Explore Multi-Step Reasoning in Video Question Answering

Xiaomeng Song, Yucheng Shi, Xin Chen, Yahong Han
School of Computer Science and Technology, Tianjin University, Tianjin, China
{songxiaomeng,yucheng,chenxin13,yahong}@tju.edu.cn

ABSTRACT

Video question answering (VideoQA) always involves visual reasoning. When answering questions composing of multiple logic correlations, models need to perform multi-step reasoning. In this paper, we formulate multi-step reasoning in VideoQA as a new task to answer compositional and logical structured questions based on video content. Existing VideoQA datasets are inadequate as benchmarks for the multi-step reasoning due to limitations such as lacking logical structure and having language biases. Thus we design a system to automatically generate a large-scale dataset, namely SVQA (Synthetic Video Question Answering). Compared with other VideoQA datasets, SVQA contains exclusively long and structured questions with various spatial and temporal relations between objects. More importantly, questions in SVQA can be decomposed into human readable logical tree or chain layouts, each node of which represents a sub-task requiring a reasoning operation such as comparison or arithmetic. Towards automatic question answering in SVQA, we develop a new VideoQA model. Particularly, we construct a new attention module, which contains spatial attention mechanism to address crucial and multiple logical sub-tasks embedded in questions, as well as a refined GRU called ta-GRU (temporal-attention GRU) to capture the long-term temporal dependency and gather complete visual cues. Experimental results show the capability of multi-step reasoning of SVQA and the effectiveness of our model when compared with other existing models.

KEYWORDS

Video Question Answering; Multi-Step Reasoning

ACM Reference Format:
Xiaomeng Song, Yucheng Shi, Xin Chen, Yahong Han. 2018. Explore Multi-Step Reasoning in Video Question Answering. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240563

1 INTRODUCTION

Video question answering (VideoQA) targets to answer natural language questions based on video content, which always involves visual reasoning. Visual reasoning is a perceptual ability to deal with object interactions, attribute comparison, or arithmetic problems [12]. When assemble multiple logical operations to answer questions, we call this process multi-step reasoning [6]. Multi-step reasoning is more challenging compared with single-step reasoning, and achieving such a high-level ability is a closer step toward artificial intelligence.

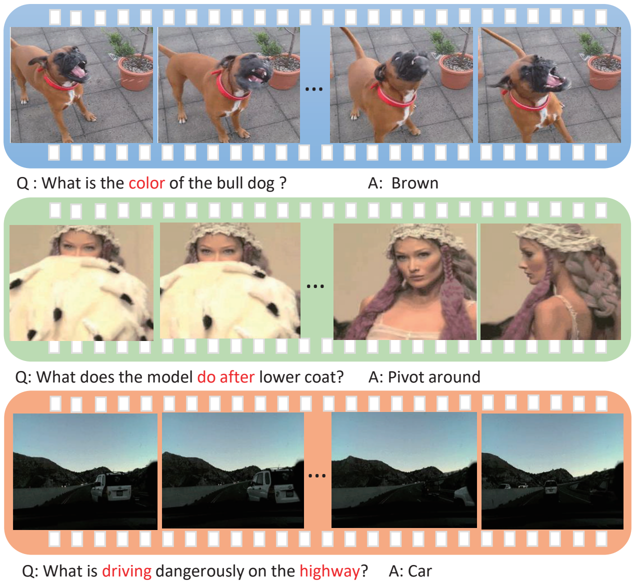

Q : What is the color of the bull dog ?
A: Brown
Q: What does the model do after lower coat?
A: Pivot around
Q: What is driving dangerously on the highway?
A: Car

Figure 1: Three examples from existing VideoQA datasets that illustrate the limitations: The top question only needs visual recognition without reasoning temporal dependency between frames; the question in the middle row lacks compositional logical structure, which only addresses single-step reasoning; the bottom question contains language bias that shadows the role of visual reasoning.

There exist prior efforts [8, 12, 19, 24] that tried to explore multi-step reasoning in image question answering (ImageQA). As the natural extension of images, videos contain richer information with additional temporal and dynamic characteristics, which brings up more challenges and difficulties. In order to explore multi-step reasoning in VideoQA, we formulate it as a new task that answers logical compositional questions based on videos. Existing VideoQA datasets [15, 17, 17, 36, 41] are inadequate to serve as benchmarks for this task due to some limitations: First, answering questions in some datasets [17, 36, 41] only requires basic visual recognition. As illustrated in the top row of Fig. 1, the answer 'brown' can be extracted directly by observing single frame without considering temporal structures of videos. Second, questions in some datasets [15, 18, 34] lack compositional logical structures, thus the answering only requires single-step reasoning, as shown in the middle row of Fig. 1. Finally, correctly answering questions in some datasets

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240563

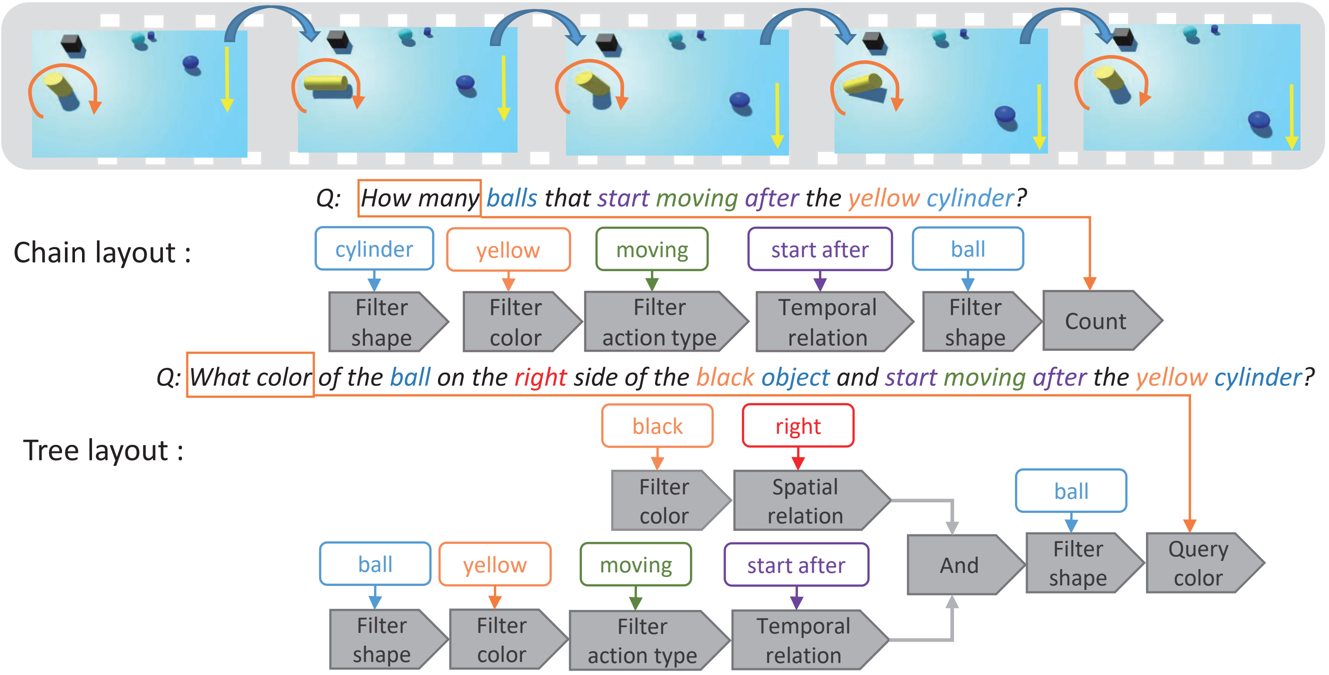


Figure 2: An example from SVQA. The top part is our synthetic video, which contains specific and various spatial and temporal relations between static or dynamic objects. We also display our questions (italic sentence) and their chain or tree layouts in the bottom, demonstrating that our questions are compositional logically structured. Specifically, each gray node represents a logical operation, and words with different color in the question correspond to the rectangle with the same color, which can be seen as the parameter of each operation. The outermost node of each layout represents the question category, which corresponds to the orange box in the question.

[10, 17] can only rely on exploiting the correlations between questions and answers, while ignoring video content completely. As shown in the bottom row of Fig. 1, the answer 'car' can be obtained straightforwardly from particular words in the question like 'driving' or 'highway'. As one kind of language bias, this drawback deviates from the original goal of VideoQA.

To provide a reasonable benchmark for this task, we develop a system to automatically generate a large-scale synthetic VideoQA dataset called SVQA (Synthetic Video Question Answering). Our design principle is consistent with the multi-step reasoning process, that solving our questions requires multiple logical operations. Compared with traditional VideoQA dataset, questions in SVQA are exclusively long and logically structured, containing various spatial and temporal relations between objects, such as their relative positions or action orders. An example of SVQA is shown in Fig. 2. More importantly, questions in SVQA can be decomposed into human readable logical forms, i.e., logical tree or chain layouts, as shown in the bottom half of Fig. 2. Each node of layouts represents a sub-task requiring to perform a logical reasoning operation. We categorize questions based on the outermost operation of their layouts. To ensure the reasonability of our dataset and eliminate above limitations, we design quality control and answer balance mechanisms in our system. We choose synthetic videos because they contain clear spatial and temporal relations between objects which may be hard to collect in real-world videos. Thus SVQA is suitable as a test bed for multi-step reasoning. Besides, annotating questions with multiple logical operations (see examples in Fig. 2) is labour consuming, whereas in SVQA such questions are automatically devised and generated.

Towards automatic question answering and visual reasoning, some earlier works [8, 11, 19] take advantage of neural modular network [1]. They design a set of neural modules to solve each sub-task embedded in questions respectively and integrate them based on the layout of the whole question. However, the layouts they use are either obtained by hand-crafted rules and annotations [8, 11], which is labor consuming and not easy to be generalized, or by prediction from modular network [19], which may be inaccurate and leads to inferior performance. Recent works propose general frameworks [20, 24, 32], but they only focus on ImageQA without exploiting temporal and dynamic characteristics videos possess.

To handle this task, we propose a new VideoQA model. Our model takes advantage of inherent spatial-temporal structure of videos without decomposing questions into logical layouts at first. We notice that most common and important sub-tasks embedded in questions can be solved by spatial attention mechanism. For example, the 'filter shape' operation is to locate corresponding objects based on the given attributes such as 'ball'. It is similar to the goal of attention [29] that forces models selectively focus on relevant regions with semantic information as guidance. It is the same with other logical operations like the 'query color' or 'count', which all need to find certain objects based on descriptions in the question first. When certain objects are located, models can perform basic visual recognition to identify or compare their attributes. From this insight, to handle multiple reasoning operations, we design a novel attention module to employ the spatial attention. We also construct a refined GRU (Gated Recurrent Unit) called ta-GRU (temporal-attention GRU) in our attention module. As an extension structure of Recurrent Neural Networks (RNN), GRU [3] is widely used in sequence encoding and decoding tasks [16, 27, 30, 33]. However, it shows poor performance when capture long-term temporal dependency. To address this problem, we refine the traditional GRU by associating temporal attention with its hidden state transfer process, which strengthens the long-term temporal dependency, thus can gather more complete temporal visual cues. The details are discussed in Section 4.

Our contribution can be summarized as three-folds: (1) We explore multi-step reasoning in VideoQA by formulating it as a new task, which targets to answer compositional logical structured questions based on video content; (2) We develop a system to automatically generate a large-scale VideoQA dataset called SVQA. Extensive experimental results show it can serve as a benchmark for multi-step reasoning in VideoQA; (3) We propose a new model for VideoQA on SVQA. Particularly, we construct a novel attention module containing spatial attention mechanism and ta-GRU to address multi-step reasoning. Experiment results show the effectiveness of it compared with existing models. [1]

2 RELATED WORK

In recent years, VideoQA has attracted a large number of researchers from computer vision and multimedia field [10, 15, 34, 36]. Many prior works tried to explore visual reasoning in VideoQA. For example, Jang *et al.* constructed a dataset called TGIF-QA [10] and defined three new tasks to discuss the spatial-temporal reasoning. Mun *et al.* [18] proposed a dataset with different levels of reasoning complexity to measure the temporal relations in videos. The work of Zhu *et al.* [41] also focused on temporal reasoning, i.e., inferring the past, describing the present and predicting the future of the video. We explore visual reasoning from another perspective that we focus on multi-step reasoning in VideoQA.

There have been benchmarks proposed towards multi-step reasoning in ImageQA [12, 13, 32]. However, compared with videos, images lack temporal and dynamic characteristics, thus reducing the complexity of multi-step reasoning to some extent. To the best of our knowledge, multi-step reasoning in VideoQA is unexplored, it is partly due to the lack of well-designed benchmarks as CLEVR for evaluation. Although there have been many VideoQA datasets [22, 25, 34, 39, 40] with diverse question forms and video sources, they are inadequate due to some limitations such as lacking compositional logic structure or having language biases. To fill this gap, we design an automatic data construct system to generate a large-scale dataset called SVQA. We use synthetic videos instead of real-world ones because it is hard to construct questions we need based on diverse visual contents. Besides, synthetic videos contain clearer and more specific spatial and temporal relations between objects, which is suitable as a test bed for multi-step reasoning. Our design principle that solving our questions requires multiple logical operations is consistent with the multi-step reasoning process.

Some methods have been put forward to investigate multi-step reasoning in ImageQA. Some of them [8, 11, 19] implement neural module networks [1], which are designed according to the inherent compositional logical structure of questions. Others constructed general frameworks [9, 20, 24, 32] to address this task. However, they all focus on ImageQA without considering temporal information which is crucial for videos as previous works mentioned [28, 38]. We propose a general reasoning model without decomposing questions into logical layouts first and utilize the spatial-temporal structure of videos. Specifically, we employ spatial attention mechanism to perform crucial and multiple logical operations and ta-GRU to capture the long-term temporal dependency by associating temporal attention with its hidden state transfer process.

We notice Zhao *et al.* [40] also proposed a refined GRU. Ours differs from them in two ways: (1) They wait until all hidden states are output and feed them into an additional bidirectional GRU to calculate attention weights. While we directly calculate attention weights as soon as each one is output, which can reduce the time complexity and parameters; (2) They only apply attention on the current output hidden state and its previous one, while we apply attention on the current one and all previous upgraded ones, which can gather more complete temporal visual cues.

3 THE PROPOSED BENCHMARK SVQA

In this section we design a system to automatically generate a large-scale dataset SVQA. Our system consists of four processes: video generation, question-answer (QA) generation, quality control, and answer balance. The constructed SVQA contains 12000 synthetic videos and around 120K QA pairs in total. Specifically, videos are generated from Unity3D and QA pairs are generated from question templates automatically, which is easy to expand its scale without large manpower consumption. The answer balance and quality control mechanisms in our system eliminate the language biases and ensure the reasonability of SVQA.

3.1 Video Generation

Our system calls APIs of Unity3D to customize the generation of videos. Each video in our dataset depicts three to eight static or dynamic 3D geometries in consistent scene. Each geometry has three basic attributes: shape, size and color. Dynamic geometries possess extra action type and action direction attributes. After generating one video, the attributes and position of each geometry inside are recorded in a JSON file, which serves as a source to generate video-specific QA pairs. The uniqueness of content within each video is ensured. We also guarantee the rationality of visual information, i.e., there is no collision or overlap between geometries, nor will a geometry move out of the scene. The duration of actions are also limited to the length of the video. These can eliminate ambiguity of spatial and temporal information in video content. Since visual content in our videos contain various and specific spatial and temporal relations between objects, such as their relative positions or action orders, it enables us to construct compositional logical structured questions which require multi-step reasoning.

[1] SVQA dataset and the code of our model are available in *https://svqa-founder.github.io/SVQA/*

Table 1: Statistics of our SVQA dataset

Question Category		Train	Val	Test
Count		19320	2760	5520
Exist		6720	960	1920
Query	Color	7560	1056	2160
	Size	7560	1056	2160
	Action Type	6720	936	1920
	Direction	7560	1056	2160
	Shape	7560	1056	2160
Integer Comparison	More	2520	600	720
	Equal	2520	600	720
	Less	2520	600	720
Attribute Comparison	Color	2520	216	720
	Size	2520	216	720
	Action Type	2520	216	720
	Direction	2520	216	720
	Shape	2520	216	720
Total QA pairs		83160	11760	23760
Total Videos		8400	1200	2400

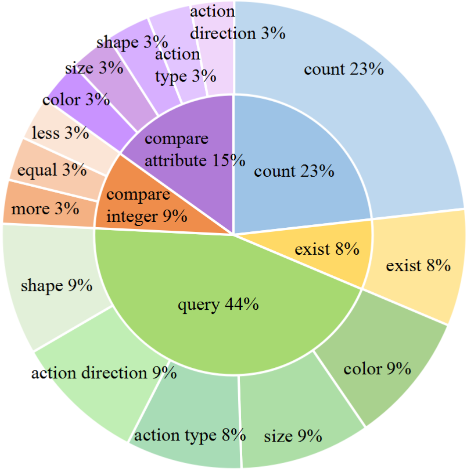


Figure 3: Distribution of each question category in our SVQA dataset.

3.2 QA Generation

Our questions are generated by predefined question templates, which are similar to natural language sentences with a few blanks. For instance, 'Is there any $\langle obj1 \rangle$ that locate to the $\langle position \rangle$ of the $\langle obj2 \rangle$?', where $\langle obj1 \rangle$, $\langle obj2 \rangle$ and $\langle position \rangle$ are blanks to be filled in. Specifically, $\langle obj1 \rangle$ and $\langle obj2 \rangle$ are filled with descriptions comprised of randomly combined attributes, which represent a geometry in the video. And $\langle position \rangle$ is filled with a randomly chosen azimuth noun. When the question is complemented, the answer can be directly extracted from the recorded information in JSON files. We also add several expressions for each question template to increase the linguistic diversity, which reduces the language regularity that can be remembered by the model. Syntax errors of questions are corrected by *Grammarly*, an open source software.

Our design principle that solving compositional logical structured questions is consistent with the multi-step reasoning process. The generated questions are compositional logically structured and exclusively long with an average length of 20 words. Each geometry within the question may be modified by three or more attributes or relations to discriminate the exact one, i.e., the content of questions contain various spatial and temporal relations. And each question can be decomposed into human readable logical chain or tree layout, where each node represents a logical reasoning operation such as comparison or arithmetic. We further categorize questions by the outermost operation they need for detailed evaluation. The statistics of SVQA is illustrated in Table. 1.

3.3 Quality Control

The automatically generated questions may address some problems. For example, *'What is the action type of the big blue ball?'* would arise *ambiguous reference* problem if there are many big blue balls in the video, or *redundant information* problem if 'big' or 'blue' can be removed without changing the questions' meaning. To prevent these problems from happening, we design an automatic quality control mechanism. First, attributes of one geometry are random combined to form all candidate descriptions. Second, each description is checked if it is the most concise one and only refers to one geometry in the video. The one that satisfies two conditions can be picked to fill in the blank. Another problem called *self-evident*, which means the answer can be straightforwardly extracted from the question as one language bias we mentioned in Section 1, such as *'What the color of the big blue cube?'*. To avoid this problem, the attribute that questions ask about are filtered out from all candidate descriptions before filling in the blanks.

3.4 Balance Answer

There is another kind of language bias that would impede objective evaluation. As mentioned in [37], some models can achieve high accuracy by simply remembering the most 'familiar' answer to certain kind of questions without using any visual information. Thus we employ answer balance mechanism to ensure the uniformity of answer distribution. Specifically, we calculate the ratio of each question category as shown in Fig. 3 by statistic methods and use random sampling to force the answer in each question category to be significantly balanced (except for the 'count' category). Since each video contains a limited number of geometries while there are a large number of attributes, the number of geometries that satisfy the questions of count category is always one or two. Under this circumstance, we don't force the answer distribution of count category to be uniform.

4 THE PROPOSED MODEL

We propose a new model towards the task of multi-step reasoning in VideoQA. The overview of our model is illustrated in Fig. 4. It is composed of four components: question embedding module, video embedding module, attention module, and answer decoder module. Specifically, the question embedding module extracts question

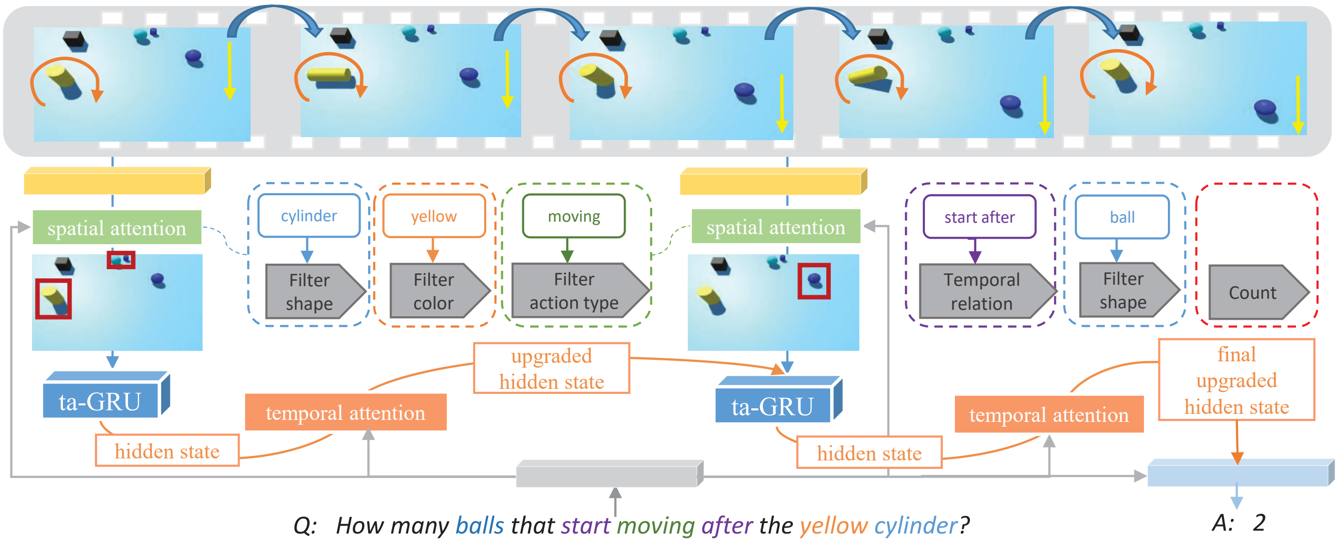


Figure 4: The overview of our model. The yellow cuboid represents video embedding module which extracts video representations. The gray cuboid represents question embedding module which extracts question representations. The wathet cuboid represents answer decoding module which generates answers. The middle main part is the attention module, containing spatial attention applying on video representations to handle each logical sub-task embedded in questions, and ta-GRU with temporal attention mechanism on its hidden state transfer process to gather more complete temporal visual cues and to capture long-term temporal dependency.

representations, the video embedding module extracts video representations, the attention module which contains spatial attention and ta-GRU is the core part of our model to conduct reasoning and the answer decoder module is a decoder to generate answers. We describe each component in detail below.

4.1 Question Embedding Module

Each question in SVQA is comprised of variable-length sequence of words (w_1,w_2,. . .,w_n). We represent each word as an one-hot vector in the size of the vocabulary first, then embed them into a semantic space through an embedding matrix($\in R^{300}$). Because questions in SVQA have less word variation compared with those based on real-world videos, we use the random initialization matrix instead of the pretrained ones. We feed embedding vectors into a 1024-dimension GRU and pick the last hidden state as the question representation $q_n(\in R^{1024})$.

4.2 Video Embedding Module

Because the movements in videos are successive and adjacent frames have little change, using all frames is time-consuming. Thus we divide each video into clips (segments) of 16 frames, with 80% overlap between successive clips (segments). This filters out redundant content and preserves the important visual information. We extract feature from each clip and aggregate features of all clips from one video to form a sequential video representation.

We first resize each frame into the size of 224 × 224. Then we employ C3D [26] which pre-trained on Sports1M [14] to get clip-level dynamic representations. We also employ ResNet-152 [7] that pre-trained on ImageNet2012 classification dataset [23] to get clip-level static representations. Specifically, we aggregate the output $v_i^d(\in R^{4\times4\times512})$ from *pool5* layer in C3D of each clip to form a dynamic video representation in sequence form $(v_1^d,\dots,v_n^d)$, where n is the total number of clips in a video. And we sample the center frame of each clip and use the output $v_i^s(\in R^{2048})$ from *pool5* layer in ResNet-152 as the clip-level static representation. Then we gather all clip-level static representations from one video to form a sequential static video representation $(v_1^s,\dots,v_n^s)$.

4.3 Attention Module

First we employ spatial attention to the dynamic visual representation v_i^d which preserve the spatial information of videos. Specifically, we use the question representation q_n to compute a spatial attention weight s_{ij} for every spatial region j in v_i^d and assign it to corresponding region. Then we add them all through spatial dimension to form $v_i(\in R^{512})$. The operation is formulated as:

$$\begin{aligned} a_{ij} &= W^T tanh(W_v v_{ij}^d + W_q q_n) \\ s_{ij} &= exp(a_{ij}) / \sum_{j=1}^{4\times4} exp(a_{ij}) \\ v_i &= \sum_{j=1}^{4\times4} s_{ij} v_{ij}^d \end{aligned} \quad (1)$$

where W_v,W_q and W are trainable parameters. W_v and W_q are used to transform both the visual and the question representations to the same attention dimension space. This spatial attention mechanism

Table 2: Accuracies of all models on each question category of SVQA (Type and Dir are abbreviations of action type and action direction)

	Exist	Count	Integer Comparison			Attribute Comparison					Query					All
			More	Equal	Less	Color	Size	Type	Dir	Shape	Color	Size	Type	Dir	Shape	
Chance	50.00	22.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	12.50	50.00	50.00	25.00	33.33	33.33
Q-Type	50.00	23.73	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	12.50	50.00	50.00	25.00	33.33	33.33
GRU-Q	52.92	32.41	**75.14**	56.39	57.81	47.73	52.56	53.12	53.55	51.56	12.27	51.07	48.65	25.23	32.70	39.95
GRU+AVG	51.77	33.18	59.66	54.12	59.38	52.27	50.00	51.13	53.27	47.58	19.78	51.91	53.33	28.26	38.29	41.43
2GRU	53.54	35.02	68.18	53.70	56.10	54.12	51.28	51.70	52.70	47.86	19.59	53.50	58.38	34.79	38.34	41.85
E-VQA [36]	53.59	35.88	72.59	51.99	61.08	51.56	51.14	50.99	49.72	49.86	23.27	62.97	62.14	41.18	43.33	42.83
E-SA [36]	52.14	34.02	68.89	54.83	55.68	52.67	55.54	52.41	53.41	49.43	17.35	54.90	54.32	38.20	39.09	43.99
ST-VQA-Tp [10]	51.46	32.54	58.46	50.39	53.52	49.74	54.56	53.12	51.95	50.39	21.23	53.81	55.70	36.08	40.60	40.47
r-STAN [40]	51.64	32.99	68.37	53.28	52.70	50.13	53.32	52.72	51.63	48.72	20.34	55.37	54.92	35.29	37.30	40.53
Unified-Att [31]	54.17	35.23	67.05	52.98	50.00	51.85	54.26	50.14	50.28	49.43	22.01	53.73	56.51	34.70	38.20	41.69
Ours(S)	51.70	36.30	72.70	54.80	58.60	52.20	53.60	52.70	52.98	52.25	**29.00**	53.97	55.69	38.10	**46.30**	43.10
Ours(T)	**54.60**	36.56	73.00	57.30	57.67	53.80	53.40	**54.80**	55.10	52.40	22.00	54.80	55.50	41.69	42.90	44.20
Ours(S+T)	52.03	**38.20**	74.28	**57.67**	**61.60**	**55.96**	**55.90**	53.40	**57.50**	**52.98**	23.39	**63.30**	**62.90**	**43.20**	41.69	**44.90**

can force model to selectively focus on semantic-relevant regions, especially locate the objects described in questions based on their attributes and relations.

We embed each generated $v_i (\in R^{512})$ and corresponding static visual representation $v_i^s (\in R^{2048})$ into the same 512-dimension space, and concatenate them to form $V_i (\in R^{1024})$. Since v_i contains semantic-relevant regional information and v_i^s contains global information, complete visual information is obtained by combining them. Then we feed them into ta-GRU, a refined GRU whose hidden state transfer process is associated with temporal attention to strengthen long-term temporal dependency. In other words, we replace the hidden state h_t which participate in the traditional GRU process with the new hidden state $h_t^{'}$ generated by temporal attention, which we called upgraded hidden state.

Specifically, once the V_i is input into ta-GRU at time i, we will get a hidden state $h_i (\in R^{1024})$. Then we use question representation q_n to calculate attention weights over it and all previous upgraded hidden states $(h_1^{'}, \ldots, h_{i-1}^{'})$. The operation is formulated as:

$$\begin{aligned} b_m &= U^T tanh(U_v h_m + U_q q_n) \qquad m = i \\ b_m &= U^T tanh(U_v h_m^{'} + U_q q_n) \qquad 1 \leq m \leq i-1 \\ t_m &= exp(b_m) / \sum_{m=1}^{i} exp(b_m) \end{aligned} \tag{2}$$

where the U_v,U_q and U are trainable parameters. U_v and U_q are used to transform the hidden state and the question representation to a same dimension space. Then we obtain the upgraded hidden state $h_i^{'}$ by weighted sum operation as the equation shown below:

$$h_i^{'} = \sum_{l=1}^{i-1} t_l h_l^{'} + t_i h_i \tag{3}$$

Since upgraded hidden state fuses all previous temporal and semantic information, transferring it instead of the traditional one through the recurrent process of GRU can facilitate the long-term temporal dependency and gather more complete visual cues.

4.4 Answer Decoder Module

The goal of this module is to select an answer from a vocabulary of words. The final hidden state $h_n^{'} (\in R^{1024})$ of ta-GRU is catenated with the question representation $q_n (\in R^{1024})$. Then we feed this concatenated vector$(\in R^{2048})$ into a one-layer perceptron to get the answer probability distribution, i.e., a confidence vector $o (\in R^{\alpha})$, where α is the size of the vocabulary. The final answer A is obtained as shown below:

$$\begin{aligned} o &= softmax(W_s^T [h_n^{'} \circ q_n]) \\ i &= argmax(o) \\ A &= y_i \end{aligned} \tag{4}$$

where W_s is a trainable parameter, y is the answer set and $[\circ]$ denotes a concatenate operation. We train this module by minimizing the softmax loss.

5 EXPERIMENTS

First, we implement some baseline methods on SVQA to suggest it eliminates the limitations which we discuss in Section 1. Then we compare our model with other VideoQA models on SVQA and conduct ablation studies to show its effectiveness. Finally, we employ our model on FrameQA task of TGIF-QA to suggest its generalization ability. We use the *accuracy* as evaluation metric, reporting the percentage of correctly answered questions.

5.1 Dataset Analysis

We use some methods described below to analyze SVQA:

- **Chance:** The correct probability of random choosing an answer from answer set.
- **Q-type [2]:** This model chooses the most frequent train set answer in each question category as the test set answer.
- **GRU-Q:** This model is a language-only model that similar to 'LSTM-Q' [2]. The question is fed into a GRU and the final hidden state is passed to a MLP to predict an answer distribution.

Table 3: Experimental results of FrameQA task in TGIF-QA

Model		FrameQA
Random chance		0.06
VIS+LSTM [21]	aggregate	34.59
	average	34.97
VQA-MCB [4]	aggregate	25.70
	average	15.49
Yu *et al.* [35]		39.64
ST-VQA-Sp [10]		45.45
ST-VQA-Tp [10]		48.56
ST-VQA-Sp.Tp [10]		47.49
Co-memory (w/o DFE) [5]		51.00
Co-memory (full) [5]		51.50
Ours(S+T)		**53.53**

- **GRU+AVG:** This model averages the sequential video representation through temporal dimension and concatenates it with question representations generated by GRU.
- **2GRU:** This model uses two GRU to encode questions and videos separately and concatenates two outputs.

From Table. 2 we observe that results of Q-Type are almost identical with the chance. It suggests that answer distribution in each category is nearly balanced, i.e., models can not achieve high accuracy by remembering which answer is more 'familiar' to certain kind of questions. The language-only model (GRU-Q) performs poorly with accuracy no better than chance, suggesting that use questions alone while ignoring videos can not handle SVQA. Models (GRU+AVG and 2GRU) utilizing visual information perform better also confirms it. However, the performance of GRU-Q on integer comparison categories are surprisingly well. We believe that it is partly due to the number of one kind of geometry is related to the length of its description. The longer description means more attribute restrictions which limit the amount of matched geometries, thus the model can generate correct answers relying on comparison of the length of descriptions without using visual information. Besides, the performance of 2GRU is better than GRU+AVG to some extent. It is probably because simply average visual cues through the temporal dimension may destroy the inherent temporal structure of videos. This suggests that answering questions in SVQA needs to take consideration on temporal dependency between frames instead of only observing single frame. All results suggest that SVQA eliminates the limitations which we discussed in Section 1. We also display some examples of SVQA in Fig. 6.

5.2 Performance Comparisons

Five video question answering models as well as ours are used on SVQA for comparisons:

- **Extended VQA (E-VQA) [36]:** This model uses two GRU to encode videos and questions and fuses two representations by element-wise multiplication.
- **Extended Soft Attention (E-SA) [36]:** This model applies soft attention [29] on frames to generate answers.
- **ST-VQA-Tp [10]:** This model uses question representations to compute temporal attention masks on sequential video representations. We use it instead of others in [10] because it gets the highest accuracy on Frame QA task whose questions form is the same as ours.
- **r-STAN [40]**: This model is a hierarchical spatio-temporal attention network that learns the joint representation of questions and videos.
- **Unified-Att [31]:** This model integrates two attention models: sequential video attention model and temporal question attention model. The former accumulates the video attention for each word in sequential order of the question, while the later accumulates the question attentions for each frame along the temporal dimension.
- **Ours(S):** A variant of our model without ta-GRU.
- **Ours(T):** A variant of our model without spatial attention.
- **Ours(S+T):** Our whole model with both spatial attention and ta-GRU.

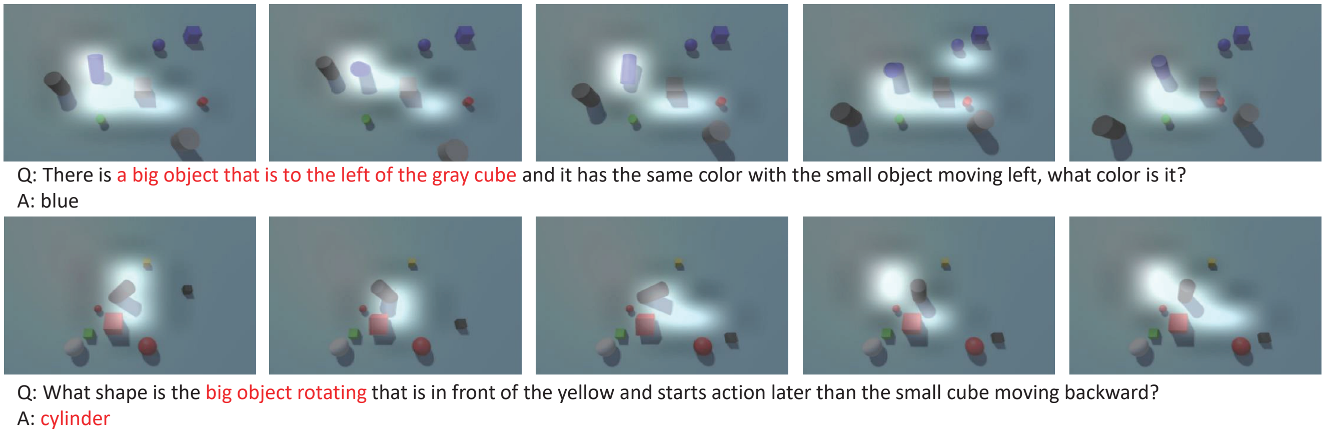
As shown in Table. 2, our whole model achieves the best (or the second best) performance, suggesting its effectiveness. We notice that E-VQA model which lacks of attention mechanism shows lower performance. Since our model with spatial attention mechanism outperforms it, the effectiveness of spatial attention can be confirmed. Besides, the improvement of E-SA compared with E-VQA also demonstrates the critical role of attention in this task. ST-VQA-Tp model shows inferior performance, mainly because it only employs temporal attention once, thus lose temporal visual cues. We attribute the lower performance of Unified-Att to traditional GRU which can poorly capture the long-term temporal dependency. While our model outperforms it suggesting the effectiveness of ta-GRU which can capture complete and long-term temporal visual cues. And we argue that the inferior performance of r-STAN probably due to overfitting.

5.3 Ablation Study

To further evaluate the effectiveness of spatial attention and ta-GRU in our attention module, we conduct ablation studies. For **Ours(S)**, we replace ta-GRU with traditional GRU. For **Ours(T)**,we simply average the dynamic video representations into one-dimension instead of employing spatial attention. Experimental results show that the whole model achieves the best performances among three models, validating the effectiveness of combinative way of spatial attention and ta-GRU. Besides, we notice that **Ours(S)** and **Ours(T)** almost outperform all other methods, suggesting that each one is useful. Furthermore, **Ours(T)** performs better than **Ours(S)** on most questions, suggesting that ta-GRU plays a key role. We also display some spatial attention visualization examples in Fig. 5. It is noticeable that the regions which have high correlations with the question (especially the objects asked in the question) are highlighted, which demonstrates that our spatial attention does work.

5.4 Generalization Ability

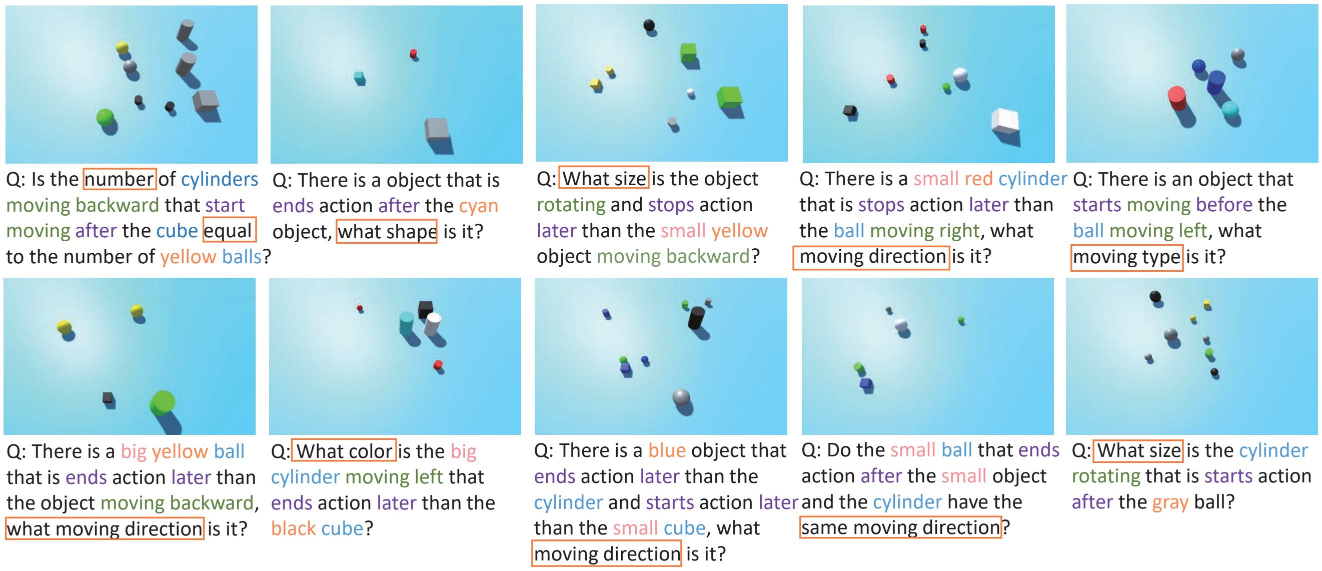
We implement our whole model on FrameQA task on TGIF-QA to verify its generalization ability. We choose FrameQA task because it has the same question form (open-ended) as that in SVQA, with which we can make a fair comparison without changing any modules in our model. Comparison results in Table. 3 show that our model outperforms other representative methods. From the



Q: There is a big object that is to the left of the gray cube and it has the same color with the small object moving left, what color is it?
A: blue

Q: What shape is the big object rotating that is in front of the yellow and starts action later than the small cube moving backward?
A: cylinder

Figure 5: Some spatial attention visualization examples of our model. We can notice that highlighted regions are corresponding to the red words in questions and answers, which can suggest that our spatial attention does work.



Q: Is the number of cylinders moving backward that start moving after the cube equal to the number of yellow balls?

Q: There is a object that is ends action after the cyan object, what shape is it?

Q: What size is the object rotating and stops action later than the small yellow object moving backward?

Q: There is a small red cylinder that is stops action later than the ball moving right, what moving direction is it?

Q: There is an object that starts moving before the ball moving left, what moving type is it?

Q: There is a big yellow ball that is ends action later than the object moving backward, what moving direction is it?

Q: What color is the big cylinder moving left that ends action later than the black cube?

Q: There is a blue object that ends action later than the cylinder and starts action later than the small cube, what moving direction is it?

Q: Do the small ball that ends action after the small object and the cylinder have the same moving direction?

Q: What size is the cylinder rotating that is starts action after the gray ball?

Figure 6: Several examples of SVQA containing center frames of each video and corresponding questions. Words with different colors represent relations and attributes described in questions which require to perform logical operations, while orange boxes which correspond to the outermost operation represent question categories.

comparison results we can see that, though developed towards VideoQA on the proposed synthetic SVQA dataset, the proposed model also shows good performance on other benchmarks like TGIF-QA. Thus the proposed model owns generalization ability for visual reasoning of VideoQA.

6 CONCLUSION

In this paper, we explore the multi-step reasoning in VideoQA by formulating it as a new task, which targets to answer compositional logical structured questions based on video content. To provide a benchmark for this task and address limitations of existing datasets, we design an automatic data construct system to generate our large-scale dataset SVQA. We also propose a general and powerful model with novel attention module to handle this task. Experimental results show the reasonability of SVQA and effectiveness of our model compared with other models. We hope SVQA and our model can provide a complement for existing literature of VideoQA and benefit further study of multi-step reasoning.

ACKNOWLEDGMENTS

This work is supported by the NSFC (Grant No. U1509206,61472276) and Tianjin Natural Science Foundation (No. 15JCYBJC15400).

REFERENCES

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Klein Dan. 2016. Neural Module Networks. In *CVPR*.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

[3] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

[4] 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 *EMNLP*.

[5] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-Appearance Co-memory Networks for Video Question Answering. In *CVPR*.

[6] Till Haug, Octavian-Eugen Ganea, and Paulina Grnarova. 2018. Neural Multi-Step Reasoning for Question Answering on Semi-Structured Tables. In *ECIR*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[8] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *ICCV*.

[9] Drew A. Hudsonand and Christopher D. Manning. 2018. Compositional Attention Networks for Machine Reasoning. In *ICLR*.

[10] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*.

[11] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. 2017. Inferring and Executing Programs for Visual Reasoning. In *ICCV*.

[12] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*. 1988–1997.

[13] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. In *ICLR*.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*. 1725–1732.

[15] 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*.

[16] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. 2016. A Context-aware Attention Network for Interactive Question Answering. In *ACM SIGKDD*.

[17] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A Dataset and Exploration of Models for Understanding Video Data Through Fill-In-The-Blank Question-Answering. In *CVPR*. 7359–7368.

[18] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering Questions by Watching Gameplay Videos. In *ICCV*.

[19] Ethan Perez, Harm de Vries, Florian Strub, Vincent Dumoulin, and Aaron C. Courville. 2017. Learning Visual Reasoning Without Strong Priors. In *ICML*.

[20] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*.

[21] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring Models and Data for Image Question Answering. In *NIPS*. 2953–2961.

[22] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher J. Pal, Hugo Larochelle, Aaron C. Courville, and Bernt Schiele. 2017. Movie Description. *IJCV* 123, 1 (2017), 94–120.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115 (2015), 211–252.

[24] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427* (2017).

[25] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*. 4631–4640.

[26] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*. 4489–4497.

[27] Huiyun Wang, Youjiang Xu, and Yahong Han. 2018. Spotting and Aggregating Salient Regions for Video Captioning. In *ACM MM*.

[28] Aming Wu and Yahong Han. 2018. Multi-modal Circulant Fusion for Video-to-Language and Backward. In *IJCAI*.

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Computer Science* (2015), 2048–2057.

[30] Youjiang Xu, Yahong Han, Richang Hong, and Qi Tian. 2018. Sequential Video VLAD: Training the Aggregation Locally and Temporally. *IEEE TIP* 27 (2018), 4933–4944.

[31] Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the Video and Question Attentions for Open-Ended Video Question Answering. *IEEE TIP* 26 (2017), 5656–5666.

[32] Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. 2018. A dataset and architecture for visual reasoning with a working memory. *arXiv preprint arXiv:1803.06092* (2018).

[33] Ziwei Yang, Yahong Han, and Zheng Wang. 2017. Catching the Temporal Regions-of-Interest for Video Captioning. In *ACM MM*.

[34] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video Question Answering via Attribute-Augmented Attention Network Learning. In *SIGIR*.

[35] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering. In *CVPR*.

[36] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging Video Descriptions to Learn Video Question Answering. In *AAAI*.

[37] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and Answering Binary Visual Questions. In *CVPR*.

[38] Shichao Zhao, Yanbin Liu, Yahong Han, and Richang Hong. 2017. Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. *TCSVT* (2017). https://doi.org/10.1109/TCSVT.2017.2682196

[39] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Dual-Level Attention Network Learning. In *ACM MM*. 1050–1058.

[40] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks. In *IJCAI*. 3518–3524.

[41] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. 2015. Uncovering Temporal Context for Video Question and Answering. *Computer Science* 124 (2015), 409–421.