning approach to obtain more fine-grained visual features.

3 BiST Model

The input includes a video V, dialogue history of $(t-1)$ turns (where t is the current turn), each including a pair of (human utterance H, dialogue agent response A) $(H_1, A_1, ..., H_{t-1}, A_{t-1})$, and current human utterance H_t. The output is a system response A_t that can address current human utterance. The input video can contain features in different modalities, including vision, audio, and text (such as video caption or subtitle). Without loss of generalization, we can denote each text input as a sequence of tokens, each represented by a unique token index from a vocabulary set V: dialogue history X_{his}, user utterance X_{que}, text input of video X_{cap} , and output response Y. We also denote L_{S} as the length of a sequence S. For instance, L_{que} is the length of X_{que}.

Our model is composed of four parts: (1) The encoders encode text sequences and video inputs, including visual, audio, and text features, into continuous representations. For non-text features such as vision and sound, we follow previous work (Lei et al., 2018; Hori et al., 2019) and assume access to pre-trained models. (2) Several neural reasoning components learn dependencies between user utterances/queries and video features of multiple modalities. For video visual features, we propose to learn dependencies at both spatial and temporal levels in two directions (see Figure 2). Specifically, we allow interaction between each token in user query and each spatial position or temporal step of the video. The outputs from spatial-based or temporal-based reasoning are sequentially incorporated in two directions, *temporal→spatial* and *spatial→temporal*. The bidirectional strategy enables information being fused dynamically and captures complex dependencies between textual signals from dialogues and visual signals from videos. (3) The decoder passes encoded system responses over multiple attention steps, each of which integrates information from textual or video representations. The decoder output is passed to a generator to generate tokens by an auto-regressive way. (4) The generator computes three distributions over the vocabulary set, one distribution as output from a linear transformation and the others based on pointer attention scores over positions of input sequences.

3.1 Encoders

Text Encoder. We use an encoder to embed text-based input X into continuous representations $Z \in \mathbb{R}^{L_X \times d}$. L_X is the length of sequence X and d is the embedding dimension. A text encoder includes a token-level embedding layer and a layer normalization (Ba et al., 2016). The embedding layer includes a trainable matrix $E \in \mathbb{R}^{|V| \times d}$, with each row representing a token in the vocabulary set V as a vector of dimension d. We denote $E(X)$ as the embedding function that looks up the vector of each token in input sequence X: $Z_{\text{emb}} = E(X) \in \mathbb{R}^{L_X \times d}$. To incorporate the positional encoding layer, we adopt the approach from (Vaswani et al., 2017) with each token position represented as a sine or cosine function. The output from positional encoding and token-level embedding is combined through element-wise summation and layer normalization. The encoder outputs include representations for dialogue history Z_{his}, user query Z_{que}, video caption Z_{cap}, and target response Z_{res}. For target response, during training, the sequence is shifted left by one position to allow prediction in the decoding step i is auto-regressive on the previous positions $1, ..., (i-1)$. We share the embedding matrix E to encode all text sequences.

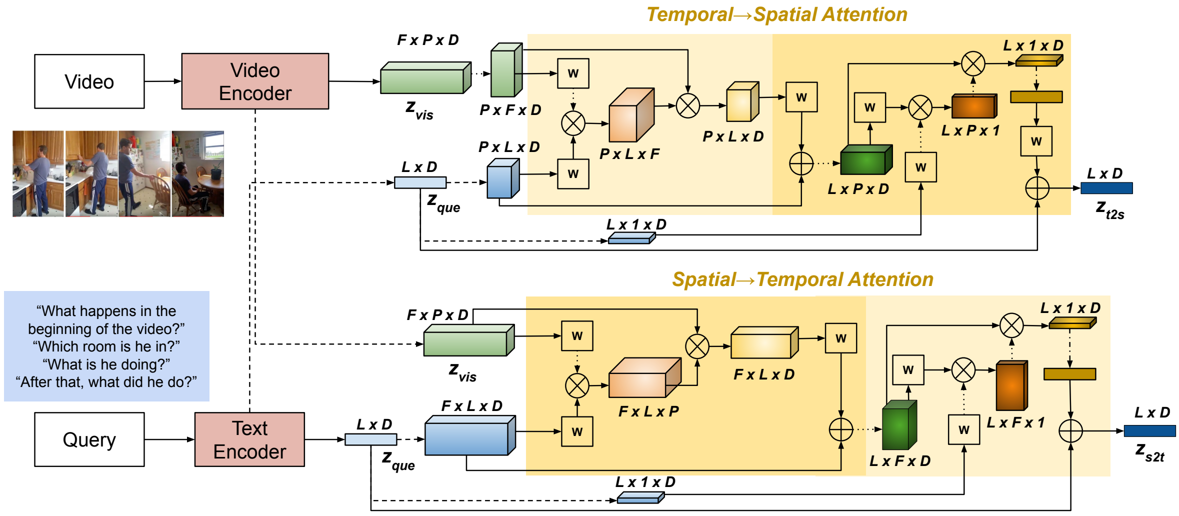


Figure 2: Our bidirectional approach models the dependencies between text and vision in two reasoning directions: spatial→temporal and temporal→spatial. $\otimes$ and $\oplus$ denote dot-product operation and element-wise summation.

Video Encoder. We make use of a 3D-CNN video model to extract spatio-temporal visual features. The dimensions of the resulting output depend on the configuration of sampling stride and clip length. We denote the output from a pretrained visual model as $Z_{\text{vis}}^{\text{pre}} \in \mathbb{R}^{F \times P \times d_{\text{vis}}^{\text{pre}}}$ where F is the number of sampled video clips, P is the spatial dimension from a 3D CNN layer, and $d_{\text{vis}}^{\text{pre}}$ is the feature dimension. We apply a linear layer with ReLU and layer normalization to reduce feature dimension to $d \ll d_{\text{vis}}^{\text{pre}}$. For audio features, we follow similar procedure to obtain audio representation $Z_{\text{aud}} \in \mathbb{R}^{F \times d}$. We keep the pretrained visual and audio models fixed and directly use extracted features to our dialogue models.

3.2 Bi-directional Reasoning

We propose a bidirectional architecture whereby the text features are used to select relevant information in both spatial and temporal dimensions in two reasoning directions (See Figure 2).

Temporal→spatial. In one direction, the user query is used to select relevant information along temporal steps of each spatial region independently. We first stack the encoded query features to P spatial positions and denote the stacked features as $Z_{\text{que}}^{\text{stack}} \in \mathbb{R}^{P \times L_{\text{que}} \times d}$. For each spatial position, the model learns the dependencies between question and each of F temporal steps through an attention mechanism as follows:

$$Z_{t2s}^{(1)} = Z_{\text{vis}}^{\mathsf{T}} W_{t2s}^{(1)} \in \mathbb{R}^{P \times F \times d_{\text{att}}} \quad (1)$$

$$Z_{t2s}^{(2)} = Z_{\text{que}}^{\text{stack}} W_{t2s}^{(2)} \in \mathbb{R}^{P \times L_{\text{que}} \times d_{\text{att}}} \quad (2)$$

$$S_{t2s}^{(1)} = \text{Softmax}(Z_{t2s}^{(2)} {Z_{t2s}^{(1)}}^{\mathsf{T}}) \in \mathbb{R}^{P \times L_{\text{que}} \times F} \quad (3)$$

where d_{att} is the dimension of the attention hidden layer, $W_{t2s}^{(1)} \in \mathbb{R}^{d \times d_{att}}$ and $W_{t2s}^{(2)} \in \mathbb{R}^{d \times d_{att}}$. The attention scores $S_{t2s}^{(1)}$ are used to obtain weighted sum along the temporal dimension of each spatial position of Z_{vis}. The resulting tensor is passed through a linear transformation and ReLU layer. The output contains temporally attended visual features and are combined with language features through skip connection. We denote the output by vector Z_{t2s}^{t}.

From the temporally attended features, user query is used again to obtain dependencies along the spatial dimension. We use a similar attention network to model the interaction between each token in query and each temporally attended spatial region.

$$Z_{t2s}^{(3)} = Z_{t2s}^{t} W_{t2s}^{(3)} \in \mathbb{R}^{L_{\text{que}} \times P \times d_{\text{att}}} \quad (4)$$

$$Z_{t2s}^{(4)} = Z_{\text{que}} W_{t2s}^{(4)} \in \mathbb{R}^{L_{\text{que}} \times d_{\text{att}}} \quad (5)$$

$$S_{t2s}^{(2)} = \text{Softmax}(Z_{t2s}^{(3)} {Z_{t2s}^{(4)}}^{\mathsf{T}}) \in \mathbb{R}^{L_{\text{que}} \times P} \quad (6)$$

where $W_{t2s}^{(3)} \in \mathbb{R}^{d \times d_{\text{att}}}$ and $W_{t2s}^{(4)} \in \mathbb{R}^{d \times d_{\text{att}}}$. The attention scores $S_{t2s}^{(2)}$ is used to obtain the weighted sum of all spatial positions from Z_{t2s}^{t}. The output is temporal-to-spatially attended visual features and is incorporated into language features through skip connection. We denote the resulting output as Z_{t2s}.

Spatial→temporal. In this reasoning direction, similar neural operations are used to compute spatially attended features followed by temporally attended features. The main difference from the other reasoning direction is that we stacked the query features to F temporal steps to obtain $Z_{\text{que}}^{\text{stack}} \in \mathbb{R}^{F \times L_{\text{que}} \times d}$. Other network components, including two attention layers, are as described in Equation 1 to 6. The final output is denoted as Z_{s2t}.

Previous approaches in video-based NLP tasks (Yu et al., 2016; Jang et al., 2017; Hori et al., 2019) focus on the interaction between global representations of questions and temporal-level representations of videos. This strategy potentially loses critical information on spatial variations in video frames. Our approach does not only emphasize both spatial and temporal feature spaces but also allows neural models to diffuse information from these feature spaces in two different ways. As we can consider spatial information as local signals and temporal information as global signals, our approach enables global-to-local and local-to-global diffusion of visual cues in video. This approach is similar to (Qiu et al., 2019) in which local and global visual signals are learned and diffused iteratively. However, different from this approach, our approach focuses on language-vision reasoning for more accurate visual information queries.

Multimodal Reasoning. In addition to language-vision reasoning, our models also consider learning of other information dependencies between queries and audio inputs or textual video inputs.

- Language→Audio Reasoning. We adopt similar neural operations from language-vision reasoning. The difference is that we directly use the query features without stacking the features into Equation 1 to 3. The resulting output of text-audio reasoning is denoted as Z_{q2a} which contains query-guided temporally attended features of Z_{aud}.
- Language→Language Reasoning. This reasoning module focuses on the unimodal dependencies between user query and video caption (if the caption is available). As the caption can contain useful information about the video content, we apply the dot-product attention mechanism similarly as with audio features to obtain Z_{q2c}.

Multimodal Fusioning. Given the attended features, we combine them to obtained query-guided video representation, incorporating information from all modalities. We denote the concatenated representation in the following:

$$Z_{\text{q2vid}} = [Z_{\text{que}}; Z_{t2s}; Z_{s2t}, Z_{\text{q2a}}, Z_{\text{q2c}}] \in \mathbb{R}^{L_{\text{que}} \times 5d}$$

where ; is the concatenation operation. The features are combined through an importance score matrix:

$$S_{\text{vid}} = \text{Softmax}(Z_{\text{q2vid}} W_{\text{q2vid}}) \in \mathbb{R}^{L_{\text{que}} \times 4}$$

where $W_{\text{q2vid}} \in \mathbb{R}^{5d \times 4}$. The scores from S_{vid} are used to obtain the weighted sum of component video modalities, resulting in a fusion vector from multiple modalities. We denote the resulting output Z_{vid}. Compared to previous work such as (Hori et al., 2019; Le et al., 2019b) which generally treat all modalities equally, our multimodal features are fused in a question-dependent manner. Potentially, our approach can avoid noisy or unnecessary signals, e.g. audio features not needed for questions only concerning visual contents.

3.3 Response Decoder

The decoder aims to decode system responses in an auto-regressive manner. During inference, a special token $\langle$sos$\rangle$ is fed to the decoder. The output token is then concatenated to this special token as input to the decoder again to decode the second token. This repeats until reaching a limit of decoding rounds or when the special token $\langle$eos$\rangle$ is predicted. We apply a similar decoding architecture as (Le et al., 2019b). The decoder includes three attention layers to incorporate contextual cues from textual components to the output token representations. The first layer is a self-attention to learn dependencies among the current tokens. Intuitively, this helps to shape a more semantically structured sequence. The second and third attention steps are used to capture contextual information from dialogue history and current user query to make the responses coherently connected to the whole dialogue context. To incorporate contextual cues from video components, our decoder is slightly different from (Le et al., 2019b). Instead of sequentially going through multiple attention layers, we only need one layer on the fused features Z_{vid}. This is more memory efficient since it only requires a single attention operation. It also does not depend on the design decision of the ordering of attention layers. At decoding step j, we denote the decoder output as $Z_{\text{dec}} \in \mathbb{R}^{j \times d}$.

3.4 Pointer Generator

Given the output from the decoder, the generator network is used to materialize responses in natural language. A linear transformation is used to obtain distribution over the vocabulary set V.

$$P_{\text{vocab}} = \text{Softmax}(Z_{\text{dec}} W_{\text{vocab}}) \in \mathbb{R}^{j \times |V|}$$

where $W_{\text{vocab}} \in \mathbb{R}^{d\times|V|}$. We share the weights between W_{vocab} and E as the semantics between source sequences and target responses are similar. To strengthen the model generation capability, we adopt pointer networks (Vinyals et al., 2015) to emphasize tokens from source sequences, i.e. user queries and video captions. We denote $\text{Ptr}(Z_1, Z_2)$ as the pointer network operation i.e. each token in Z_2 is "pointed" to all tokens in Z_1 through a learnable probability distribution. The resulting probability distribution is aggregated by all tokens in Z_1 to obtain $\text{Ptr}(Z_1, Z_2) \in L_{Z_2} \times |V|$. The final output distribution, denoted $P_{\text{out}} \in \mathbb{R}^{j\times|V|}$, is the weighted sum of three distributions: P_{vocab}, $\text{Ptr}(Z_{\text{que}}, Z_{\text{dec}})$, and $\text{Ptr}(Z_{\text{cap}}, Z_{\text{dec}})$. The weights for this fusion are learned via a linear transformation with softmax: $\alpha = \text{Softmax}(Z_{\text{gen}}W_{\text{gen}}) \in \mathbb{R}^{L_{\text{res}}\times 3}$ where $Z_{\text{gen}} = [Z_{\text{res}}; Z_{\text{dec}}; Z^{\text{exp}}_{\text{que}}; Z^{\text{exp}}_{\text{cap}}] \in \mathbb{R}^{j\times 4d}$, $W_{gen} \in \mathbb{R}^{4d\times 3}$, and $Z^{\text{exp}}_{\text{que}}$ and $Z^{\text{exp}}_{\text{cap}}$ are the stacked tensors of caption and user queries to j dimensions.

Optimization. During training, we learn all model parameters by minimizing the generation loss:

$$\mathcal{L} = \sum_{j=0}^{L_Y} -\log(P_{\text{out}}(y_j)).$$

4 Experiments

4.1 Experimental Setups

Datasets. We use the AVSD benchmark from DSTC7 (Alamri et al., 2018, 2019) which contains dialogues grounded on the Charades videos (Sigurdsson et al., 2016). In addition, we adapt our models to the video QA benchmark TGIF-QA (Jang et al., 2017). (See Table 1 for a summary of the two datasets). To extract visual and audio features, we used 3D-CNN ResNext-101 (Xie et al., 2017) pretrained on Kinetics (Hara et al., 2018) to obtain spatio-temporal visual features and VGGish pretrained on YouTube videos (Hershey et al., 2017) to extract (temporal) audio features. We sample video clips to extract visual features with a window size of 16 frames, and stride of 16 and 4 in AVSD and TGIF-QA respectively. In TGIF-QA experiments, we also extract visual features from pretrained ResNet-152 (He et al., 2016) for a fair comparison with existing work. In AVSD experiments, we make use of the video summary as the video-dependent text input X_{cap}.

Training Procedure. We adopt the Adam optimizer (Kingma and Ba, 2015) and the learning rate strategy from (Vaswani et al., 2017). We set the learning rate *warm-up* steps equivalent to 5 epochs and train models up to 50 epochs. We select the best models based on the average loss per epoch in the validation set. We initialize all model parameters with uniform distribution (Glorot and Bengio, 2010). During training, we adopt the auxiliary auto-encoder loss function from (Le et al., 2019b). We adopt Transformer attention (Vaswani et al., 2017) in our models and select the following hyper-parameters: $d = d_{\text{att}} = 128$, $N_{\text{att}} = N_{\text{dec}} = 3$, and $h_{\text{att}} = 8$ where N_{att} and N_{dec} are the number of Transformer blocks in multimodal reasoning and decoder networks and h_{att} is the number of attention heads. We tuned other hyper-parameters following grid-search over the validation set. In AVSD experiments, we train our models by applying label smoothing (Szegedy et al., 2016) on the target system responses Y. We adopt a beam search technique with a beam size 5.

Benchmark	#	**Train**	**Val.**	**Test**
AVSD	Dialogs	7,659	1,787	1,710
	Turns	153,180	35,740	13,490
	Words	1,450,754	339,006	110,252
TGIF-QA	Count QA	24,159	2,684	3,554
	Action QA	18,428	2,047	2,274
	Trans. QA	47,434	5,270	6,232
	Frame QA	35,453	3,939	13,691

Table 1: Summary of DSTC7 AVSD and TGIF-QA benchmark. The TGIF-QA contains 4 different tasks: (1) Count: open-ended QA which counts the number of repetitions of an action. (2) Action: multi-choice (MC) QA about a certain action occurring a fixed number of times. (3) Transition: MC QA about the temporal variation of video. (4) Frame: open-ended QA which can be answered from one video frame.

4.2 Modifications for Video QA

In many Video QA benchmarks such as TGIF-QA (Jang et al., 2017), the tasks are retrieval-based (e.g. output a single score for each output candidate) rather than generation-based as in many dialogue tasks. Following (Fan et al., 2019), we first concatenate the question with each candidate answer individually and treat this as Z_{que} to our models. As there is no target response to be decoded, we adapt our models to this setting by using a trainable vector $z_j \in \mathbb{R}^d$ to represent a candidate response R_j, replacing $Z_{\text{res}} \in \mathbb{R}^{j\times d}$ in a dialogue, as input to the decoder. The output, denoted $Z_{j,\text{dec}} \in \mathbb{R}^d$, is passed to a linear transformation layer to obtain a score $s_{j,\text{out}} = Z_{j,\text{dec}}W_{\text{out}} \in \mathbb{R}$

Model	Z_{vis}	Z_{aud}	Z_{cap}	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
Baseline (Hori et al., 2019)	I3D	-	-	0.621	0.480	0.379	0.305	0.217	0.481	0.733
MTN (Le et al., 2019b)	I3D	-	-	0.654	0.521	0.420	0.343	0.247	0.520	0.936
MTN (Le et al., 2019b)	ResNext	-	-	0.688	0.550	0.444	0.363	0.260	0.541	0.985
BiST	ResNext	-	-	**0.711**	**0.578**	**0.475**	**0.394**	**0.261**	**0.550**	**1.050**
Video Sum. (Sanabria et al., 2019)	ResNext	-	✓	0.718	0.584	0.478	0.394	0.267	0.563	1.094
Video Sum.+How2 (Sanabria et al., 2019)	ResNext	-	✓	0.723	0.586	0.476	0.387	0.266	0.564	1.087
MTN (Le et al., 2019b)	I3D	-	✓	0.715	0.581	0.476	0.392	0.269	0.559	1.066
MTN (Le et al., 2019b)	ResNext	-	✓	0.731	0.597	0.490	0.406	0.271	0.564	1.127
BiST	ResNext	-	✓	**0.754**	**0.622**	**0.515**	**0.430**	**0.284**	**0.584**	**1.190**
Baseline (Hori et al., 2019)	I3D	VGGish	-	0.626	0.485	0.383	0.309	0.215	0.487	0.746
Baseline+GRU+HierAttn. (Le et al., 2019a)	I3D	VGGish	-	0.631	0.491	0.390	0.315	0.239	0.509	0.848
FA+HRED (Nguyen et al., 2018)	I3D	VGGish	-	0.648	0.505	0.399	0.323	0.231	0.510	0.843
Student-Teacher (Hori et al., 2019)	I3D	VGGish	-	0.675	0.543	0.446	0.371	0.248	0.527	0.966
MTN (Le et al., 2019b)	I3D	VGGish	-	0.692	0.556	0.450	0.368	0.259	0.537	0.964
MTN (Le et al., 2019b)	ResNext	VGGish	-	0.688	0.554	0.452	0.372	0.251	0.531	0.950
BiST	ResNext	VGGish	-	**0.715**	**0.560**	**0.477**	**0.390**	**0.259**	**0.552**	**1.030**
Baseline+GRU+HierAttn. (Le et al., 2019a)	I3D	VGGish	✓	0.633	0.490	0.386	0.310	0.242	0.515	0.856
FA+HRED (Nguyen et al., 2018)	I3D	VGGish	✓	0.695	0.553	0.444	0.360	0.249	0.544	0.997
Student-Teacher (Hori et al., 2019)	I3D	VGGish	✓	0.727	0.593	0.488	0.405	0.273	0.566	1.118
MTN (Le et al., 2019b)	I3D	VGGish	✓	0.731	0.597	0.494	0.410	0.274	0.569	1.129
MTN (Le et al., 2019b)	ResNext	VGGish	✓	0.735	0.600	0.498	0.413	0.275	0.571	1.137
BiST	ResNext	VGGish	✓	**0.755**	**0.619**	**0.510**	**0.429**	**0.284**	**0.581**	**1.192**

Table 2: Evaluation results on the test split of the AVSD benchmark. The results are presented in 4 settings by video feature components: (1) visual-only, (2) visual and text, (3) visual and audio, and (4) visual, audio, and text.

where $W_{\text{out}} \in \mathbb{R}^{d\times 1}$. In this setting, we remove the language→language and language→audio reasoning modules. The loss function is the summed pairwise hinge loss (Jang et al., 2017) between scores of positive answer s^p_{out} and each negative answer $s^n_{j,\text{out}}$. $\mathcal{L} = \sum_{j=1}^{K} max(0, m - (s^p_{\text{out}} - s^n_{j,\text{out}}))$ where K is the total number of candidate answers and m is a hyper-parameter used as a margin between positive and negative answers.

Training. Multiple-choice tasks, including *Action* and *Transition*, are trained following the pairwise loss with $K = 5$ and $m = 1$. *Count* task is trained with similar approach but as a regression problem with a single output score s_{out}. The loss function is measured as mean square error between output s_{out} and label y. The open-ended *Frame* task is trained as a generation task, similarly to the dialogue response generation task, with a single-token output. We use the the vector $z \in \mathbb{R}^d$ as input to the decoder. The generator includes a single linear layer with $W_{\text{out}} \in \mathbb{R}^{d\times|v|}$. We do not apply pointer network in this case as the output is only a single-token response.

4.3 Results

AVSD Results. We report the objective scores, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). These metrics, which formulate lexical overlaps between generated and ground-truth dialogue responses, are borrowed from language generation tasks such as machine translation and captioning. We compare our generated responses with 6 reference responses. Major baseline models are: (1) *Baseline* (Alamri et al., 2018; Hori et al., 2019) consists of LSTM-based encoder-encoder with attention layers between user queries and temporal-level visual and audio features. (2) *Baseline+GRU+HierAttn.* (Le et al., 2019a) extends (1) through GRU and question-guided self-attention and caption attention. (3) *FA+HRED* (Nguyen et al., 2018) adopts FiLM neural blocks for language-vision dependency learning. (4) *Video Summarization* (Sanabria et al., 2019) reformulates the task as a video summarization task and enhances the models with transfer learning from a large-scale summarization benchmark. (5) *Student-Teacher* (Hori et al., 2019) adopts dual network architecture in which a student network is trained to mimic a teacher network trained with additional video-dependent text input. (6) *MTN* (Le et al., 2019b) fuses temporal features of different modalities sequentially through a Transformer decoder architecture. (7) *FGA* (Schwartz et al., 2019) consists of attention networks between all pairs of modalities and the models aggregate attention scores along edges of an attention graph.

In Table 2, we present the scores by different combinations of features, including vision Z_{vis}, audio Z_{aud}, and text Z_{cap}. In all settings, our models outperform the existing approaches. The performance of our models in the visual-only setting shows the performance gain coming from our bidirectional language-vision reasoning approach. We

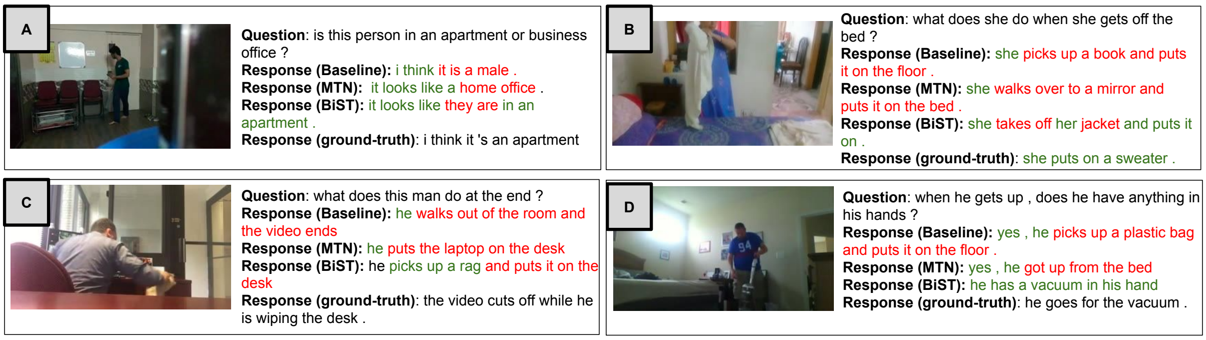
Model	Z_{vis}	Count (Loss)	Action (Acc)	Trans. (Acc)	Frame (Acc)
VIS (aggr) (Ren et al., 2015)	R	5.09	0.468	0.569	0.346
VIS (avg) (Ren et al., 2015)	R	4.80	0.488	0.348	0.350
MCB (aggr) (Fukui et al., 2016)	R	5.17	0.589	0.243	0.257
MCB (avg) (Fukui et al., 2016)	R	5.54	0.291	0.330	0.155
Yu *et al.* (Yu et al., 2017)	R	5.13	0.561	0.640	0.396
ST-VQA (s) (Jang et al., 2017)	R+C	4.28	0.573	0.637	0.455
ST-VQA (t) (Jang et al., 2017)	R+C	4.40	0.608	0.671	0.493
ST-VQA (st) (Jang et al., 2017)	R+C	4.56	0.570	0.596	0.478
Co-Mem (Gao et al., 2018)	R+F	4.10	0.682	0.743	0.515
PSAC (Li et al., 2019c)	R	4.27	0.704	0.769	0.515
HME (Fan et al., 2019)	R+C	4.02	0.739	0.778	0.538
STA (Gao et al., 2019b)	R	4.25	0.723	0.790	0.566
CRN+MAC (Le et al., 2019c)	R	4.23	0.713	0.787	0.592
MQL (Lei et al., 2020)	V	-	-	-	0.598
QueST (Jiang et al., 2020)	R	4.19	0.759	0.810	0.597
HGA (Jiang and Han, 2020)	R+C	4.09	0.754	0.810	0.551
GCN (Huang et al., 2020)	R+C	3.95	0.743	0.811	0.563
HCRN (Le et al., 2020)	R+RX	3.82	0.750	0.814	0.559
BiST	R	2.40	0.839	0.817	0.630
BiST	RX	**2.19**	**0.847**	**0.819**	**0.648**

Table 3: Evaluation results on the test split of the TGIF-QA benchmark. Visual features are: R(ResNet), C(C3D), F(FlowCNN), RX(ResNext).

also observe a performance boost whenever the text feature from video is considered. When we add the audio features, however, the performance gain is not significant. This reveals a potential future extension in our work to better combine visual and audio feature representations. FGA (Schwartz et al., 2019) reports the CIDEr score of 0.806 in the visual-only setting. Compared to FGA, our performance gain indicates the efficacy of learning fine-grained dependencies between query and visual features at both spatial and temporal levels to select relevant information from video.

TGIF-QA Results. We give the L2 loss for *Count* task and accuracy for the other three QA tasks (See Appendix A for description of baseline models). From Table 3, our model outperforms existing approaches across all QA tasks, using either frame-level (appearance) feature, ResNet, or sequence-level feature, ResNext. Our models perform better with ResNext as we expect sequence-level feature is more consistent than frame-level feature. Experiments on this benchmark show clearer performance gain of our bidirectional language-vision reasoning approach as the performance is not affected by errors of generation components as in the AVSD experiments. By focusing on learning high-resolution dependencies from spatio-temporal features, our models can fully exploit contextual cues and select better answers for video QA tasks.

Impacts of Spatio-temporal Learning. We consider model variants based on the spatio-temporal dynamics and report the results in Table 4. We noted that when using a single reasoning direction, the model with *temporal→spatial* performs better than one with the reverse reasoning direction. This observation is different from prior approaches of spatio-temporal learning such as (Jang et al., 2017) which are limited to the reasoning order *spatial→temporal*. This can be explained as the videos in the AVSD benchmark are typically longer than other QA benchmarks. It is practical to focus on temporal locations in frame sequences first before selecting spatial regions in individual frames. In addition, dialogue queries are positioned in a multi-turn setting whereby each turn is relevant to different video segments as the dialogue evolves. Potentially, this observation indicates an important difference of video-grounded dialogues compared to video QA. Secondly, we also observe that our model performance improves when we use both reasoning directions rather than only one of them. Our motivation for this approach is similar to (Schuster and Paliwal, 1997) who proposes a bidirectional strategy to process sequences in both forward and backward directions. Similarly, our approach exploits visual information through a bidirectional information diffusion strategy that can interpret information from both spatial or temporal aspects based on language input. Finally, we observe that using spatio-temporal features is better



Question: is this person in an apartment or business office ?
Response (Baseline): i think it is a male .
Response (MTN): it looks like a home office .
Response (BiST): it looks like they are in an apartment .
Response (ground-truth): i think it 's an apartment

Question: what does she do when she gets off the bed ?
Response (Baseline): she picks up a book and puts it on the floor .
Response (MTN): she walks over to a mirror and puts it on the bed .
Response (BiST): she takes off her jacket and puts it on .
Response (ground-truth): she puts on a sweater .

Question: what does this man do at the end ?
Response (Baseline): he walks out of the room and the video ends
Response (MTN): he puts the laptop on the desk
Response (BiST): he picks up a rag and puts it on the desk
Response (ground-truth): the video cuts off while he is wiping the desk .

Question: when he gets up , does he have anything in his hands ?
Response (Baseline): yes , he picks up a plastic bag and puts it on the floor .
Response (MTN): yes , he got up from the bed
Response (BiST): he has a vacuum in his hand
Response (ground-truth): he goes for the vacuum .

Figure 3: Comparison of dialogue response outputs of BiST against the baseline models. Parts of the outputs that match and do not match the ground truth are highlighted in green and red respectively.

than only using one of them, demonstrating the importance of information in both dimensions. To obtain Z_{vis} for spatial-only or temporal-only features, the spatio-temporal features are passed through an average pooling operation along the temporal or spatial dimensions respectively.

t2s	s2t	BLEU4	METEOR	ROUGE-L	CIDEr
✓	✓	**0.430**	**0.284**	**0.584**	**1.190**
✓		0.422	0.281	0.581	1.183
	✓	0.420	0.282	0.579	1.177
t only		0.419	0.278	0.573	1.156
s only		0.418	0.276	0.570	1.150

Table 4: Ablation analysis on the AVSD benchmark with variants of BiST by spatio-temporal dynamics.

Ablation Analysis. We conduct experiments with model variants of different hyper-parameter settings. Specifically, we vary the the number of attention rounds N_{att} and attention heads h_{att}. From Table 5, we noted the contribution of the multi-round architecture to language-vision reasoning as the performance improves with larger reasoning steps, i.e. up to three attention rounds. However, we observe that as we increase to more than 3 reasoning steps, the model performance only improves slightly. We also note that using a multi-head attention mechanism is suitable for tasks dealing with information-intensive media such as video and dialogues. The multi-head structure enables feature projection to multiple subspaces and capture complex language-vision dependencies.

Qualitative Analysis. In Figure 3, we present some example outputs. We note that the predicted dialogue responses of BiST models are closer to the ground-truth responses. Particularly for complex questions that query specific segments (example B, C, D), and/or specific spatial locations (Example D), our approach can generally produce better responses. Another observation is that for ambiguous examples such as Example C (where the visual appearance is not clear to differentiate "apartment" and "business office"), our model can return the correct answer. Potentially this can be explained by the extracted signals from spatial-level feature representations. Finally, we note that there are still some errors that make the output sentences partially wrong, such as mismatching subjects (example A), wrong entities (Example B), or wrong actions (Example C). For detailed qualitative analysis, please refer to Appendix B.

N	h_{att}	BLEU4	METEOR	ROUGE-L	CIDEr
3-3	8	**0.430**	**0.284**	**0.584**	1.190
1-1	8	0.418	0.280	0.574	1.171
2-2	8	0.422	0.278	0.576	1.171
3-3	1	0.414	0.278	0.580	1.173
3-3	2	0.418	0.280	0.579	1.174
3-3	4	0.428	0.280	**0.584**	**1.195**

Table 5: Performance of model variants by $N = N_{\text{att}} = N_{\text{dec}}$, and h_{att} on the AVSD benchmark

5 Conclusion

We proposed BiST, a novel deep neural network approach for video-grounded dialogues and video QA, which exploits the complex visual nuances of videos through a bidirectional reasoning framework in both spatial and temporal dimensions. Our experimental results show that BiST can extract relevant, high-resolution visual cues from videos and generate quality dialogue responses/answers.

References

Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12487–12496.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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*, volume 2.

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.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019a. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

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

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019b. Structured two-stream attention network for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6391–6398.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

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.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

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.

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.

C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356.

Chiori Hori, Anoop Cherian, Tim K Marks, and Takaaki Hori. 2019. Joint student-teacher learning for audio-visual scene-aware dialog. *Proc. Interspeech 2019*, pages 1886–1890.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*, volume 1.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Hung Le, S Hoi, Doyen Sahoo, and N Chen. 2019a. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019b. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2019c. Learning to reason with relational video representation for question answering. *arXiv preprint arXiv:1907.04553*.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. *arXiv preprint arXiv:2002.10698*.

Chenyi Lei, Lei Wu, Dong Liu, Zhao Li, Guoxin Wang, Haihong Tang, and Houqiang Li. 2020. Multi-question learning for visual question answering. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2019a. Collaborative spatiotemporal feature learning for video action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. 2019b. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019c. Beyond rnns: Positional self-attention with co-attention for video question answering. In *The 33rd AAAI Conference on Artificial Intelligence*, volume 8.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2018. From film to video: Multi-turn question answering with multi-modal context. In *AAAI 2019 Dialog System Technology Challenge (DSTC7) Workshop*.

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.

Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. 2019. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12056–12065.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961.

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbad's submission for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*, volume 6.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

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.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

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)*.

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, Ł ukasz 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, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. 2019. Spatiotemporal cnn for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1379–1388.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, 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.

Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. 2016. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 24th ACM international conference on multimedia*, pages 978–987. ACM.

Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. 2019. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–272.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173.

A TGIF-QA Baselines

In TGIF-QA experiments, we compare our models with the following baselines: (1) *VIS* (Ren et al., 2015) and (2) *MCB* (Fukui et al., 2016) are two image-based VQA baselines which were adapted to TGIF-QA by (Jang et al., 2017). (3) Yu *et al.* (Yu et al., 2017) uses a high-level concept word detector and the detected words are used for semantic reasoning. (4) *ST-VQA* (Jang et al., 2017) integrates temporal and spatial features by first pretraining temporal part and then finetuning the spatial part. (5) *Co-Mem* (Gao et al., 2018) includes a co-memory mechanism on two video streams based on motion and appearance features. (6) *PSAC* (Li et al., 2019c) uses multi-head attention layers to exploit the dependencies between text and temporal variation of video. (7) *HME* (Fan et al., 2019) is a memory network with read and write operations to update global context representations. (8) *STA* (Gao et al., 2019b) divides video into N segments and uses temporal attention modules on each segment independently. (9) *CRN+MAC* (Le et al., 2019c) is a clip-based reasoning framework by aggregating frame-level features into clips through temporal attention. (10) *MQL* (Lei et al., 2020) exploits the semantic relations among questions and proposes a multi-label prediction task. (11) *QueST* (Jiang et al., 2020) has two types of question embeddings: spatial and temporal embeddings based on attention guided by video features. (12) *HGA* (Jiang and Han, 2020) is a graph alignment network consisting of inter- and intra-modality edges to model the interaction between video and question. (13) *GCN* (Huang et al., 2020) is a similar approach with graph network but utilizes the video object-level features as node representations. (14) *HCRN* (Le et al., 2020) extends (Le et al., 2019c) with a hierarchical relation network over temporal-level video features.

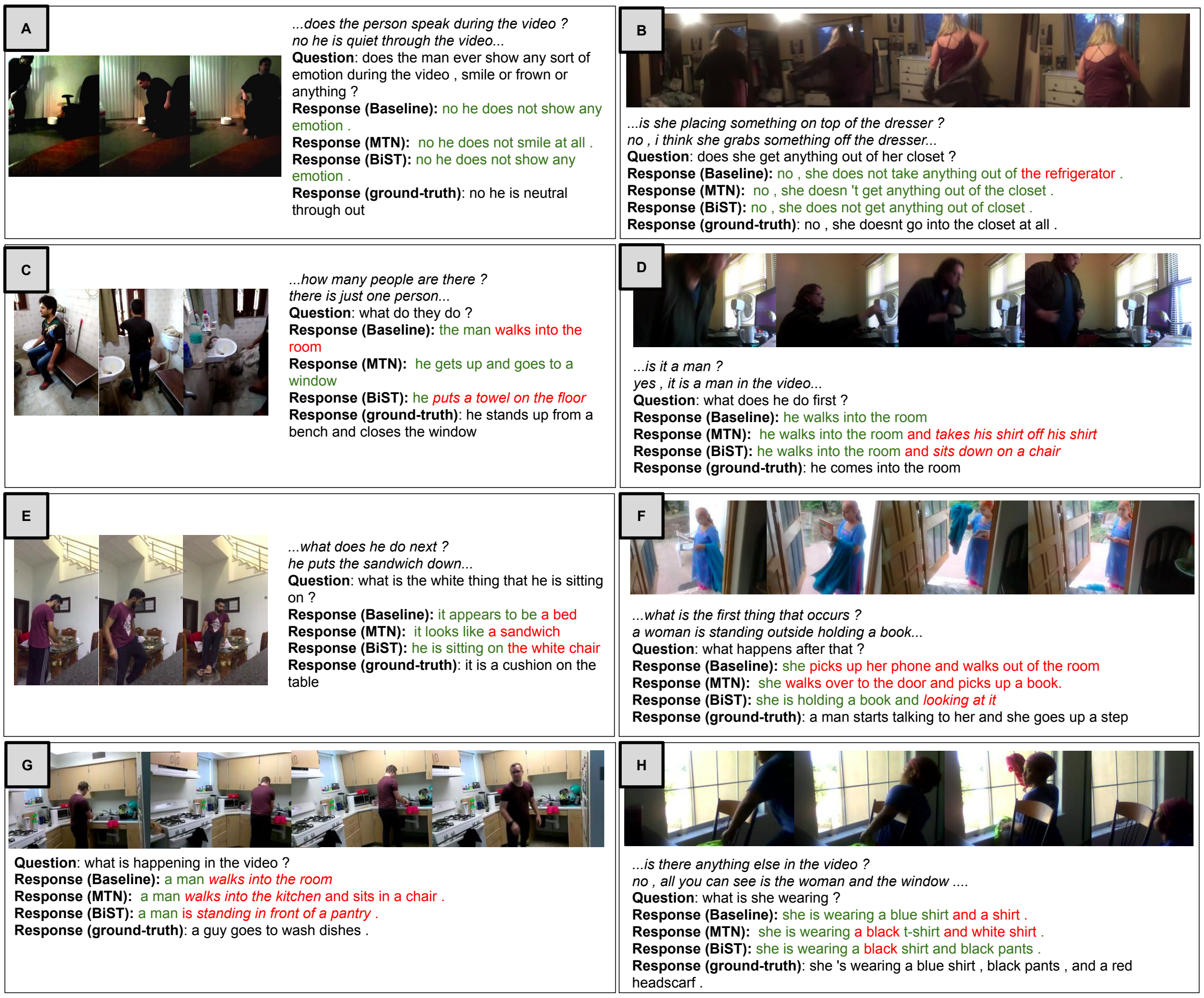
B Qualitative Analysis

We present additional example outputs in Figure 4. For each examples, we include the last dialogue turn from the dialogue history. In general, BiST can generate responses that better match the ground truth than the Baseline (Hori et al., 2019) and MTN (Le et al., 2019b) (example A, B). Furthermore, we analyze both negative and positive outputs and have the following observations:

- In cases where the videos contain more than one actions, our models can predict responses that describe multiple actions in their correct *orders of appearance*. For instance, in example D, even though our model response does not completely match the ground truth, it is still correctly explaining the sequence of actions, including first "walking into the room" and "sits down on a chair", matching the visual input from video. MTN response in the same example can express multiple actions but fail to detect the second action before "takes his shirt off". A similar observation can be found in the example F.

- In cases where the entities are hard to detect due to weak *visual distinction*, BiST can materialize the correct entity in its responses, e.g. in example C, "a towel" was seen in the last sampled video frame. Another example is example H where BiST detects both "shirt" and "pants" entities (even though their color attributes are not totally correct). However, in example E, all models fail to identify the entity "a cushion", possibly because of the ambiguous and subdue visual features of this object in the video. This displays an important challenge for more fine-grained information extraction in video-grounded dialogues.

- We noted our model fails in the following complex cases. First, for case with *ambiguous questions* such as example C, BiST emphasizes an action in the later part of the video (3^{rd} sampled frame) rather than the early part of the video (1^{st} and 2^{nd} sampled frame). This error might be due to the implied temporal specification in the question. Similarly, in example G, the ambiguous question results in generated responses of different action-level granularity from all models and some responses are partially correct. Secondly, in cases where the ground-truth answer involves *unseen entity* (example F with the entity "a man" without any visual appearance but possibly detected by his voice in the audio input), our model fails to include this entity in the response. A possible explanation for this example is that our model is not able to detect the entity based on audio input, i.e. "a man talking". This presents the retaining challenge to fully combine multiple modalities into natural language responses in dialogues.



A

...does the person speak during the video ?
no he is quiet through the video...
Question: does the man ever show any sort of emotion during the video , smile or frown or anything ?
Response (Baseline): no he does not show any emotion .
Response (MTN): no he does not smile at all .
Response (BiST): no he does not show any emotion .
Response (ground-truth): no he is neutral through out

B

...is she placing something on top of the dresser ?
no , i think she grabs something off the dresser...
Question: does she get anything out of her closet ?
Response (Baseline): no , she does not take anything out of the refrigerator .
Response (MTN): no , she doesn 't get anything out of the closet .
Response (BiST): no , she does not get anything out of closet .
Response (ground-truth): no , she doesnt go into the closet at all .

C

...how many people are there ?
there is just one person...
Question: what do they do ?
Response (Baseline): the man walks into the room
Response (MTN): he gets up and goes to a window
Response (BiST): he *puts a towel on the floor*
Response (ground-truth): he stands up from a bench and closes the window

D

...is it a man ?
yes , it is a man in the video...
Question: what does he do first ?
Response (Baseline): he walks into the room
Response (MTN): he walks into the room and *takes his shirt off his shirt*
Response (BiST): he walks into the room and *sits down on a chair*
Response (ground-truth): he comes into the room

E

...what does he do next ?
he puts the sandwich down...
Question: what is the white thing that he is sitting on ?
Response (Baseline): it appears to be a bed
Response (MTN): it looks like a sandwich
Response (BiST): he is sitting on the white chair
Response (ground-truth): it is a cushion on the table

F

...what is the first thing that occurs ?
a woman is standing outside holding a book...
Question: what happens after that ?
Response (Baseline): she picks up her phone and walks out of the room
Response (MTN): she walks over to the door and picks up a book.
Response (BiST): she is holding a book and *looking at it*
Response (ground-truth): a man starts talking to her and she goes up a step

G

Question: what is happening in the video ?
Response (Baseline): a man *walks into the room*
Response (MTN): a man *walks into the kitchen* and sits in a chair .
Response (BiST): a man is *standing in front of a pantry* .
Response (ground-truth): a guy goes to wash dishes .

H

...is there anything else in the video ?
no , all you can see is the woman and the window ....
Question: what is she wearing ?
Response (Baseline): she is wearing a blue shirt and a shirt .
Response (MTN): she is wearing a black t-shirt and white shirt .
Response (BiST): she is wearing a black shirt and black pants .
Response (ground-truth): she 's wearing a blue shirt , black pants , and a red headscarf .

Figure 4: Comparison of dialogue response outputs of BiST against the baseline models: *Baseline* (Hori et al., 2019) and *MTN* (Le et al., 2019b). Parts of the outputs that match and do not match the ground truth are highlighted in green and red respectively.