TVQA: Localized, Compositional Video Question Answering

Jie Lei **Licheng Yu** **Mohit Bansal** **Tamara L. Berg**
Department of Computer Science
University of North Carolina at Chapel Hill
{jielei, licheng, mbansal, tlberg}@cs.unc.edu

Abstract

Recent years have witnessed an increasing interest in image-based question-answering (QA) tasks. However, due to data limitations, there has been much less work on video-based QA. In this paper, we present TVQA, a large-scale video QA dataset based on 6 popular TV shows. TVQA consists of 152,545 QA pairs from 21,793 clips, spanning over 460 hours of video. Questions are designed to be compositional in nature, requiring systems to jointly localize relevant moments within a clip, comprehend subtitle-based dialogue, and recognize relevant visual concepts. We provide analyses of this new dataset as well as several baselines and a multi-stream end-to-end trainable neural network framework for the TVQA task. The dataset is publicly available at http://tvqa.cs.unc.edu.

1 Introduction

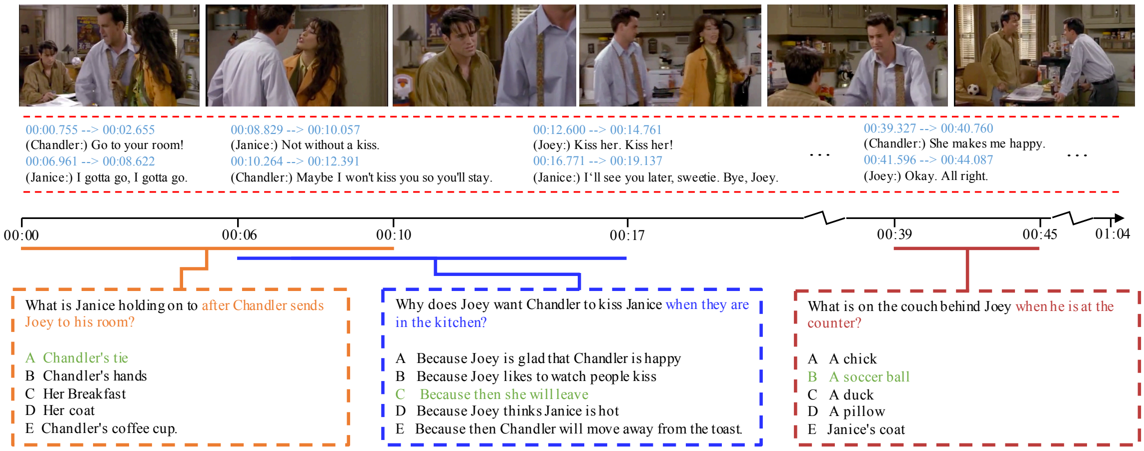
Now that algorithms have started to produce relevant and realistic natural language that can describe images and videos, we would like to understand what these models truly comprehend. The Visual Question Answering (VQA) task provides a nice tool for fine-grained evaluation of such multimodal algorithms. VQA systems take as input an image (or video) along with relevant natural language questions, and produce answers to those questions. By asking algorithms to answer different types of questions, ranging from object identification, counting, or appearance, to more complex questions about interactions, social relationships, or inferences about why or how something is occurring, we can evaluate different aspects of a model's multimodal semantic understanding.

As a result, several popular image-based VQA datasets have been introduced, including DAQUAR (Malinowski and Fritz, 2014), COCO-QA (Ren et al., 2015a), FM-IQA (Gao et al., 2015), Visual Madlibs (Yu et al., 2015), VQA (Antol et al., 2015), Visual7W (Zhu et al., 2016), etc. In addition, multiple video-based QA datasets have also been collected recently, e.g., MovieQA (Tapaswi et al., 2016), MovieFIB (Maharaj et al., 2017a), PororoQA (Kim et al., 2017), TGIF-QA (Jang et al., 2017), etc. However, there exist various shortcomings for each such video QA dataset. For example, MovieFIB's video clips are typically short (∼4 secs), and focused on purely visual concepts (since they were collected from audio descriptions for the visually impaired); MovieQA collected QAs based on text summaries only, making them very plot-focused and less relevant for visual information; PororoQA's video domain is cartoon-based; and TGIF-QA used predefined templates for generation on short GIFs.

With video-QA in particular, as opposed to image-QA, the video itself often comes with associated natural language in the form of (subtitle) dialogue. We argue that this is an important area to study because it reflects the real world, where people interact through language, and where many computational systems like robots or other intelligent agents will ultimately have to operate. As such, systems will need to combine information from what they see with what they hear, to pose and answer questions about what is happening.

We aim to provide a dataset that merges the best qualities from all of the previous datasets as well as focus on multimodal compositionality. In particular, we collect a new large-scale dataset that is built on natural video content with rich dynamics and realistic social interactions, where question-answer pairs are written by people observing both videos and their accompanying dialogues, encouraging the questions to require both vision and language understanding to answer. To further encourage this multimodal-QA quality, we ask people to write compositional questions consisting

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1369–1379
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics



00:00.755 --> 00:02.655
(Chandler:) Go to your room!
00:06.961 --> 00:08.622
(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057
(Janice:) Not without a kiss.
00:10.264 --> 00:12.391
(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761
(Joey:) Kiss her. Kiss her!
00:16.771 --> 00:19.137
(Janice:) I'll see you later, sweetie. Bye, Joey.

. . .

00:39.327 --> 00:40.760
(Chandler:) She makes me happy.
00:41.596 --> 00:44.087
(Joey:) Okay. All right.

. . .

What is Janice holding on to after Chandler sends Joey to his room?

A Chandler's tie
B Chandler's hands
C Her Breakfast
D Her coat
E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice when they are in the kitchen?

A Because Joey is glad that Chandler is happy
B Because Joey likes to watch people kiss
C Because then she will leave
D Because Joey thinks Janice is hot
E Because then Chandler will move away from the toast.

What is on the couch behind Joey when he is at the counter?

A A chick
B A soccer ball
C A duck
D A pillow
E Janice's coat

Figure 1: Examples from the TVQA dataset. All questions and answers are attached to 60-90 seconds long clips. For visualization purposes, we only show a few of the most relevant frames here. As illustrated above, some questions can be answered using subtitles or videos alone, while some require information from both modalities.

of two parts, a main question part, e.g. "What are Leonard and Sheldon arguing about" and a grounding part, e.g. "when they are sitting on the couch". This also leads to an interesting secondary task of QA temporal localization.

Our contribution is the TVQA dataset, built on 6 popular TV shows spanning 3 genres: medical dramas, sitcoms, and crime shows. On this data, we collected 152.5K human-written QA pairs (examples shown in Fig.1). There are 4 salient advantages of our dataset. First, it is large-scale and natural, containing 21,793 video clips from 925 episodes. On average, each show has 7.3 seasons, providing long range character interactions and evolving relationships. Each video clip is associated with 7 questions, with 5 answers (1 correct) for each question. Second, our video clips are relatively long (60-90 seconds), thereby containing more social interactions and activities, making video understanding more challenging. Third, we provide the dialogue (character name + subtitle) for each QA video clip. Understanding the relationship between the provided dialogue and the question-answer pairs is crucial for correctly answering many of the collected questions. Fourth, our questions are compositional, requiring algorithms to localize relevant moments (START and END points are provided for each question).

With the above rich annotation, our dataset supports three tasks: QA on the grounded clip, question-driven moment localization, and QA on the full video clip. We provide baseline experiments on both QA tasks and introduce a state-of-the-art language and vision-based model (leaving moment localization for future work).

2 Related Work

Visual Question Answering: Several image-based VQA datasets have recently been constructed, e.g., DAQUAR (Malinowski and Fritz, 2014), VQA (Antol et al., 2015), COCO-Q (Ren et al., 2015a), FM-IQA (Gao et al., 2015), Visual Madlibs (Yu et al., 2015), Visual7W (Zhu et al., 2016), CLEVR (Johnson et al., 2017), etc. Additionally, several video-based QA datasets have also been proposed, e.g. TGIF-QA (Jang et al., 2017), MovieFIB (Maharaj et al., 2017b), VideoQA (Zhu et al., 2017), LSMDC (Rohrbach et al., 2015), TRECVID (Over et al., 2014), MovieQA (Tapaswi et al., 2016), PororoQA (Kim et al., 2017) and MarioQA (Mun et al., 2017). However, none of these datasets provides a truly realistic, multimodal QA scenario where both visual and language understanding are required to answer a large portion of questions, either due to unrealistic video sources (PororoQA, MarioQA) or data collection strategy being more focused on either visual (MovieFIB, VideoQA, TGIF-QA) or language (MovieQA) sources. In comparison, our TVQA collection strategy takes a directly multimodal approach to construct a large-scale, real-video dataset by letting humans ask and answer questions while watching TV-show videos with associated dialogues.

Text Question Answering: The related task of text-based question answering has been extensively explored (Richardson et al., 2013; Weston et al., 2015; Rajpurkar et al., 2016; Hermann et al., 2015; Hill et al., 2015). Richardson et al. (2013) collected MCTest, a multiple choice QA dataset intended for open-domain reading comprehension.

With the same goal in mind, Rajpurkar et al. (2016) introduced the SQuAD dataset, but their answers are specific spans from long passages. Weston et al. (2015) designed a set of tasks with automatically generated QAs to evaluate the textual reasoning ability of artificial agents and Hermann et al. (2015); Hill et al. (2015) constructed the cloze dataset on top of an existing corpus. While questions in these text QA datasets are specifically designed for language understanding, TVQA questions require both vision understanding and language understanding. Although methods developed for text QA are not directly applicable to TVQA tasks, they can provide inspiration for designing suitable models.

Natural Language Object Retrieval: Language grounding addresses the task of object or moment localization in an image or video from a natural language description. For image-based object grounding, there has been much work on phrase grounding (Plummer et al., 2015; Wang et al., 2016b; Rohrbach et al., 2016) and referring expression comprehension (Hu et al., 2016; Yu et al., 2016; Nagaraja et al., 2016; Yu et al., 2017, 2018b). Recent work (Vasudevan et al., 2018) extends the grounding task to the video domain. Most recently, moment localization was proposed in (Hendricks et al., 2017; Gao et al., 2017), where the goal is to localize a short moment from a long video sequence given a query description. Accurate temporal grounding is a necessary step to answering our compositional questions.

3 TVQA Dataset

3.1 Dataset Collection

We collected our dataset on 6 long-running TV shows from 3 genres: 1) sitcoms: *The Big Bang Theory, How I Met Your Mother, Friends*, 2) medical dramas: *Grey's Anatomy, House*, 3) crime drama: *Castle*. There are in total 925 episodes spanning 461 hours. Each episode was then segmented into short clips. We first created clips every 60/90 seconds, then shifted temporal boudaries to avoid splitting subtitle sentences between clips. Shows that are mainly conversational based, e.g., *The Big Bang Theory*, were segmented into 60 seconds clips, while shows that are less cerebral, e.g. *Castle*, were segmented into 90 seconds clips. In the end, 21,793 clips were prepared for QA collection, accompanied with subtitles and aligned with transcripts to add character names. A sample clip is shown in Fig. 1.

Amazon Mechanical Turk was used for VQA collection on video clips, where workers were presented with both videos and aligned named subtitles, to encourage multimodal questions requiring both vision and language understanding to answer. Workers were asked to create questions using a compositional-question format: [What/How/Where/Why/...] ______ [when/before/after] ______. The second part of each question serves to localize the relevant video moment within a clip, while the first part poses a question about that moment. This compositional format also serves to encourage questions that require both visual and language understanding to answer, since people often naturally use visual signals to ground questions in time, e.g. *What was House saying before he leaned over the bed?* During data collection, we only used prompt words (when/before/after) to encourage workers to propose the desired, complex compositional questions. There were no additional template constraints. Therefore, most of the language in the questions is relatively free-form and complex.

Ultimately, workers pose 7 different questions for each video clip. For each question, we asked workers to annotate the exact video portion required to answer the question by marking the START and END timestamps as in Krishna et al. (2017). In addition, they provide 1 correct and 4 wrong answers for each question. Workers get paid $1.3 for a single video clip annotation. The whole collection process took around 3 months.

To ensure the quality of the questions and answers, we set up an online checker in our collection interface to verify the question format, allowing only questions that reflect our two-step format to be submitted. The collection was done in batches of 500 videos. For each harvested batch, we sampled 3 pairs of submitted QAs from each worker and checked the semantic correctness of the questions, answers, and timestamps.

3.2 Dataset Analysis

Multiple Choice QAs: Our QAs are multiple choice questions with 5 candidate answers for each question, for which only one is correct. Table 1 provides statistics of the QAs based on the first question word. On average, our questions contain 13.5 words, which is fairly long compared to other datasets. In general, correct answers tend

QType	#QA	Q. Len.	CA. Len.	WA. Len.
what	84768	13.3	4.9	4.3
who	17654	13.4	3.1	3.0
where	17777	12.5	5.2	4.8
why	15798	14.5	9.0	7.7
how	13644	14.4	5.7	5.1
others	2904	15.2	4.9	4.7
total	152545	13.5	5.2	4.6

Table 1: Statistics for different question types based on first question word. Q = question, CA = correct answer, WA = wrong answer. Length is defined as the number of words in the sentence.

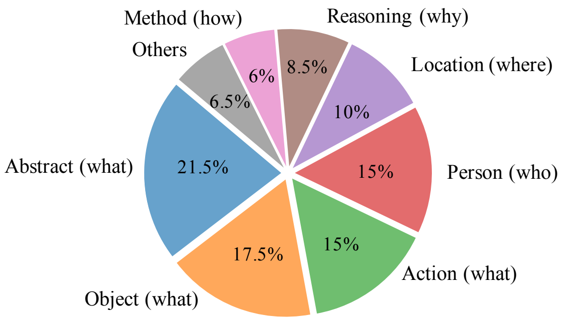


Figure 2: Distribution of question types based on answer types.

to be slightly longer than wrong answers. Fig. 2 shows the distribution of different questions types. Note "what" (Abstract, Object, Action), "who" (Person), "why" (Reasoning) and "where" (Location) questions form a large part of our data.

The negative answers in TVQA are written by human annotators. They are instructed to write false but relevant answers to make the negatives challenging. Alternative methods include sampling negative answers from other questions' correct answers, either based on semantic similarity (Das et al., 2017; Jang et al., 2017) or randomly (Antol et al., 2015; Das et al., 2017). The former is prone to introducing paraphrases of the ground-truth answer (Zhu et al., 2016). The latter avoids the problem of paraphrasing, but generally produces irrelevant negative choices. We show in Table 8 that our human written negatives are more challenging than randomly sampled negatives.

Moment Localization: The second part of our question is used to localize the most relevant video portion to answer the question. The prompt of "when", "after", "before" account for 60.03%, 30.19% and 9.78% respectively of our dataset. TVQA provides the annotated START and END timestamps for each QA. We show the annotated segment lengths in Fig. 3. We found most of the questions rely on relatively short moments (less than 15 secs) within a longer clip (60-90 secs).

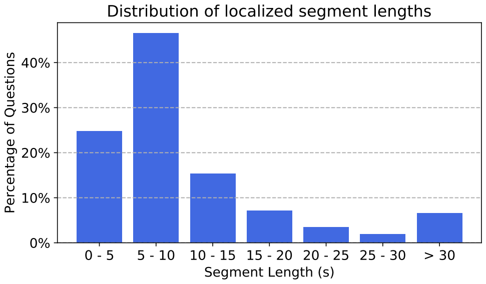


Figure 3: Distribution of localized segment lengths. The majority of our questions have timestamp localized segment with length less than 15 seconds.

Show	Genre	#Sea.	#Epi.	#Clip	#QA
BBT	sitcom	10	220	4,198	29,384
Friends	sitcom	10	226	5,337	37.357
HIMYM	sitcom	5	72	1,512	10,584
Grey	medical	3	58	1,427	9,989
House	medical	8	176	4,621	32,345
Castle	crime	8	173	4,698	32,886
Total	—	44	925	21,793	152,545

Table 2: Data Statistics for each TV show. BBT = *The Big Bang Theory*, HIMYM = *How I Met You Mother*, Grey = *Grey's Anatomy*, House = *House M.D.*, Epi = Episode, Sea. = Season

Show	Top unique nouns
BBT	game, mom, laptop, water, store, dinner, book, stair, computer, food, wine, glass, couch, date
Friends	shop, kiss, hair, sofa, jacket, counter, coffee, everyone, coat, chair, kitchen, baby, apartment
HIMYM	bar, beer, drink, job, dad, sex, restaurant, wedding, party, booth, dog, story, bottle, club, painting
Grey	nurse, side, father, hallway, scrub, chart, wife, window, life, family, chief, locker, head, surgery
House	cane, team, blood, test, brain, pill, office, pain, symptom, diagnosis, hospital, coffee, cancer, drug
Castle	gun, victim, picture, case, photo, body, murder, suspect, scene, crime, money, interrogation

Table 3: Top unique nouns in questions and correct answers.

Differences among our 6 TV Shows: The videos used in our dataset are from 6 different TV shows. Table 2 provides statistics for each show. A good way to demonstrate the difference among questions from TV shows is to show their top unique nouns. In Table 3, we present such an analysis. The top unique nouns in sitcoms (*BBT*, *Friends*, *HIMYM*) are mostly daily objects, scenes and actions, while medical dramas (*Grey, House*) questions contain more medical terms, and crime shows (*Castle*) feature detective terms. Although similar, there are also notable differences among shows in the same genre. For example, *BBT* con-

Dataset	V. Src.	QType	#Clips / #QAs	Avg. Len.(s)	Total Len.(h)	Q. Src. text	Q. Src. video	Timestamp annotation
MovieFIB (Maharaj et al., 2017a)	Movie	OE	118.5k / 349k	4.1	135	✓	-	-
Movie-QA (Tapaswi et al., 2016)	Movie	MC	6.8k / 6.5k	202.7	381	✓	-	✓
TGIF-QA (Jang et al., 2017)	Tumblr	OE&MC	71.7k / 165.2k	3.1	61.8	✓	✓	-
Pororo-QA (Kim et al., 2017)	Cartoon	MC	16.1k / 8.9k	1.4	6.3	✓	✓	-
TVQA (our)	TV show	MC	21.8k / 152.5k	76.2	461.2	✓	✓	✓

Table 5: Comparison of TVQA to various existing video QA datasets. OE = open-ended, MC = multiple-choices. Q. Src. = Question Sources, it indicates where the questions are raised from. TVQA dataset is unique since its questions are based on both text and video, with additional timestamp annotation for each of them. It is also significantly larger than previous datasets in terms of total length of videos.

Character	Top unique nouns
Sheldon	Arthur, train, Kripke, flag, flash, Wil, logo, Barry, superhero, Spock, trek, sword
Leonard	Leslie, helium, robe, Dr, team, Kurt university, key, chess, Stephen
Howard	NASA, trick, van, language, summer, letter, Mike, station, peanut, Missy
Raj	Lucy, Claire, parent, music, nothing, Isabella, bowl, sign, back, India, number
Penny	basket, order, mail, mouth, cheesecake, factory shower, pizza, cream, Alicia, waitress, ice
Amy	Dave, meemaw, tablet, birthday, monkey, coat, brain, ticket, laboratory, theory, lip, candle
Bernadette	song, sweater, wedding, child, husband, everyone, necklace, stripper, weekend, airport

Table 4: Top unique nouns for characters in BBT.

VQA source	Human accuracy on test.
Question	31.84
Video and Question	61.73
Subtitle and Question	72.88
Video, Subtitle, and Question	89.41

Table 5: Human accuracy on test set based on different sources. As expected, humans get the best performance when given both videos and subtitles.

tains "game" and "laptop" while *HIMYM* contains "bar" and "beer", indicating the different major activities and topics in each show. Additionally, questions about different characters also mention different words, as shown in Table 4.

Comparison with Other Datasets: Table 5 presents a comparison of our dataset to some recently proposed video question answering datasets. In terms of total length of videos, TVQA is the largest, with a total of 461.2 hours of videos. MovieQA (Tapaswi et al., 2016) is most similar to our dataset, with both multiple choice questions and timestamp annotation. However, their questions and answers are constructed by people posing questions from a provided plot summary, then later aligned to the video clips, which makes most of their questions text oriented.

Human Evaluation on Usefulness of Video and Subtitle in Dataset: To gain a better understanding of the roles of videos and subtitles in the our dataset, we perform a human study, asking different groups of workers to complete the QA task in settings while observing different sources (subsets) of information:

- Question only.
- Video and Question.
- Subtitle and Question.
- Video, Subtitle, and Question.

We made sure the workers that have written the questions did not participate in this study and that workers see only one of the above settings for answering each question. Human accuracy on our test set under these 4 settings are reported in Table 5. As expected, compared to human accuracy based only on question-answer pairs (Q), adding videos (V+Q), or subtitles (S+Q) significantly improves human performance. Adding both videos and subtitles (V+S+Q) brings the accuracy to 89.41%. This indicates that in order to answer the questions correctly, both visual and textual understanding are essential. We also observe that workers obtain 31.84% accuracy given question-answer pairs only, which is higher than random guessing (20%). We ascribe this to people's prior knowledge about the shows. Note, timestamp annotations are not provided in these experiments.

4 Methods

We introduce a multi-stream end-to-end trainable neural network for Multi-Modal Video Question Answering. Fig. 4 gives an overview of our model. Formally, we define the inputs to the model as: a 60-90 second video clip V, a subtitle S, a question q, and five candidate answers $\{a_i\}_{i=0}^{4}$.

4.1 Video Features

Frames are extracted at 3 fps. We run Faster R-CNN (Ren et al., 2015b) trained on the Visual

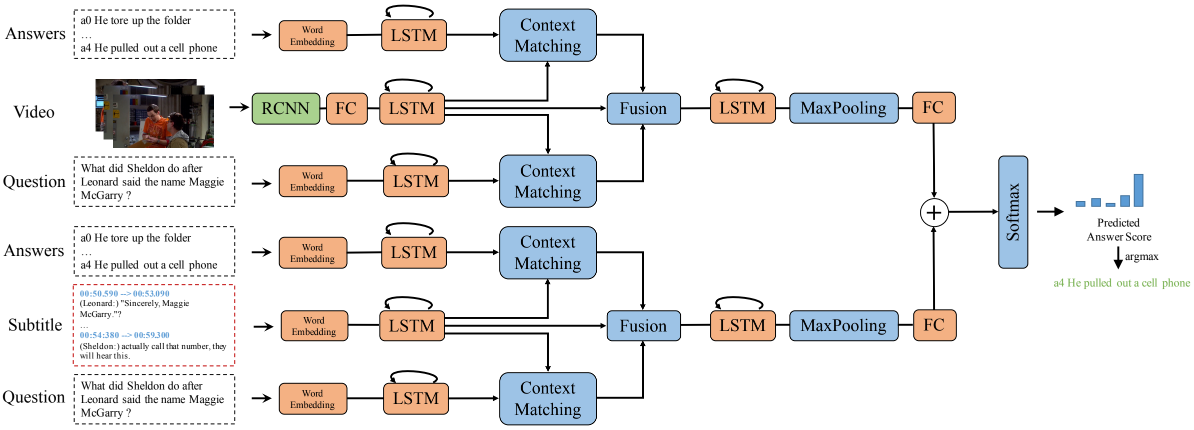
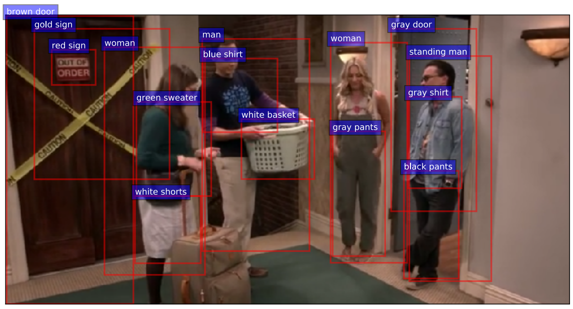


Figure 4: Illustration of our multi-stream model for Multi-Modal Video QA. Our full model takes different contextual sources (regional visual features, visual concept features, and subtitles) along with question-answer pair as inputs to each stream. For brevity, we only show regional visual features (upper) and subtitle (bottom) streams.



brown door, gold sign, red sign, woman, white shorts,
green sweater, man, blue shirt, white basket, woman,
gray pants, gray door, standing man, gray shirt, black pants

Figure 5: Faster R-CNN detection example. The detected object labels and attributes can be viewed as a description to the frame, which is potentially helpful to answer a visual question.

Genome (Krishna et al., 2017) to detect object and attribute regions in each frame. Both regional features and predicted detection labels can be used as model inputs. We also use ResNet101 (He et al., 2016) trained on ImageNet (Deng et al., 2009) to extract whole image features.

Regional Visual Features: On average, our videos contain 229 frames, with 16 detections per frame. It is not trivial to model such long sequences. For simplicity, we follow (Anderson et al., 2018; Karpathy and Fei-Fei, 2015) selecting the top-K regions[1] from each detected label across all frames. Their regional features are L2-normalized and stacked together to form our visual representation $V^{reg} \in \mathbb{R}^{n_{reg} \times 2048}$. Here n_{reg} is the number of selected regions.

Visual Concept Features: Recent work (Yin and Ordonez, 2017) found that using detected object labels as input to an image captioning system gave comparable performance to using CNN features directly. Inspired by this work, we also experiment with using detected labels as visual inputs. As shown in Fig. 5, we are able to detect rich visual concepts, including both objects and attributes, e.g. ”white basket”, which could be used to answer “What is Sheldon holding in his hand when everyone is at the door”. We first gather detected concepts over all the frames to represent concept presence. After removing duplicate concepts, we use GloVe (Pennington et al., 2014) to embed the words. The resulting video representation is denoted as $V^{cpt} \in \mathbb{R}^{n_{cpt} \times 300}$, where n_{cpt} is the number of unique concepts.

ImageNet Features: We extract the pooled 2048D feature of the last block of ResNet101. Features from the same video clip are L2 normalized and stacked, denoted as $V^{img} \in \mathbb{R}^{n_{img} \times 2048}$, where n_{img} is the number of frames extracted from the video clip.

4.2 LSTM Encoders for Video and Text

We use a bi-directional LSTM (BiLSTM) to encode both textual and visual sequences. A subtitle S, which contains a set of sentences, is flattened into a long sequence of words and GloVe (Pennington et al., 2014) is used to embed the words. We stack the hidden states of the BiLSTM from both directions at each timestep to obtain the subtitle representation $H^S \in \mathbb{R}^{n_S \times 2d}$, where n_S is the number of subtitle words, d is the hidden size of the BiLSTM (set to 150 in our experiments). Similarly, we encode question $H^q \in \mathbb{R}^{n_q \times 2d}$, candidate answers $H^{a_i} \in \mathbb{R}^{n_{a_i} \times 2d}$, and visual con-

[1] Based on cross-validation, we find K=6 to perform best.

cepts $H^{cpt} \in \mathbb{R}^{n_{cpt} \times 2d}$. n_q and n_{a_i} are the number of words in question and answer a_i, respectively. Regional features V^{reg} and ImageNet features V^{img} are first projected into word vector space using a non-linear layer with tanh activation, then encoded using the same BiLSTM to obtain the regional representations $H^{reg} \in \mathbb{R}^{n_{reg} \times 2d}$ and $H^{img} \in \mathbb{R}^{n_{img} \times 2d}$, respectively.

4.3 Joint Modeling of Context and Query

We use a context matching module and BiLSTM to jointly model the contextual inputs (subtitle, video) and query (question-answer pair). The context matching module is adopted from the context-query attention layer from previous works (Seo et al., 2017; Yu et al., 2018a). It takes context vectors and query vectors as inputs and produces a set of context-aware query vectors based on the similarity between each context-query pair.

Taking the regional visual feature stream as an example (Fig. 4 upper stream), where H^{reg} is used as context input[2]. The question embedding, H^q, and answer embedding, H^{a_i}, are used as queries. After feeding context-query pairs into the context matching module, we obtain a video-aware-question representation, $G^{reg,q} \in \mathbb{R}^{n_{reg} \times 2d}$, and video-aware-answer representation, $G^{reg,a_i} \in \mathbb{R}^{n_{reg} \times 2d}$, which are then fused with video context:

$$\begin{aligned} M^{reg,a_i} = [&H^{reg}; G^{reg,q}; G^{reg,a_i}; \\ &H^{reg} \odot G^{reg,q}; H^{reg} \odot G^{reg,a_i}], \end{aligned}$$

where $\odot$ is element-wise product. The fused feature, $M^{reg,a_i} \in \mathbb{R}^{n_{reg} \times 10d}$, is fed into another BiLSTM. Its hidden states, $U^{reg,a_i} \in \mathbb{R}^{n_{reg} \times 10d}$, are max-pooled temporally to get the final vector, $u^{reg,a_i} \in \mathbb{R}^{10d}$, for answer a_i. We use a linear layer with softmax to convert $\{u^{reg,a_i}\}_{i=0}^{4}$ into answer probabilities. Similarly, we can compute the answer probabilities given subtitle as context (Fig. 4 bottom stream). When multiple streams are used, we simply sum up the scores from each stream as the final score (Wang et al., 2016a).

5 Experiments

For evaluation, we introduce several baselines and compare them to our proposed model.

In all experiments, setup is as follows. We split the TVQA dataset into 80% training, 10% validation, and 10% testing splits such that videos and their corresponding QA pairs appear in only one split. This results in 122,039 QA pairs for training, 15,253 QA pairs for validation, and 15,253 QA pairs for testing. We evaluate each model using multiple-choice question answering accuracy.

5.1 Baselines

Longest Answer: Table 1 indicates that the average length of the correct answers is longer than the wrong ones; thus, our first baseline simply selects the longest answer for each question.

Nearest Neighbor Search: In this baseline, we use Nearest Neighbor Search (NNS) to compute the closest answer to our question or subtitle. We embed sentences into vectors using TFIDF, SkipThought (Kiros et al., 2015), or averaged GloVe (Pennington et al., 2014) word vectors, then compute the cosine similarity for each question-answer pair or subtitle-answer pair. For TFIDF, we use bag-of-words to represent the sentences, assigning a TFIDF value for each word.

Retrieval: Due to the size of TVQA, there may exist similar questions and answers in the dataset. Thus, we also implement a baseline two-step retrieval approach: given a question and a set of candidate answers, we first retrieve the most relevant question in the training set, then pick the candidate answer that is closest to the retrieved question's correct answer. Similar approaches have also been used in dialogue systems (Jafarpour and Burges, 2010; Leuski and Traum, 2011), picking the appropriate responses to an utterance from a predefined human conversational corpus. Similar to NNS, we use TFIDF, SkipThought, and GloVe vectors with cosine similarity.

5.2 Results

Table 6 shows results from baseline methods and our proposed neural model. Our main results are obtained by using full-length video clips and subtitles, without using timestamps (*w/o ts*). We also run the same experiments using the localized video and subtitle segment specified by the ground truth timestamps (*w/ ts*). If not indicated explicitly, the numbers described below are from the experiments on full-length video clips and subtitles.

Baseline Comparison: Row 1 shows results of the longest answer baseline, achieving 30.41%

[2] For visual concept features and ImageNet features, we simply replace H^{reg} with H^{cpt} or H^{img} as the context.

	Method	Video Feature	Test Accuracy	
			w/o ts	w/ ts
0	Random	-	20.00	20.00
1	Longest Answer	-	30.41	30.41
2	Retrieval-Glove	-	22.48	22.48
3	Retrieval-SkipThought	-	24.24	24.24
4	Retrieval-TFIDF	-	20.88	20.88
5	NNS-Glove Q	-	22.40	22.40
6	NNS-SkipThought Q	-	23.79	23.79
7	NNS-TFIDF Q	-	20.33	20.33
8	NNS-Glove S	-	23.73	29.66
9	NNS-SkipThought S	-	26.81	37.87
10	NNS-TFIDF S	-	49.94	51.23
11	Our Q	-	43.34	43.34
12	Our V+Q	img	42.67	43.69
13	Our V+Q	reg	42.75	44.85
14	Our V+Q	cpt	43.38	45.41
15	Our S+Q	-	63.14	66.23
16	Our S+V+Q	img	63.57	66.97
17	Our S+V+Q	reg	63.19	67.82
18	Our S+V+Q	cpt	**65.46**	**68.60**

Table 6: Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation. Human performance without timestamp annotation is reported in Table 5.

(compared to random chance at 20%). As expected, the retrieval-based methods (row 2-4) and the answer-question similarity based methods (row 5-7) perform rather poorly, since no contexts (video or subtitle) are considered. When using subtitle-answer similarity to choose correct answers, Glove, SkipThought, and TFIDF based approaches (row 8-10) all achieve significant improvement over question-answer similarity. Notably, TFIDF (row 10) answers 49.94% of the questions correctly. Since our questions are raised by people watching the videos, it is natural for them to ask questions about specific and unique objects/locations/etc., mentioned in the subtitle. Thus, it is not surprising that TFIDF based similarity between answer and subtitle performs so well.

Variants of Our Model: Rows 11-18 show results of our model with different contextual inputs and features. The model that only uses question-answer pairs (row 11) achieves 43.34% accuracy. Compared to the subtitle model (row 15), adding video as additional sources (row 16-18) improves performance. Interestingly, adding video to the question only model (row 11) do not work as well (row 12-14). Our hypothesis is that the video feature streams may be struggling to learn models for answering textual questions, which degrades their ability to answer visual questions. Overall, the best performance is achieved by using all the contextual sources, including subtitles and videos (using concept features, row 18).

	Q	S+Q	V+Q			S+V+Q		
			img	reg	cpt	img	reg	cpt
what (55.62%)	44.11	62.29	44.96	45.93	47.44	63.88	65.28	66.05
who (11.55%)	36.55	68.33	35.75	34.85	34.68	67.76	67.20	67.99
where (11.67%)	42.58	56.97	47.13	48.43	48.20	61.97	63.71	61.46
how (8.98%)	41.17	71.97	41.17	42.41	40.95	71.17	70.80	71.53
why (10.38%)	45.23	78.65	46.05	45.36	45.48	78.33	77.13	78.77
other (1.80%)	36.50	74.45	37.23	36.50	33.58	73.72	72.63	74.09
all (100%)	42.77	65.15	43.78	44.40	45.03	66.44	67.17	67.70

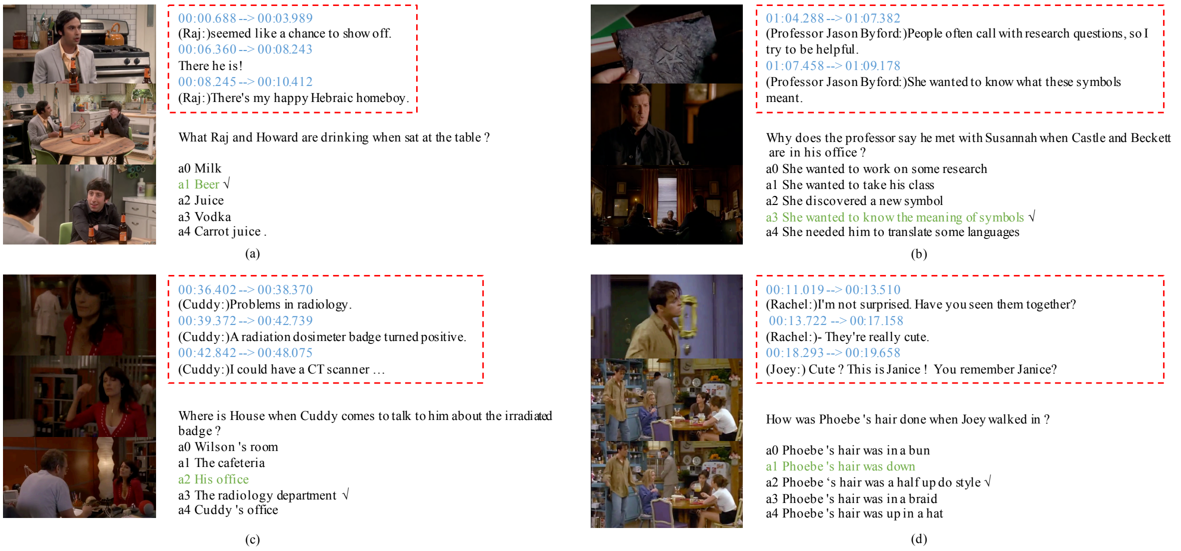
Table 7: Accuracy of each question type using different models (w/ ts) on TVQA Validation set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features. The percentage of each question type is shown in brackets.

Comparison with Human Performance: Human performance without timestamp annotation is shown in Table 5. When using only questions (Table 6 row 11), our model outperforms humans (43.34% *vs* 31.84%) as it has access to all statistics of the questions and answers. When using videos or subtitles or both, humans perform significantly better than the models.

Models with Timestamp Annotation: Columns under *w/o ts* and *w/ ts* show a comparison between the same model using full-length videos/subtitles and using timestamp localized videos/subtitles. With timestamp annotation, the models perform consistently better than their counterpart without this information, indicating that localization is helpful for question answering.

Accuracy for Different Question Types: To gain further insight, we examined the accuracy of our models on different question types on the validation set (results in Table 7), all models using timestamp annotation. Compared to S+Q model, S+V+Q models get the most improvements on "what" and "where" questions, indicating these questions require additional visual information. On the other hand, adding video features did not improve S+Q performance on questions relying more on textual reasoning, e.g., "how" questions.

Human-Written Negatives vs. Randomly-Sampled Negatives For comparison, we create a new answer set by replacing the original human written negative answers with randomly sampled negative answers. To produce relevant negative answers, for each question, negatives are sampled (from the other QA pairs) within the same show.



00:00.688 --> 00:03.989
(Raj:)seemed like a chance to show off.
00:06.360 --> 00:08.243
There he is!
00:08.245 --> 00:10.412
(Raj:)There's my happy Hebraic homeboy.

What Raj and Howard are drinking when sat at the table ?

a0 Milk
a1 Beer √
a2 Juice
a3 Vodka
a4 Carrot juice .

(a)

01:04.288 --> 01:07.382
(Professor Jason Byford:)People often call with research questions, so I try to be helpful.
01:07.458 --> 01:09.178
(Professor Jason Byford:)She wanted to know what these symbols meant.

Why does the professor say he met with Susannah when Castle and Beckett are in his office ?
a0 She wanted to work on some research
a1 She wanted to take his class
a2 She discovered a new symbol
a3 She wanted to know the meaning of symbols √
a4 She needed him to translate some languages

(b)

00:36.402 --> 00:38.370
(Cuddy:)Problems in radiology.
00:39.372 --> 00:42.739
(Cuddy:)A radiation dosimeter badge turned positive.
00:42.842 --> 00:48.075
(Cuddy:)I could have a CT scanner …

Where is House when Cuddy comes to talk to him about the irradiated badge ?
a0 Wilson 's room
a1 The cafeteria
a2 His office
a3 The radiology department √
a4 Cuddy 's office

(c)

00:11.019 --> 00:13.510
(Rachel:)I'm not surprised. Have you seen them together?
00:13.722 --> 00:17.158
(Rachel:)- They're really cute.
00:18.293 --> 00:19.658
(Joey:) Cute ? This is Janice ! You remember Janice?

How was Phoebe 's hair done when Joey walked in ?

a0 Phoebe 's hair was in a bun
a1 Phoebe 's hair was down
a2 Phoebe ‘s hair was a half up do style √
a3 Phoebe 's hair was in a braid
a4 Phoebe 's hair was up in a hat

(d)

Figure 6: Example predictions from our best model. Top row shows correct predictions, bottom row shows failure cases. Ground truth answers are in green, and the model predictions are indicated by ✓. Best viewed in color.

Method	N.A. Src.	Video	Val Accuracy	
		Feature	w/o ts	w/ ts
V+Q	Rand	cpt	84.64	85.01
S+Q	Rand	-	90.94	90.72
S+V+Q	Rand	cpt	91.55	92.00
V+Q	Human	cpt	43.03	45.03
S+Q	Human	-	62.99	65.15
S+V+Q	Human	cpt	64.70	67.70

Table 8: Accuracy on TVQA validation set with negative answers collected using different strategies. Negative Answer Source (N.A. Src.) indicates the collection method of the negative answers. Q = Question, S = Subtitle, V = Video, cpt = visual concept features, ts = timestamp annotation. All the experiments are conducted using the proposed multi-stream neural model.

Results are shown in Table 8. Performance on randomly sampled negatives is much higher than that of human written negatives, indicating that human written negatives are more challenging.

Qualitative Analysis: Fig. 6 shows example predictions from our S+V+Q model (row 18) using full-length video and subtitle. Fig. 6a and Fig. 6b demonstrate its ability to solve both grounded visual questions and textual reasoning question. Bottom row shows two incorrect predictions. We found that wrong inferences are mainly due to incorrect language inferences and the model's lack of common sense knowledge. For example, Fig. 6c, the characters are talking about radiology, the model is distracted to believe they are in the radiology department, while Fig. 6d shows a case of questions that need common sense to answer, rather than simply textual or visual cues.

6 Conclusion

We presented the TVQA dataset, a large-scale, localized, compositional video question answering dataset. We also proposed two QA tasks (with/without timestamps) and provided baseline experiments as a benchmark for future comparison. Our experiments show both visual and textual understanding are necessary for TVQA.

There is still a significant gap between the proposed baselines and human performance on the QA accuracy. We hope this novel multimodal dataset and the baselines will encourage the community to develop stronger models in future work. To narrow the gap, one possible direction is to enhance the interactions between videos and subtitles to improve multimodal reasoning ability. Another direction is to exploit human-object relations in the video and subtitle, as we observe that a large number of questions involve such relations. Additionally, temporal reasoning is crucial for answering the TVQA questions. Thus, future work also includes integrating better temporal cues.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and discussions. This research is supported by NSF Awards #1633295, 1562098, 1405822 and a Google Faculty Research Award, Bloomberg Data Science Research Grant, and ARO-YIP Award #W911NF-18-1-0336. The views contained in this article are those of the authors and not of the funding agency.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and vqa. *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. *ICCV*, pages 2425–2433.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. *CVPR*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *CVPR*, pages 248–255.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. 2017. Tall: Temporal activity localization via language query. *ICCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. *ICCV*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *ICLR*.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. *CVPR*.

Sina Jafarpour and Chris J.C. Burges. 2010. Filter, rank, and transfer the knowledge: Learning to chat. Technical report.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *CVPR*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. In *IJCAI*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Fei fei Li, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. *ICCV*, pages 706–715.

Anton Leuski and David R. Traum. 2011. Npceditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32:42–56.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017a. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *CVPR*.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017b. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *CVPR*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. Marioqa: Answering questions by watching gameplay videos. In *ICCV*.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *ECCV*.

Paul Over, Jon Fiscus, Greg Sanders, David Joy, Martial Michel, George Awad, Alan Smeaton, Wessel Kraaij, and Georges Quénot. 2014. Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*, page 52.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

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

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *NIPS*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015b. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39:1137–1149.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *CVPR*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *ICLR*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. *CVPR*.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. *CVPR*.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qi Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016a. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016b. Learning deep structure-preserving image-text embeddings. In *CVPR*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *ICLR*.

Xuwang Yin and Vicente Ordonez. 2017. Obj2text: Generating visually descriptive language from object layouts. *EMNLP*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018a. Qanet: Combining local convolution with global self-attention for reading comprehension. *ICLR*.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018b. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank image generation and question answering. *ICCV*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speakerlistener-reinforcer model for referring expressions. In *CVPR*.

Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *IJCV*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *CVPR*.