Bridge to Answer: Structure-aware Graph Interaction Network for Video Question Answering

Jungin Park Jiyoung Lee Kwanghoon Sohn*
Yonsei University, Seoul, South Korea
{ newrun, easy00, khsohn }@yonsei.ac.kr

Abstract

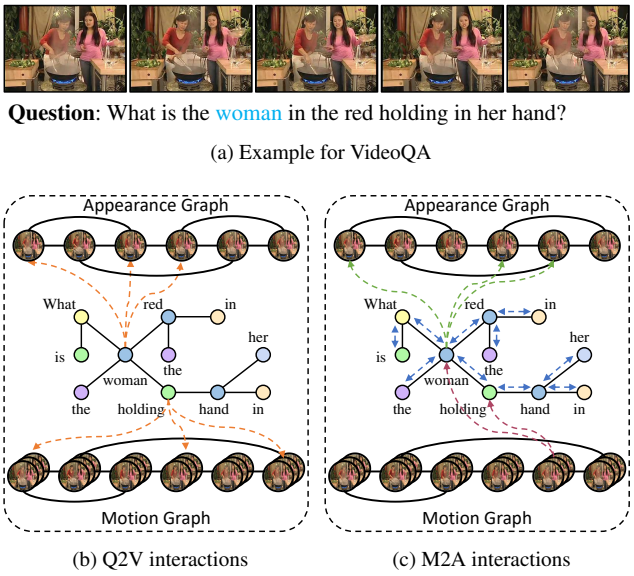
This paper presents a novel method, termed Bridge to Answer, to infer correct answers for questions about a given video by leveraging adequate graph interactions of heterogeneous crossmodal graphs. To realize this, we learn question conditioned visual graphs by exploiting the relation between video and question to enable each visual node using question-to-visual interactions to encompass both visual and linguistic cues. In addition, we propose bridged visual-to-visual interactions to incorporate two complementary visual information on appearance and motion by placing the question graph as an intermediate bridge. This bridged architecture allows reliable message passing through compositional semantics of the question to generate an appropriate answer. As a result, our method can learn the question conditioned visual representations attributed to appearance and motion that show powerful capability for video question answering. Extensive experiments prove that the proposed method provides effective and superior performance than state-of-the-art methods on several benchmarks.

1. Introduction

Video question answering (VideoQA) is a task to answer the question regarding a given video in a natural language form. Over the past few years, several methods have been focused on manipulating spatio-temporal visual representations conditioned by linguistic cues for VideoQA [20, 31, 32, 35]. However, because of its specificities such as dynamic spatiotemporal dependencies of the video and sophisticated compositional semantics of the question, the VideoQA still remains a challenging problem.

Recent works [11, 27, 6, 2, 5, 18] have adopted the encoder-decoder structure. Typically, LSTM-based encoders [11, 6, 2, 5] are used to encode the representations of video frames and a question into the visual and word sequence. The encoded representations are then incorporated to provide the answer with an attention mechanism. The several types of attention have shown promising results by learning the temporal relations between video frames [38, 2], the spatial relations between regions in every single frame [15, 37, 27], or spatiotemporal relations using appearance and motion representations [11, 6]. Although these methods have suggested how to use the visual relationship for VideoQA, they still rely on learning positional relationships, not on in-depth semantic meaning, which makes capturing sophisticated appearance-motion or visual-question relations difficult.

Meanwhile, methods to understand cross-modal relationships have been proposed for vision-language interac-



(a) Example for VideoQA

(b) Q2V interactions

(c) M2A interactions

Figure 1. (a) An example for VideoQA. (b) Question-to-Visual (Q2V) interactions that each question node are propagated to visual nodes. (c) Visual-to-Visual (V2V) interactions that each visual node are associated with the relative visual nodes using the question bridge. Only motion-to-appearance (M2A) interaction is shown.

*Corresponding author.

This research was supported by the Yonsei University Research Fund of 2021 (2021-22-0001).

tion tasks, such as image-text matching [21, 26] or video-text matching [1], exploiting global [22, 29, 36] or local [19, 25] representations for visual and textual information. Similar approaches have also been adopted in VideoQA. The global question representation has been used as a condition to learn a question-specific visual representation [40, 6, 18]. For example, the global question feature vector was used to update the memory network to learn attention that attributed to the question in [6]. Le *et al*. [18] proposed a hierarchical architecture that transforms a sequence of objects into a new array conditioned on the global question feature. The word-level features of the question have been treated as sequential data in the local approach [11, 38, 5, 24, 10]. These approaches leveraged each word representation to learn visual attention [11, 38, 2] or co-attention [27, 5, 24, 10] by fusing visual and word representations. However, the global approaches have learned coarse relations that frequently fail to capture video-word relations. The local approaches associate visual and word information based on co-occurrence statistics, not compositional semantics of the question. For instance, without semantic relations, the word "*woman*" of the question in Fig. 1-(a) can incorrectly be correlated with all women in the video. On the other hand, compositional semantics clearly indicate from the phrase "*in the red*" that the question point to the left woman.

To address these limitations, the consideration of grammatical dependencies between sentence words [3, 4] has been raised. For visual question answering (VQA), Teney *et al*. [33] proposed structured representations that the input image and question are encoded as graphs to leverage compositional semantics of the question. For image-text matching, Liu *et al*. [26] proposed a graph-structured network that construct graphs for the image and corresponding captions to find the fine-grained image-text correspondences using node-level and structure-level matching. Although the effectiveness of structured representations for image-text relations has been extensively demonstrated, it is still underexplored in VideoQA.

In this paper, we propose a novel method, called *Bridge to Answer*, that formulates structure-aware interaction for semantic relation modeling between crossmodal information, including appearance, motion, and question. Contrary to existing approaches [6, 10], we construct not only appearance and motion graphs for video but also the question graph that represents compositional semantic relations between words. We perform question-to-video (Q2V) interactions that propagate the question node to its relevant visual nodes to learn question conditioned visual representations with visual-question relations, as shown in Fig. 1-(b). Also, we apply visual-to-visual (V2V) interactions to visual graphs delivering each visual node to nodes in the relative visual graph to model appearance-motion relations. To utilize compositional semantic structure of the question, we use the question graph as an intermediate bridge, as shown in Fig. 1-(c). We demonstrate the capability of the proposed method through extensive ablation studies and comparison with state-of-the-art methods on three datasets, including TGIF-QA [11], MSVD-QA [38], and MSRVTT-QA [39].

2. Related Works

Video question answering (VideoQA) has attracted intense attention over the past few years due to its applicability to human-robot interaction and video retrieval, etc. The existing methods have mostly been proposed to learn visual representations by leveraging video-question interactions. We summarize recent works according to the types of utilized interactions. Typically, the temporal attention has been learned by exploiting relationships between the appearance and question [20, 43, 23]. Li *et al*. [23] proposed to learn co-attention between the appearance and question, and Li *et al*. [24] enhanced co-attention by using self-attention [34] mechanism. Some researchers have presented to capture fine-grained appearance-question interactions. Jin *et al*. [13] introduced object-aware temporal attention that learns object-question interactions. Huang etal [10] also utilize frame and object features to enhance co-attention between the appearance and question.

Since Jang *et al*. [11] proposed a two-stream architecture using appearance and motion features, researchers have focused on learning spatiotemporal attention that leverages motion, appearance, and question interactions. Developments of spatiotemporal attention have successfully been applied to various approaches including a multimodal fusion memory [5], co-memory attention [6], hierarchical attention [41, 40], multi-head attention [16], and multi-step progressive attention [14, 38, 31]. The hierarchical structure that capture appearance-question and motion-question relations from the frame-level to segment-level have also been proposed by Zhao *et al*. [42] and Le *et al*. [18].

Although they have suggested methods that effectively learn relations between appearance and question or even motion, they still rely on positional relationships [24]. Moreover, there have not been presented for capturing the relationship between appearance and motion with compositional semantics of the question. To address these limitations, we explicitly model appearance, motion, and the question as graphs. Our model learns question conditioned visual representations and mutually enhances appearance and motion representations by leveraging compositional semantics of the question.

Graph-structured vision-language interaction has recently been studied to learn semantic relations between visual and textual information. Teney *et al*. [33] firstly proposed to learn graph-structured representations of the input image and question for visual question answering (VQA).

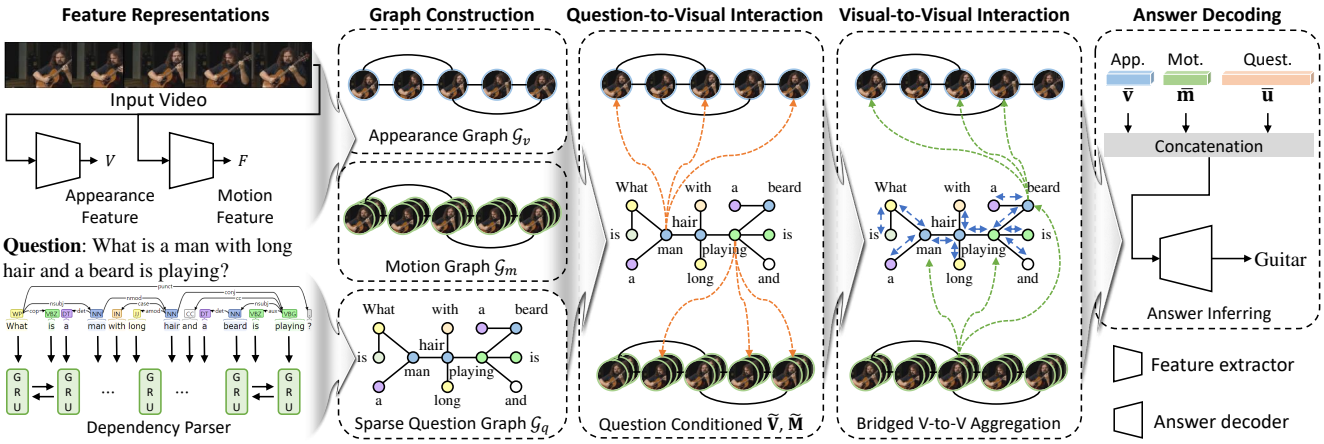


Figure 2. The overall architecture of the proposed method for VideoQA. The visual and question representations are extracted to construct appearance, motion and question graphs. The graph nodes in each graph are propagated to nodes in another graph differentially through question-to-visual interactions and visual-to-visual interactions to learn question conditioned visual representation attributed to appearance and motion.

More recently, Li *et al*. [21] proposed to learn relationships between regions in the input image using graph convolutional network (GCN) [17] and capture image-phrase correspondence for image-text matching. To enable fine-grained image-text matching, Liu *et al*. [26] constructed a visual graph for the input image and textual graph with the compositional semantics of the caption, respectively. They successfully achieved state-of-the-art performance by learning correspondences between two structured graphs. Similarly, Chen *et al*. [1] recently proposed a hierarchical graph reasoning method that learns fine-grained video-text correspondence. They composed hierarchical graphs for video and caption according to semantic roles, and performed global and local matching between two graphs.

While these works have suggested methods that effectively learn visual-text relations with structured representations, they cannot be directly applied to VideoQA. To our knowledge, our work is the first attempt to perform relation reasoning between appearance and motion information of the video with compositional semantics of the question.

3. Method

Given a video $\mathcal{V}$ and a question q, the VideoQA problem is generally formulated as follows:

$$\tilde{a} = \arg\max_{a \in \mathcal{A}} \mathcal{F}_\theta(a|q, \mathcal{V}), \tag{1}$$

where $\tilde{a}$ is the answer that can be inferred in answer space $\mathcal{A}$. θ is the set of model parameters of mapping function $\mathcal{F}$, which maps a pair of the video and question to the answer. We illustrate proposed method in Fig. 2. We first extract feature representations from the video and the question, and construct graphs for appearance, motion, and question, respectively (Sec. 3.1). The question nodes propagated to the visual graphs using question-to-visual interactions to learn question conditioned visual representations (Sec. 3.2). Thereafter, the nodes in each visual graph are aggregated into relevant nodes in the relative visual graph over a question bridge to enhance visual representations by learning appearance-motion relations (Sec. 3.3). Lastly, the final visual and question representations are concatenated and fed into the decoder to infer the answer (Sec. 3.4). Tab. 1 summarizes the notations used over our method. Following subsections, we depict the proposed method in details.

3.1. Feature Extraction and Graph Construction

Visual representations and visual graphs. Similar to the previous works for videoQA [18, 10], we divide the video $\mathcal{V}$ of L frames into N uniform length clips $C = \{C_1, ..., C_N\}$, such that the length of each clip is $T = L/N$. To represent two types of information of the video, we extract frame-wise appearance feature vectors $\mathbf{V}$ and clip-wise motion feature vectors $\mathbf{M}$. In our work, $\mathbf{V}$ and $\mathbf{M}$ are extracted from the pretrained feature extractor (*e.g*., ResNet [9] and ResNeXt-101 [8]). The extracted features are fed into the linear feature transformation layers to project $\mathbf{V}$ and $\mathbf{M}$ into d'-dimensional feature space to obtain $\hat{\mathbf{V}} = \{\hat{\mathbf{v}}_l | \hat{\mathbf{v}}_l \in \mathbb{R}^{\mathrm{d}'}\}_{l=1}^{L}$ and $\hat{\mathbf{M}} = \{\hat{\mathbf{m}}_n \in \mathbb{R}^{\mathrm{d}'}\}_{n=1}^{N}$, respectively.

With appearance and motion representations, we construct an appearance graph $\mathcal{G}_v$ and a motion graph $\mathcal{G}_m$ as undirected fully-connected graphs. The frames and clips are set to nodes, and each node is connected with all the other nodes in each graph with edges. The weight matrices $\mathbf{W}^v$ and $\mathbf{W}^m$, which represent node connections and their edge weights are computed by the affinity between node

Notation	Role
$\hat{\mathbf{V}}$, $\hat{\mathbf{M}}$	Input visual node representations
$\mathbf{U}$	Input question node representation
$\mathbf{W}^v$, $\mathbf{W}^m$, $\mathbf{W}^q$	Weight matrices of graphs
$\mathbf{S}^v$, $\mathbf{S}^m$	Q2V interaction matrix
$\tilde{\mathbf{V}}$, $\tilde{\mathbf{M}}$	Output of Q2V interaction
$\mathbf{U}_b^v$, $\mathbf{U}_b^v$	Bridged visual representations
$\hat{\mathbf{U}}_b^v$, $\hat{\mathbf{U}}_b^m$	Aggregated question representations
$\mathbf{S}_b^v$, $\mathbf{S}_b^m$	V2V interaction matrix
$\mathbf{V}^f$, $\mathbf{M}^f$	Output of V2V interaction

Table 1. Notations of Bridge to Answer

representations of $\hat{\mathbf{V}}$ and $\hat{\mathbf{M}}$ as

$$w_{ij}^v = \frac{\exp(\lambda \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_j)}{\sum_{j=0}^{L} \exp(\lambda \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_j)}, \quad w_{ij}^m = \frac{\exp(\lambda \hat{\mathbf{m}}_i^T \hat{\mathbf{m}}_j)}{\sum_{j=0}^{T} \exp(\lambda \hat{\mathbf{m}}_i^T \hat{\mathbf{m}}_j)}, \tag{2}$$

where w_{ij}^v and w_{ij}^m indicate the edge weights between i-th and j-th node in each graph. λ is a scaling factor.

Linguistic representations and question graph. For the linguistic representation, we first embed all words in the question into 300-dimensional vectors with pretrained word embeddings (*e.g.*, GloVe [30]). In the case of multiple choice questions, the words in answer candidates are also embedded. The embedded vectors are passed through a Bidirectional Gated Recurrent Unit (BiGRU) to establish the context dependency between words and are projected into the d'-dimensions feature space. The linguistic representations $\mathbf{U} = \{\mathbf{u}_i | \mathbf{u}_i \in \mathbb{R}^{\mathrm{d}'}\}_{i=1}^{K}$ are obtained by concatenating the hidden states of forward and backward GRU at each time step, and applying a linear feature transformation, where K is the number of words in a question.

To take compositional semantic structure of the question into the question graph $\mathcal{G}_q$, we identify the semantic dependency within the question (and answer candidates) using Stanford CoreNLP [28]. This parser is used to analyze the components in a sentence (*e.g.*, nouns, verbs, or quantifiers) and parse their semantic dependencies (*e.g.*, nominal subject or adjectival modifier). For example, given a question "What is the woman in the red holding in her hand?", "What", "red", and "holding" are semantically dependent with "woman". Based on these dependencies, we set each word as a node and connect two nodes if they are semantically dependent. To obtain the weight matrix of the question graph, we compute the affinity matrix $\mathbf{E}$ of the question representation $\mathbf{U}$ as

$$e_{ij} = \frac{\exp(\lambda \hat{\mathbf{u}}_i^T \hat{\mathbf{u}}_j)}{\sum_{j=0}^{K} \exp(\lambda \hat{\mathbf{u}}_i^T \hat{\mathbf{u}}_j)}, \tag{3}$$

where e_{ij} indicates the affinity between the i-th and j-th question node and λ is a scaling factor. Then the weight matrix $\mathbf{W}^q$ is represented by a Hadamard product between $\mathbf{E}$ and the adjacency matrix $\mathbf{A}$, followed by L_2 normalization, such that,

$$\mathbf{W}^q = ||\mathbf{E} \circ \mathbf{A}||_2, \tag{4}$$

where the adjacency matrix $\mathbf{A}$ represents the connectivity of the question graph.

3.2. Question-to-Visual Interactions

The goal of question-to-visual (Q2V) interactions is to learn question conditioned visual representations by associating the question nodes with visual nodes and propagating question representations along visual edges through graph convolution layers [17]. The Q2V interactions are performed as question-to-appearance (Q2A) and question-to-motion (Q2M) interaction, respectively. Since Q2V interactions are symmetric operations on each graph except for the number of nodes, we describe Q2A interaction in detail and then roughly depict that on Q2M interaction. Specifically, we first obtain an interaction matrix $\mathbf{S}^v$ by applying softmax function to the affinity matrix between $\hat{\mathbf{V}}$ and $\mathbf{U}$ along the question axis, such that $\mathbf{S}^v = \mathrm{softmax}_{\mathbf{U}}(\lambda \hat{\mathbf{V}} \mathbf{U}^T)$. The interaction value s_{ij}^v represents how much the j-th question node is associated with the i-th appearance node. All the question nodes are aggregated to the corresponding visual node with $\mathbf{S}^v$, followed by a fully connected (FC) layer, so that the aggregated appearance node representation is formulated as

$$\hat{\mathbf{v}}_i' = \sigma(\mathbf{W}_f^v(\hat{\mathbf{v}}_i + \sum_{j=1}^{K} s_{ij}^v \mathbf{u}_j) + b), \tag{5}$$

where $\hat{\mathbf{v}}_i'$ is the i-th node representation of the aggregated appearance graph, $\mathbf{W}_f^v$ and b are the learnable parameters of FC layer, and $\sigma(\cdot)$ is an activate function such as ReLU.

Subsequently, we apply consecutive graph convolution layers that take $\hat{\mathbf{V}}'$ and the weight matrix $\mathbf{W}^v$ as inputs to propagate the aggregated node to its neighborhoods along the appearance edges. Formally, the output of Q2A interaction is represented as

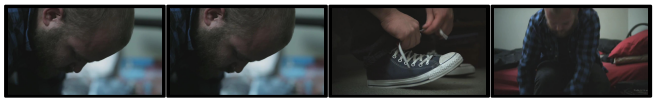
$$\tilde{\mathbf{V}} = \mathcal{F}(\mathbf{W}^v, \hat{\mathbf{V}}' | \mathbf{W}_g^v), \tag{6}$$

where $\mathbf{W}_g^v$ is the set of parameters of graph convolution layers.[1]

Symmetrically, question-to-motion (Q2M) interaction, which is performed to obtain the question conditioned motion representation $\tilde{\mathbf{M}}$, can be formulated as

$$\begin{aligned} \hat{\mathbf{m}}_i' &= \sigma(\mathbf{W}_f^m(\hat{\mathbf{m}}_i + \sum_{j=1}^{K} s_{ij}^m \mathbf{u}_j) + b), \\ \tilde{\mathbf{M}} &= \mathcal{F}(\mathbf{W}^m, \hat{\mathbf{M}}' | \mathbf{W}_g^m), \end{aligned} \tag{7}$$

[1] We denote consecutive graph convolution layers as a feed-forward process $\mathcal{F}$.



(a) **Question**: What is a man doing?

Baseline: Look
Ours: Tie
Groundtruth: Tie

(b) **Question**: What is a woman applies a concealer to the lower portion of her right cheek and blends it doing?

Baseline: Talk
Ours: Makeup
Groundtruth: Use

Figure 3. Example questions for challenging conditions. (a) Sudden transitions of the scene lead to confusion in the model capturing the visual relation. (b) Long and complex question composition makes it difficult for the model to learn a properly conditioned visual representation. Our model with the capacity to capture relations of heterogeneous cross-modal graphs copes with these challenging cases.

where $\mathbf{S}^m = \text{softmax}_{\mathbf{U}}(\lambda \hat{\mathbf{M}}\mathbf{U}^T)$ and $\mathbf{W}_g^m$ is the set of parameters of graph convolution layers.

3.3. Visual-to-Visual Interactions

One of the most important capabilities for VideoQA is to capture and incorporate the relations between appearance and motion information. To realize this, we present visual-to-visual (V2V) interaction that learns semantic relationships between appearance and motion. Different from previous works [6, 5] that appearance and motion information directly interact, we use the question graph as a bridge to leverage compositional semantics of the question. Since the structure of the question graph reflects semantic dependencies between words, the question conditioned visual node can effectively be delivered to the relative visual nodes along the question edges.

The V2V interaction can be summarized as three-fold: 1) one visual graph is bridged to the question graph, 2) the bridged node representation is propagated along the question edges through graph convolution layers and aggregated to the question graph, and 3) the aggregated question node is delivered to the relative visual graph. Concretely, motion-to-appearance (M2A) interaction begins by bridging motion and question graphs. A bridged motion representation, denoted as $\mathbf{U}_b^m$, can be obtained as a weighted combination of the question conditioned motion representation $\tilde{\mathbf{M}}$, where the weight is the interaction between $\mathbf{U}$ and $\tilde{\mathbf{M}}$, such that

$$\mathbf{U}_b^m = \text{softmax}_{\tilde{\mathbf{M}}}(\lambda \mathbf{U}\tilde{\mathbf{M}}^T)\tilde{\mathbf{M}}. \tag{8}$$

The bridged representation is propagated to its neighbors along the question edges through graph convolution layers and aggregated to the question representation as

$$\hat{\mathbf{U}}_b^m = \mathbf{U} + \mathcal{F}(\mathbf{W}^q, \mathbf{U}_b^m | \mathbf{W}_{gb}^m), \tag{9}$$

where $\mathbf{W}_{gb}^m$ is the set of trainable parameters of graph convolution layers. This form of the aggregated question graph enables the representation to have motion and question information simultaneously.

Finally, the aggregated question node is delivered to the appearance graph to obtain a question conditioned appearance representation attributed to motion. The output of M2A interaction can be formulated by following equation:

$$\begin{aligned} \mathbf{v}_i^f &= \sigma(\mathbf{W}_b^v(\tilde{\mathbf{v}}_i + \sum_{j=1}^{K}(s_b^v)_{ij}(\hat{\mathbf{u}}_b^m)_j) + b), \\ \mathbf{S}_b^v &= \text{softmax}_{\hat{\mathbf{U}}_b^m}(\lambda \tilde{\mathbf{V}}(\hat{\mathbf{U}}_b^m)^T), \end{aligned} \tag{10}$$

where $\mathbf{v}_i^f$ is the i-th node representation of the final appearance graph, $\mathbf{W}_b^v$ and b are the parameters of FC layer.

As a symmetric process, the node representation of the final motion graph $\mathbf{M}^f$ can be obtained with A2M interaction by following equations:

$$\begin{aligned} \mathbf{U}_b^v &= \text{softmax}_{\tilde{\mathbf{V}}}(\lambda \mathbf{U}\tilde{\mathbf{V}}^T)\tilde{\mathbf{V}}, \\ \hat{\mathbf{U}}_b^v &= \mathbf{U} + \mathcal{F}(\mathbf{W}^q, \mathbf{U}_b^v | \mathbf{W}_{gb}^v), \\ \mathbf{m}_i^f &= \sigma(\mathbf{W}_b^m(\tilde{\mathbf{m}}_i + \sum_{j=1}^{K}(s_b^m)_{ij}(\hat{\mathbf{u}}_b^v)_j) + b), \\ \mathbf{S}_b^m &= \text{softmax}_{\hat{\mathbf{U}}_b^v}(\lambda \tilde{\mathbf{M}}(\hat{\mathbf{U}}_b^v)^T), \end{aligned} \tag{11}$$

where $\mathbf{W}_{gb}^v$ and $\mathbf{W}_b^m$ are the weight parameters of graph convolution layers and FC layer, respectively.

We apply an average pooling to the final visual node representations along the temporal axis to vectorize the representations, and concatenate them to make the incorporated visual representation $\mathbf{o}$:

$$\begin{aligned} \bar{\mathbf{v}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{v}_l^f, \quad \bar{\mathbf{m}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{m}_n^f, \\ \mathbf{o} = [\bar{\mathbf{v}}; \bar{\mathbf{m}}], \end{aligned} \tag{12}$$

where $[\cdot;\cdot]$ denotes concatenation operation.

3.4. Answer Decoder and Loss Functions

Following previous works [5, 18], we use different answer decoders depending on the type of question. Specifically, we treat open-ended questions as a multi-label classification problem. The decoder takes the final visual representation $\mathbf{o}$ and the averaged question representation $\bar{\mathbf{u}}$ to

Model	Action	Trans.	Frame	Count
ST-TP [11]	62.9	69.4	49.5	4.32
Co-mem [6]	68.2	74.3	51.5	4.10
PSAC [24]	70.4	76.9	55.7	4.27
HME [5]	73.9	77.8	53.8	4.02
L-GCN [10]	74.3	81.1	56.3	3.95
QueST [12]	75.9	81.0	**59.7**	4.19
HCRN [18]	75.0	81.4	55.9	3.82
Ours	**75.9**	**82.6**	57.5	**3.71**

Table 2. Performance comparison for several tasks on TGIF-QA [11] dataset with state-of-the-art methods. The lower the better for count.

compute label probabilities $p \in \mathbb{R}^{|\mathcal{A}|}$:

$$\begin{aligned} y &= \sigma(\mathbf{W}_2[\mathbf{o}, \mathbf{W}_1\bar{\mathbf{u}} + b] + b), \\ y' &= \sigma(\mathbf{W}_y y + b), \\ p &= \text{softmax}(\mathbf{W}_{y'} y' + b), \end{aligned} \tag{13}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_y$, and $\mathbf{W}_{y'}$ are of learnable parameters of the decoder. We employ the cross-entropy loss for open-ended questions.

We treat repetition count task as a linear regression problem that the decoder takes y' in Eq. (13) as an input and applies a rounding function for integer count results. The Mean Squared Error (MSE) is employed as the loss function.

For multiple choice question types (*i.e.*, repeating action and state transition), $|\mathcal{A}|$ answer candidates are used to make a set of visual representations corresponding to each candidate in the same way with the question. As a result, we have a set of final visual representations, $\mathbf{o}$ conditioned by the question, and $\{\mathbf{o}_i^a\}_{i=1}^{|\mathcal{A}|}$ conditioned by answer candidates. The classifier for multiple choice question takes $\mathbf{o}$, $\mathbf{o}_i^a$, $\bar{\mathbf{u}}$, and answer candidates $\bar{\mathbf{a}}_i$ to output probabilities for candidates as follows:

$$\begin{aligned} y_i &= [\mathbf{o}, \mathbf{o}_i^a, \mathbf{W}_w\bar{\mathbf{u}} + b, \mathbf{W}_a\bar{\mathbf{a}}_i + b], \\ y_i' &= \sigma(\mathbf{W}_y y_i + b), \\ s_i &= \mathbf{W}_{y'} y_i' + b, \end{aligned} \tag{14}$$

where $\mathbf{W}_1$, $\mathbf{W}_a$, $\mathbf{W}_y$, and $\mathbf{W}_{y'}$ are of learnable parameters the decoder. Then, the candidate with the largest s value is selected as the answer such that,

$$\tilde{a} = \arg\max_i s_i. \tag{15}$$

We employ the hinge loss [7] between incorrect answer score s^n and correct answer score s^p, $\max(0, 1 + s^n - s^p)$, as the loss function.

4. Experiment

4.1. Datasets

TGIF-QA [11] dataset contains $72K$ animated GIF files and $165K$ question answer pairs. The dataset provides four kinds of tasks that address the unique properties of videos. *Repetition Count* is to retrieve number of occurrences of an action. *Repeating action* is a task to identify an action repeated for a given number of times among multiple choices. *State Transition* is a multiple choice task to identify an action regarding the temporal order of action state. *Frame QA* is to find a particular frame in a video that can answer the questions.

Model	MSVD-QA	MSRVTT
AMU [38]	32.0	32.5
HRA [2]	34.4	35.0
Co-mem [6]	31.7	31.9
HME [5]	33.7	33.0
L-GCN [10]	34.3	-
QueST [12]	36.1	34.6
HCRN [18]	36.1	35.6
Ours	**37.2**	**36.9**

Table 3. Performance comparison for open-ended questions on MSVD-QA [38] and MSRVTT-QA [39] datasets with state-of-the-art methods.

MSVD-QA [38] dataset contains $1,970$ short clips and $50,505$ question answer pairs. The questions are composed of five types, including what, who, how, when, and where.

MSRVTT-QA [39] dataset contains $10K$ videos and $243K$ question answer pairs. While types of questions are the same with MSVD-QA dataset, the contents of the videos in MSRVTT-QA are more complex and the lengths of the videos are much longer from 10 to 30 seconds.

For the evaluation metrics, we use Mean Squared Error (MSE) for repetition count on TGIF-QA dataset and use accuracy for all the other experiments.

4.2. Implementation Details

We divide the video into 8 clips containing 16 frames in each clip by default. Following the previous work [18], the long videos in MSRVTT-QA are additionally divided into 24 clips. We train our model for 25 epochs with a batch size of 16 for TGIF-QA and MSVD-QA datasets, and of 4 for MSRVTT-QA dataset. The learning rate is set to 10^{-4} and decayed by half for every 5 epochs. The reported results are at the epoch showing the best validation accuracy.

4.3. Experimental Results

We compare our model with several state-of-the-art methods on TGIF-QA, MSVD-QA, and MSRVTT-QA datasets.[2] For TGIF-QA dataset, we compare our model with [11, 6, 24, 5, 10, 12, 18] in Tab. 2. We display evaluation results over four tasks, including repeating action, state transition, frameQA, and repetition count. The results

[2]Reported values of the other methods are taken from the original papers and [18]

Model	Act.	Trans.	F.QA	Count
Input conditioning				
w/o appearance	72.8	77.2	-	4.01
w/o motion	68.2	75.5	57.3	4.21
Interaction				
w/o Q2V, V2V	69.4	75.3	51.4	3.97
w/o Q2A	73.9	80.1	52.7	3.87
w/o Q2M	73.1	78.2	56.8	3.90
w/o Q2V	72.3	76.3	52.6	3.95
w/o A2M	74.7	81.4	56.1	3.84
w/o M2A	74.8	80.9	56.9	3.86
w/o V2V	74.1	78.5	56.5	3.89
Bridge conditioning				
w/o bridge	75.1	81.7	56.9	3.83
Parameter λ				
$\lambda = 1$	75.1	81.5	56.2	3.80
$\lambda = 5$	75.3	81.8	57.2	3.77
$\lambda = 10$	75.9	82.6	57.5	3.71
$\lambda = 20$	75.4	82.2	57.1	3.73
Full model	75.9	82.6	57.5	3.71

Table 4. Ablation studies for input conditioning, interaction, and the value of λ on TGIF-QA dataset [11]. Act.: Action; Trans.: Transition; F.QA: Frame QA. When not explicitly specified, we use $\lambda = 10$. The lower the better for count.

show that our model achieves state-of-the-art performance and outperforms the existing methods on all tasks except for FrameQA task.

For MSVD-QA and MSRVTT-QA datasets, our model is compared with [38, 2, 6, 5, 10, 12, 18] in Tab. 3. Since these datasets provide open-ended questions, they are referred to as highly challenging benchmarks compared to the TGIF-QA dataset. Our model achieves 37.2% and 36.9% accuracy, outperforming the existing approaches by 1.1% and 1.3% for accuracy, respectively.

We provide qualitative results for two challenging examples in Fig. 3. The first example shows that a sudden transition of the scene causes the problem of capturing semantic relations. Our model handles this problem by learning in-depth semantic relations, not positional relations. The second example reflects the case in which a long and complex question has given[3]. Our model successfully analyzes this long and complex question by explicitly modeling the compositional semantics of the question.

4.4. Ablation Studies

To validate the effectiveness of components within our model, we conduct extensive ablation studies on TGIF-QA [11], as shown in Tab. 4. Ablation studies widely cover the results according to input conditioning, interactions, question bridge, and the value of the parameter λ.

[3]Groundtruth is probably incorrectly annotated.

Parameter λ	MSVD-QA	MSRVTT
$\lambda = 1$	35.3	33.8
$\lambda = 5$	36.9	35.0
$\lambda = 10$	**37.2**	36.6
$\lambda = 20$	36.5	**36.9**

Table 5. Performance comparison with different λ values on MSVD-QA [38] and MSRVTT-QA [39] datasets.

The overall result verify that all components affect performance, and even any direction of interaction contribute to the performance improvement. The detailed analyzes are described below.

Effect of input conditioning. We study the effect of the input condition with following settings:
► *w/o appearance*: Remove appearance feature from full model. Q2A and V2V interaction also been removed. Since the frames are not used in this setting, we do not measure the performance for FrameQA. ► *w/o motion*: Remove motion feature from full model. Q2M and V2V interaction also been removed. We find that while the absence of either appearance or motion feature is critical to the performance, motion contributes more to the performance in action-related tasks. The results of FrameQA show that motion information does not play an important role in tasks where appearance information of the frame is important.

Effect of interactions. To investigate the effectiveness of each interaction (*e.g.*, Q2A or A2M), we evaluate our model with all possible combination of interactions:
► *w/o Q2V*: Not using question conditioned visual representations and performing V2V interaction only with the question bridge. ► *w/o Q2A* or *w/o Q2M*: Using only one question conditioned visual representation (*i.e.*, using $\hat{\mathbf{V}}$ and $\tilde{\mathbf{M}}$, or $\tilde{\mathbf{V}}$ and $\hat{\mathbf{M}}$). ► *w/o V2V*: No interaction between the two visual representations. ► *w/o A2M* or *w/o M2A*: Using one-way V2V interaction (*i.e.*, motion to appearance or appearance to motion only).

Overall we find that the absence of any direction of interaction significantly degrades the performance on all tasks. Specifically, Q2V works as a primary component for VideoQA. This is not surprising given that learning question conditioned visual representations is one of the most important ingredient for VideoQA. The notable performance degradation due to the ablation of Q2M is shown in all the tasks except for FrameQA.

We also find that A2M and M2A are complemented each other from the results that promising performance can be achieved with only one-way interaction. To analyze V2V interaction, we display the visualization for the connectivity of motion-question and question-appearance as shown in Fig. 4. We depict M2A interaction only and represent the clips as frames sampled at each clip for visibility. The clips and words are placed according to temporal and word

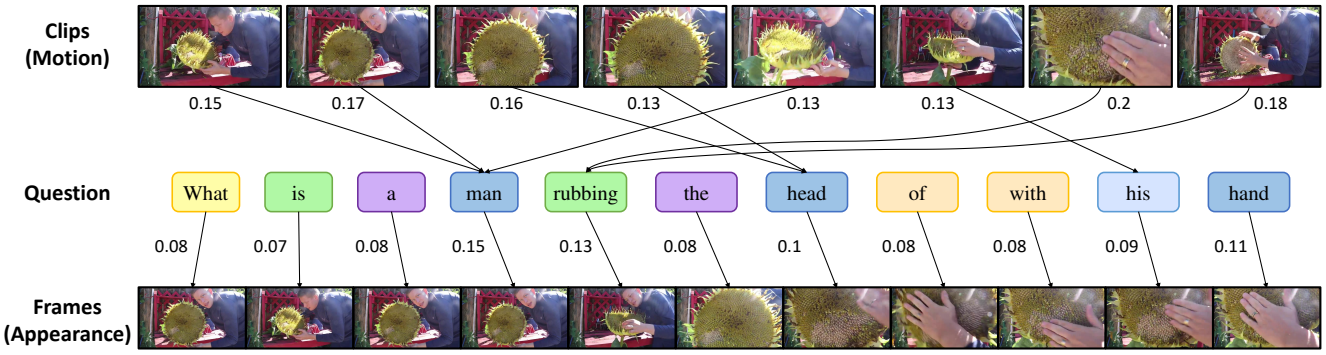


Figure 4. Visualization of M2A interaction. Although any two graphs are fully connected by interaction value, we only indicate the connection with the largest value in each interaction matrix for visibility. Note that the clips are represented by frames sampled from each clip.

order, respectively, and corresponding frames of each word are placed regardless of temporal order. The connections indicate that two nodes are associated with the maximum interaction value. For example, the first clip is associated with the word "man" by the interaction value of 0.15, and the word "man" is related to the fourth frame by the interaction value of 0.15. Note that, when all the nodes are connected with uniformly distributed weights, the motion-question interaction value and the question-appearance interaction value is 0.09 and 0.008, respectively. The results show that the relavant nodes are connected with relatively high interaction values, and the question node is also connected with the appearance node through semantic relation.

Effect of the question bridge. The question bridge is a key component to leverage the compositional semantics of the question. To verify the effectiveness, we conduct ablation study for the question bridge. The V2V interactions without the question bridge are performed by directly aggregating the relative node representations based on the affinity value between appearance and motion nodes. For instance, the output of the M2A interaction without the bridge is obtained by

$$\mathbf{v}_i^f = \sigma(\mathbf{W}_{wob}^v(\tilde{\mathbf{v}}_i + \sum_{j=1}^{T}(s_{wob}^v)_{ij}\tilde{\mathbf{m}}_j) + b), \qquad (16)$$

where $\mathbf{S}_{wob}^v = \text{softmax}_{\tilde{\mathbf{M}}}(\lambda\tilde{\mathbf{V}}\tilde{\mathbf{M}}^T)$.

The results at each task demonstrate the advantage of the bridged architecture with the performance improvement of 0.5%, 0.9%, 0.6% for accuracy, and 0.12 for MSE value, respectively.

Effect of λ. The scaling parameter λ adjusts the relative weight of different nodes in Q2V and V2V interactions and the edge weight of graphs. The large value of λ distills nodes highly correlated to the specific node and filters out irrelevant nodes. Contrary to this, the small value of λ is difficult to distinguish relevant nodes. Therefore, it is important to properly set the value of λ. To investigate the performance with various λ values, we measure VideoQA performance by setting the λ as 1, 5, 10, 20. Not surprisingly, Lager λ shows better performance compared to $\lambda = 1$.

We additionally evaluate our model according to λ values on MSVD-QA [38] and MSRVTT-QA [39] datasets. As shown in Tab. 5, our model yield highest performance when $\lambda = 10$ on MSVD-QA. The results on MSRVTT-QA show that $\lambda = 20$ brings out the best performance. The different optimal value of λ on two datasets might be caused by different lengths of videos.

5. Conclusion

We proposed a novel method for VideoQA, called Bridge to Answer, that constructs heterogeneous multi-modal graphs and learns relations between visual and question graphs to learn question conditioned visual representations attributed to appearance and motion. In the process, in-depth semantic relations between visual and question graphs are encapsulated to visual representations using question-visual interactions. The relations between appearance and motion graphs are modulated by compositional semantics of the question as a bridge to effectively enhance each relative visual representation. This bridged structure allows a model robust to the scene composition and sophisticated structure of the question. Our model was evaluated on several VideoQA benchmarks, including TGIF-QA, MSVD-QA, and MSRVTT-QA, achieving state-of-the-art performance.

Acknowledgement

This work was supported by Institute of Information communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00056, To create AI systems that act appropriately and effectively in novel situations that occur in open worlds)

References

[1] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10638–10647, 2020. 2, 3

[2] Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. *IEEE Int. Conf. Image Process.*, pages 599–603, 2018. 1, 2, 6, 7

[3] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. *Int. Conf. Lang. Resour. Eval.*, 2006. 2

[4] Marie-Catherine de Marneffe and Christopher D. Manning. The stanford typed dependencies representation. *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 2008. 2

[5] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1999–2007, 2019. 1, 2, 5, 6, 7

[6] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 5, 6, 7

[7] Claudio Gentile and Manfred K. Warmuth. Linear hinge loss and average margin. *Adv. Neural Inform. Process. Syst.*, pages 225–231, 1999. 6

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6546–6555, 2018. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 3

[10] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. *AAAI*, 2020. 2, 3, 6, 7

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2758–2766, 2017. 1, 2, 6, 7

[12] Jianwen Jiang, Ziqiang Chen, HaoJie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. *AAAI*, pages 11101–11108, 2020. 6, 7

[13] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Multi-interaction network with object relation for video question answering. *ACM Int. Conf. Multimedia*, pages 1193–1201, 2019. 2

[14] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8337–8346, 2019. 2

[15] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. *Adv. Neural Inform. Process. Syst.*, 2016. 1

[16] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and ByoungTak Zhang. Multimodal dual attention memory for video story question answering. *Eur. Conf. Comput. Vis.*, pages 673–688, 2018. 2

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Int. Conf. Learn. Represent.*, 2017. 3, 4

[18] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9969–9978, 2020. 1, 2, 3, 5, 6, 7

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *Eur. Conf. Comput. Vis.*, pages 201–216, 2018. 2

[20] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering. *Conference on Empirical Methods in Natural Language Processing*, 2018. 1, 2

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. *Int. Conf. Comput. Vis.*, pages 4654–4662, 2019. 2, 3

[22] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. *Int. Conf. Comput. Vis.*, pages 1908–1917, 2017. 2

[23] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. Learnable aggregating net with diversity learning for video question answering. *ACM Int. Conf. Multimedia*, pages 1166–1174, 2019. 2

[24] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. *AAAI*, 2019. 2, 6

[25] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. *ACM Int. Conf. Multimedia*, pages 3–11, 2019. 2

[26] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10921–10930, 2020. 2, 3

[27] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory-augmented networks. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2

[28] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. *ACL Sys. Demonst.*, pages 55–60, 2014. 4

[29] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 299–307, 2017. 2

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 4

[31] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. *ACM Int. Conf. Multimedia*, pages 239–247, 2018. 1, 2

[32] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4631–4640, 2016. 1

[33] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 2017. 2

[35] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. *AAAI*, 2018. 1

[36] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):394–407, 2019. 2

[37] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *Int. Conf. Mach. Learn.*, 2016. 1

[38] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. *ACM Int. Conf. Multimedia*, pages 1645–1653, 2017. 1, 2, 6, 7, 8

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5288–5296, 2016. 2, 6, 7, 8

[40] Zhou Zhao, Xinghua Jiang, Deng Cai, Xiaofei He Jun Xiao, and Shiliang Pu. Multi-turn video question answering via multi-stream hierarchical attention context network. *Int. Joint. Conf. Art. Intell.*, pages 3690–3696, 2018. 2

[41] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatiotemporal attention networks. *IJCAI*, pages 3518–3524, 2017. 2

[42] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhenxin Xiao, Xiaohui Yan, Jun Yu, Deng Cai, and Fei Wu. Long-form video question answering via dynamic hierarchical reinforced networks. *IEEE Trans. Image Process.*, 28(12):5939–5952, 2019. 2

[43] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.*, 124(3):409–421, 2017. 2