Keyword-Aware Relative Spatio-Temporal Graph Networks for Video Question Answering

Yi Cheng, Hehe Fan, Dongyun Lin, Ying Sun, *Member, IEEE*, Mohan Kankanhalli, *Fellow, IEEE*, and Joo-Hwee Lim, *Senior Member, IEEE*

arXiv:2307.13250v1 [cs.CV] 25 Jul 2023

Abstract—The main challenge in video question answering (VideoQA) is to capture and understand the complex spatial and temporal relations between objects based on given questions. Existing graph-based methods for VideoQA usually ignore keywords in questions and employ a simple graph to aggregate features without considering relative relations between objects, which may lead to inferior performance. In this paper, we propose a Keyword-aware Relative Spatio-Temporal (KRST) graph network for VideoQA. First, to make question features aware of keywords, we employ an attention mechanism to assign high weights to keywords during question encoding. The keyword-aware question features are then used to guide video graph construction. Second, because relations are relative, we integrate the relative relation modeling to better capture the spatio-temporal dynamics among object nodes. Moreover, we disentangle the spatio-temporal reasoning into an object-level spatial graph and a frame-level temporal graph, which reduces the impact of spatial and temporal relation reasoning on each other. Extensive experiments on the TGIF-QA, MSVD-QA and MSRVTT-QA datasets demonstrate the superiority of our KRST over multiple state-of-the-art methods.

Index Terms—Video question answering, relative relation reasoning, spatial-temporal graph.

I. INTRODUCTION

VIDEO question answering (VideoQA) is a challenging task in Multimedia Intelligence [1]–[3], and it aims to answer the question based on a thorough understanding of the given video. The task requires the powerful cognitive capability of spatio-temporal visual representations guided by the compositional semantics of the given question. In recent years, VideoQA has drawn increasing attention due to its wide applications in various domains, *e.g.*, human-robot interaction and autonomous driving.

Despite its recent achievements, VideoQA still remains challenging as it requires effective reasoning about complex spatio-temporal relations based on the vision and language modalities [4], [5].

Yi Cheng is with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), and also with the School of Computing, National University of Singapore, Singapore (e-mail: cheng_yi@i2r.a-star.edu.sg).

Dongyun Lin, Ying Sun are with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore (e-mail: {lin_dongyun, liu_yanzhu, suny}@i2r.a-star.edu.sg).

Joo-Hwee Lim is with the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore, and also with the SCSE, Nanyang Technological University, Singapore (e-mail: joohwee@i2r.a-star.edu.sg).

H. Fan and M. Kankanhalli are with the School of Computing, National University of Singapore, Singapore (e-mail: hehe.fan.cs@gmail.com, mohan@comp.nus.edu.sg).

Fig. 1. Illustration of the proposed Keyword-aware Relative Spatio-Temporal (KRST) graph network. **(a)** We employ an attention mechanism to assign high weights to keywords during question embedding, which is then integrated into the object-level graph. **(b)** Relations are relative. Choosing different subjects (center objects) may lead to different understandings. For example, "a man is *talking* to a dog" can also be understood as "a dog is *listening* to a man". Therefore, we integrate the relative relation modeling into graph reasoning. **(c)** Different from most existing methods that employ a unified graph to model spatial-temporal relations together, we use an object-level graph for capturing spatial relation and a frame-level graph for temporal relation reasoning.

To model the dynamics of relationships in videos, one solution is to capture the video structure at the frame level. For example, HCRN [6] proposes a relation network to select relevant frames for each element in the context of the question query. Graph neural network (GNN) methods [7], [8] are also used to model the temporal relations across frames by building graphs over video segments. However, because the object-level information is largely ignored, these frame-level methods are only able to model a limited number of objects and may fail to generalize to scenarios where multiple objects interact with each other.

The second solution is to capture relations at the object level, which is more flexible and able to model the complex spatial and temporal interactions among more objects. To this end, GNNs are usually employed to model the relation structure among local objects [9]–[11]. However, those methods treat all objects equally and potentially do not distinguish or recognize keywords in questions. For the same video, different questions tend to focus on the interactions of different objects. Therefore, keyword-aware question embeddings should be exploited to guide the object-level graph construction.

Moreover, those GNNs usually perform simple aggregation on neighbouring nodes (*e.g.*, via an affinity matrix), in which the relative relations among objects are largely ignored. Without relative relation reasoning, networks may fail to properly capture the spatio-temporal structure.

In this paper, we propose a Keyword-aware Relative Spatio-Temporal (KRST) graph network for VideoQA. The graph is built over objects with guidance from questions. First, as shown in Fig. 1(a), to make question embeddings aware of keywords, we employ an attention mechanism to assign high weights to keywords during question encoding. The keyword-aware question features are then integrated into object features. Our graph is based on those keyword-aware object representations. Second, we integrate relative relation modeling into our graph. Our motivation is that relations are relative, whether for semantics or positions. For the same scene, choosing different subjects may lead to different relation understanding. As shown in Fig. 1(b), when reasoning about the relation between the man and the dog, it can be understood as "a man is *talking* to a dog" if we focus on the man or "a dog is *listening* to a man" if we select the dog as the target object: $man - dog = talk$ and $dog - man = listen$. Also, from the perspective of position, the scene can be understood as "the dog is left to the man" or "the man is right to the dog": $man - dog = right$ and $dog - man = left$. Therefore, it is important to equip graph reasoning with the relative relation modeling ability. Third, instead of modeling spatio-temporal relations jointly, we disentangle the modeling into an object-level graph for spatial reasoning and a frame-level graph for temporal structure capture, as shown in Fig. 1(c). In this way, the object graph only focuses on extracting the object spatial relations of interest regarding the question while the frame-level graph only needs to capture the dynamics of the attended object relations, which reduces the burden of networks for reasoning. The contributions of this paper are threefold:

- We propose a Keyword-aware Relative Spatio-Temporal (KRST) graph network for VideoQA, which performs relation reasoning over objects with keyword-aware question features.
- We introduce relative relation modeling for VideoQA, which enables graph networks to better understand spatial-temporal relationships among objects.
- We conduct extensive experiments on TGIF-QA, MSVD-QA and MSRVTT-QA datasets, and the results demonstrate the superiority of our KRST to the state-of-the-arts.

II. RELATED WORK

Recently, there has been a rapid progress in vision-language tasks [12]–[19], such as image captioning, visual question answering, visual dialog, text-video retrieval, *etc*. According to the types of visual information, question answering (QA) can be classified into image QA and video QA. Compared with image QA, video QA is more challenging because both spatial and temporal relations are required to be modelled for correct answer prediction [20].

Inspired by the recent advances in large vision and language models, some existing VideoQA methods [21]–[25] attempt to extract richer information from videos and questions by applying large-scale pre-trained models. However, this work focuses on exploiting the interactions between semantic clues from the visual contents and linguistic questions. Existing methods on VideoQA can be grouped into two categories: attention-based and graph-based methods.

a) Attention-based methods: Typical attention-based methods [4], [26]–[28] learn temporal attention by modeling the correlation between appearance and question. For example, co-attention [27] is proposed to model appearance and question interaction, while object-level attention [28] is proposed to learn object-question interaction. Some other works learn spatio-temporal attention by leveraging both appearance and motion features, including co-memory attention [29], hierarchical attention [30], multi-head attention [31] and multi-step progressive attention [32]. HCRN [6] proposes a hierarchical structure to model the temporal relation of a single object with a stack of conditional relation blocks. However, these methods may fail to handle cases where multiple objects interact.

b) Graph-based methods: Graph convolutional network (GCN) has been widely applied to various vision-language tasks [33]–[35]. For VideoQA, L-GCN [9] proposes a location-aware graph to model the relation between different objects in each video frame. GMN [10] designs a holistic spatio-temporal graph to model the relations between different objects by taking object trajectories as input. MASN [11] extends the spatio-temporal graph over objects by exploiting the complementary nature between appearance and motion information. HOSTR [36] leverages the question features to compute the correlation matrix between objects for relation reasoning. HGA [7] proposes a uniform heterogeneous graph with video shots and words as nodes to incorporate both inter- and intra-modality interactions. B2A [8] constructs multi-modal graphs for appearance, motion and questions, where the question graph is used to model the relation between video and question.

Different from these approaches, our method builds spatio-temporal graphs by attending to keyword-relevant objects. Moreover, we integrate relative relation modeling into spatio-temporal graph reasoning. Last, to explicitly model the spatial and temporal relations in videos, we decompose the spatio-temporal graphs into a spatial graph over objects within each frame and a temporal graph across frames.

III. METHODOLOGY

Given a video $\mathcal{I}$ and a linguistic question q, the VideoQA problem can be formulated as follows,

$$\bar{a} = \arg\max_{a \in \mathcal{A}} \mathcal{F}_\theta \left(a \mid q, \mathcal{I} \right), \qquad (1)$$

where $\bar{a}$ is the predicted answer from an answer set $\mathcal{A}$. $\mathcal{F}_\theta(.)$ is the mapping function and θ is the set of parameters. Fig. 2 illustrates the overall architecture of the proposed KRST method. First, the question embeddings are extracted via GloVe [37] and Bidirectional LSTM (BiLSTM). We design a keyword-aware module to guide question embedding, which promotes a deep understanding of object interactions by reducing the noise from question-irrelevant objects. Second, the

Fig. 2. The architecture of the proposed Keyword-Aware Relative Spatio-Temporal graph network for VideoQA. The video and question features are firstly extracted as described in Section III-A. Then, then we construct the keyword-aware graph by designing a keyword attention module to identify and augment relevant object features for the subsequent relation reasoning. Next, we perform relative spatio-temporal graph reasoning over objects using disentangled spatial and temporal graphs. Finally, the question features and visual features are fused to predict the final answer.

object-level features are generated using pre-trained models based on RoIAlign. Those object-level features are then used to build the graph, guided by the keyword-aware question embedding. We integrate the relative relation modeling into graph reasoning. Finally, both object- and frame-level features are fused with question features to predict the final answer.

A. Video and Question Representation

1) Video-level Appearance and Motion Representation: We follow the common practice in previous works to extract both appearance and motion features from videos. Specifically, T frames are uniformly sampled from each video. Then, pre-trained ResNet-152 [38] and I3D [39] models are employed to extract the global context appearance feature and motion feature, respectively. For motion feature extraction, because I3D takes multiple frames as input, a set of 8 neighboring frames around each sampled frame is concatenated and fed into I3D model. Based on these two types of features, we build a two-stream architecture [40].

Note that, for simplicity, in the following descriptions, we do not make notes or subscripts to distinguish appearance and motion features but treat them as visual features. In this case,

we let $\boldsymbol{I} \in \mathbb{R}^{T \times C}$ denote the video-level appearance or motion features, where C denotes the feature dimension.

2) Object-level Appearance and Motion Representation: To obtain object-level features, we first generate K object bounding boxes from each frame by employing a pre-trained object detector Faster R-CNN [41]. In this case, there are TK objects in each video. Then, the object semantic features $\boldsymbol{O}_s \in \mathbb{R}^{TK \times C_s}$ for appearance or motion are obtained by applying RoIAlign onto the video-level ResNet-152 or I3D feature maps, respectively, where C_s is the dimension of the semantic feature maps. To effectively model object relations, it is important to integrate the object location information into object features. In this paper, we use the x and y coordinates of the upper-left corner, such coordinates of the lower-right corner and the height and width of the bounding box, *i.e.*, $(x_1, y_1, x_2, y_2, w, h)$, as the object location information: $\boldsymbol{O}_p \in \mathbb{R}^{TK \times 6}$. Similar to L-GCN [9], we first concatenate the object semantic features and the bounding box location. Then, the object features are obtained by projecting the concatenated features into C-dimension using a linear transformation,

$$\boldsymbol{O} = \boldsymbol{W}_o \cdot [\boldsymbol{O}_s, \boldsymbol{O}_p], \qquad (2)$$

where $\boldsymbol{W}_o \in \mathbb{R}^{C \times (C_s+6)}$, $\cdot$ is matrix multiplication and $[\cdot, \cdot]$ denotes the concatenation operation along rows.

3) Video-level and Object-level Representation Fusion:
To capture the background contextual information and compensate for potentially undetected objects, we augment the object features with video-level features. The final object-level features $\hat{\boldsymbol{O}} \in \mathbb{R}^{TK \times C_o}$ are generated as follows,

$$\hat{\boldsymbol{O}} = \mathrm{MLP}([\boldsymbol{O}, \mathrm{tile}(\boldsymbol{I})]), \qquad (3)$$

where $\mathrm{tile}(\cdot)$ repeats $\boldsymbol{I}$ by K times to $\mathbb{R}^{TK \times C}$ and C_o is the dimension of object node features. MLP is a multilayer perception to project $[\boldsymbol{O}, \mathrm{tile}(\boldsymbol{I})]$ from $\mathbb{R}^{TK \times 2C}$ to $\mathbb{R}^{TK \times C_o}$.

4) Question Representation.: Given a question, we first embed its words into vectors $\boldsymbol{E} \in \mathbb{R}^{L \times 300}$ using a pre-trained GloVe, where L denotes the number of words in the question. Then, we feed the embedded vectors into a BiLSTM to generate a contextualized word-level question embedding $\boldsymbol{Q}_w \in \mathbb{R}^{L \times C_w}$, where C_w is the question feature dimension, and a global or sentence-level question embedding $\boldsymbol{Q}_s \in \mathbb{R}^{1 \times C_w}$, which is the last hidden state of the BiLSTM.

B. Keyword-aware Graph Construction

To correctly answer the question based on a given video, the agent is required to reason about the complex interactions between objects. As videos usually contain multiple objects, it is important to attend to question-relevant objects, which can be seen as keywords, without having the noise from irrelevant objects. However, most existing VideoQA methods with object-level relation modeling treat all the objects equally without considering their relevance to the keywords. Consequently, the redundant information from irrelevant objects may reduce reasoning accuracy. To overcome this limitation, we design a keyword attention module to identify and augment relevant object features for the subsequent relation reasoning. Specifically, we first compute the attended question feature $\hat{\boldsymbol{E}} \in \mathbb{R}^{1 \times 300}$ with the attention mechanism,

$$\boldsymbol{A}_w = \mathrm{softmax}(\mathrm{MLP}(\boldsymbol{Q}_w)), \quad \hat{\boldsymbol{E}} = \boldsymbol{A}_w^T \cdot \boldsymbol{E}, \qquad (4)$$

where $\boldsymbol{A}_w \in \mathbb{R}^{L \times 1}$ is the attention weights. Then, we can use the attended question feature to guide the generation of object attention,

$$\boldsymbol{A}_o = \sigma\big(\hat{\boldsymbol{O}} \cdot (\boldsymbol{W}_q \cdot \hat{\boldsymbol{E}}^T)\big), \qquad (5)$$

where σ is the sigmoid function, $\boldsymbol{W}_q \in \mathbb{R}^{C_o \times 300}$ and $\boldsymbol{A}_o \in \mathbb{R}^{TK \times 1}$. Thus, the keyword-relevant objects are highlighted by assigning higher attention scores. Consequently, these objects will be attended in relation reasoning. Finally, the object node features are generated as follows,

$$\boldsymbol{V} = \hat{\boldsymbol{O}} + \boldsymbol{A}_o \odot \hat{\boldsymbol{O}}, \qquad (6)$$

where $\odot$ denotes the Hadamard product and $\boldsymbol{V} \in \mathbb{R}^{TK \times C_o}$.

C. Relative Spatial-Temporal Graph Reasoning

Most existing VideoQA methods model the complex object relations by representing a video as a holistic spatio-temporal graph over all the detected object proposals [10], [11]. These methods usually simply aggregate neighboring nodes with

affinity metrics, in which the relative relation information is largely ignored. However, relations are relative. Choosing different objects of interest may lead to different semantic and position relation reasoning. Taking the example of a writer writing a book, if the writer is chosen as the subject and the book as the object, the relation is "write". Conversely, the relation is "author", *i.e.*, the book's author is the writer. Similarly, for the position relation, a cup on the table can also be understood as the table under the cup. Therefore, we integrate the relative relation modeling into our graph.

Let $\boldsymbol{v}_t^i \in \mathbb{R}^{1 \times C_o}$ denote the feature of the i-th object or node in the t-th frame of $\boldsymbol{V}$. Our relative-relation-augmented graph models the relations between the object and its neighbors as follows,

$$\boldsymbol{v'}_t^i = \max_{\boldsymbol{v}_{t'}^{i'} \in \mathcal{N}(\boldsymbol{v}_t^i)} \boldsymbol{W}_a \cdot \boldsymbol{v}_{t'}^{i'} + \boldsymbol{W}_r \cdot (\boldsymbol{v}_{t'}^{i'} - \boldsymbol{v}_t^i), \qquad (7)$$

where $\mathcal{N}(\boldsymbol{v}_t^i)$ denotes the k-nearest neighbors of $\boldsymbol{v}_t^i$ in the representation space, $\boldsymbol{W}_a \in \mathbb{R}^{C \times C_o}$ is for absolute relation modeling and $\boldsymbol{W}_r \in \mathbb{R}^{C \times C_o}$ is for relative relation reasoning. Thus, our graph will be able to realize the relativity in relations.

D. Disentangled Spatial-Temporal Graph

Spatial and temporal relations are two different types of relations. Modeling them together may confuse networks and reduce the reasoning efficacy. Moreover, representing a video as a graph over all the objects is computationally expensive. Efficient information flow in such a large graph is challenging. Therefore, we propose to disentangle the graph in Eq. (7) into a spatial graph and a temporal graph,

$$\begin{aligned} \text{spatial}: \quad & \boldsymbol{s}_t^i = \max_{\boldsymbol{v}_t^{i'} \in \mathcal{N}(\boldsymbol{v}_t^i)} \boldsymbol{W}_a^s \cdot \boldsymbol{v}_t^{i'} + \boldsymbol{W}_r^s \cdot (\boldsymbol{v}_t^{i'} - \boldsymbol{v}_t^i), \\ \text{aggregation}: \quad & \boldsymbol{f}_t = \max_{i=1}^{K} \boldsymbol{s}_t^i, \\ \text{temporal}: \quad & \boldsymbol{t}_t = \sum_{\boldsymbol{f}_{t'} \in \mathcal{N}(\boldsymbol{f}_t)} \boldsymbol{W}_a^t \cdot \boldsymbol{f}_{t'} + \boldsymbol{W}_r^t \cdot (\boldsymbol{f}_{t'} - \boldsymbol{f}_t), \end{aligned} \qquad (8)$$

where $\boldsymbol{W}_a^s, \boldsymbol{W}_r^s, \boldsymbol{W}_a^t, \boldsymbol{W}_r^t \in \mathbb{R}^{C \times C_o}$ are the parameters for spatial and temporal reasoning.

In the spatial reasoning and aggregation parts, we use the max pooling operation to keep the most relevant nodes to the question and ignore the noise from the other irrelevant objects. In the temporal reasoning part, we employ the sum pooling operation. This is because in a short video clip, the object of interest usually appears in the entire video and the spatial relation in each frame is important to the overall dynamics reasoning. As shown in experiments, this max-sum pooling combination can achieve better accuracy than others, also demonstrating the benefits of this disentangled spatio-temporal modeling.

To generate the final output vector for answer prediction, we use bilinear attention [42] to project spatial graph presentations $(\{\boldsymbol{s}_t^i\}, \boldsymbol{Q}_w)$ and temporal graph presentations $(\{\boldsymbol{t}_t\}, \boldsymbol{Q}_w)$ to the same space as the question word-level features $\boldsymbol{Q}_w$'s $\mathbb{R}^{L \times C_w}$, respectively. Then, we employ the attention mechanism in [11] to fuse the projected spatial features, temporal features and the question sentence-level features $\boldsymbol{Q}_s$ to a vector, which is finally used to answer the question.

TABLE I
RESULTS ON THE TGIF-QA DATASET. GLOVE IS USED FOR WORD
EMBEDDING.

Model	Action↑	Transition↑	Frame↑	Count↓
ST-TP [43]	62.9	69.4	49.5	4.32
Co-mem [29]	68.2	74.3	51.5	4.10
PSAC [44]	70.4	76.9	55.7	4.27
QueST [45]	75.9	81.0	59.7	4.19
HCRN [6]	75.0	81.4	55.9	3.82
HGA [7]	75.4	81.0	55.1	4.09
B2A [8]	75.9	82.6	57.5	3.71
L-GCN [9]	74.3	81.1	56.3	3.95
GMN [10]	73.0	81.7	57.5	4.16
HOSTR [36]	75.0	83.0	58.0	3.65
MASN [11]	84.4	87.4	59.5	3.75
HQGA [46]	76.9	85.6	**61.3**	-
KRST (ours)	**85.0**	**88.8**	60.9	**3.62**

TABLE II
RESULTS ON THE MSVD-QA AND MSRVTT-QA DATASETS. GLOVE IS
USED FOR WORD EMBEDDING.

Model	MSVD-QA	MSRVTT-QA
AMU [32]	32.0	32.5
ST-TP [43]	30.9	31.3
Co-mem [29]	31.7	31.9
QueST [45]	36.1	34.6
HCRN [6]	36.1	35.6
HGA [7]	34.7	35.5
B2A [8]	37.2	36.9
L-GCN [9]	34.3	-
GMN [10]	35.4	36.1
HOSTR [36]	39.4	35.9
MASN [11]	38.0	35.2
HQGA [46]	41.2	37.2
KRST (ours)	**41.5**	**37.4**

E. Answer Decoder

VideoQA generally includes two types of tasks: multi-choice and open-ended. In this section, we design the answer decoders to deal with different tasks. The multi-choice task aims to choose the correct answer from M candidates. In this case, we concatenate the question with each answer and obtain M question-answer sequences. For each sequence, we feed it into the network to compute the final output vector, and then employ a fully-connected layer to generate the predicted score a_i. Suppose a^+ is the score of the correct answer and $(a_1^-, \ldots, a_{M-1}^-)$ are the scores of wrong answers. We use the pairwise hinge loss $\sum_{i=1}^{M-1} \max\left(0, 1 - (a^+ - a_i^-)\right)$ to train networks.

The open-ended task aims to choose the correct answer from a pre-defined answer set, which can be treated as a multi-label classification task. Therefore, we feed the output of our GNN into a classifier with two fully-connected layers to compute the class probabilities and train the network using the cross-entropy loss. For counting, *i.e.*, open-ended numbers, we treat it as a regression task and train the network using the Mean Squared Error (MSE) loss.

IV. EXPERIMENTS

A. Datasets

We evaluate the proposed method on the three most commonly used benchmarks for VideoQA task.

a) TGIF-QA: It is a large-scale VideoQA dataset with $165K$ Q&A pairs from $72K$ animated GIF videos [43]. There are four types of questions in this dataset: 1) *Action*: multiple-choice questions to identify the action repeated for certain times; 2) *Transition*: multiple-choice questions to identify the action regarding state transition; 3) *FrameQA*: open-ended questions that can be inferred from one frame in a video; 4) *Count*: open-ended questions to count the number of occurrences of an action. Multiple-choice questions have five options, while open-ended questions have a pre-defined answer set of size $1,746$.

b) MSVD-QA & MSRVTT-QA: Both datasets contain $50K$ Q&A pairs from 2K short videos [32] and $243K$ Q&A pairs from 10K short videos [47], respectively. They cover five different question types: what, who, how, when and where. The questions are open-ended with a pre-defined answer set of size $1K$ and $4K$, respectively.

c) Evaluation Metrics: We use MSE for the *Count* task in TGIF-QA and accuracy for other tasks.

B. Implementation Details

We extract videos at 10 frames per second for all the datasets, and uniformly sample T frames from each video. According to the average video length, we set T as 30 for MSVD-QA & MSRVTT-QA datasets and 20 for TGIF-QA dataset. For the global context feature extraction, we apply ResNet-152 model pretrained on the ImageNet [48] dataset for appearance features and I3D model pre-trained on the Kinetics [39] dataset for motion features. For the local feature extraction, we select $K = 10$ object bounding boxes with the highest confidence scores for each frame, where the bounding boxes are generated using Faster R-CNN [41] pretrained on the Visual Genome [49] dataset. We set the number of graph layers H in both spatial and temporal graphs as 2 to capture spatial and temporal information. The ratio α to control number of neighboring nodes in spatial and temporal graphs is set as 0.6 and 0.8, respectively. Moreover, the model's hidden size d is set as 512.

The proposed model is trained on PyTorch [50]. During training, we set batch size as 64 for multi-choice questions and 128 for open-ended and count questions. The dropout ratio is set as 0.3 to prevent overfitting. We train the model for 30 epochs using the Adam optimizer with a constant learning rate of 10^{-4}. Experiments are implemented on Ubuntu 20.04 by NVIDIA GTX 3090 GPUs.

C. Comparison with the State-of-the-art

Table I and Table II present the performance comparison of our KRST with multiple SOTA methods on three popular benchmarks. The comparing methods can be classified

Fig. 3. Impact of the ratio α of neighbors for graph local relation modeling. The blue (orange) bar labelled as spatial (temporal) indicates changing α in spatial (temporal) graphs and fixing α in temporal (spatial) graphs. The lower the better for *Count*.

into three categories: **(1) attention-based methods**, including ST-TP [43], Co-mem [29], PSAC [44], QueST [45] and HCRN [6]; **(2) graph-based methods on frame relations**, including HGA [7] and B2A [8]; **(3) graph-based methods on object relations**, including L-GCN [9], GMN [10], HOSTR [36], MASN [11] and HQGA [46]. The experimental results demonstrate that our model consistently outperforms the SOTA models on three datasets. Note that it is challenging to achieve the best accuracies on all of the three datasets. For example, MASN [11] achieves quite a high accuracy on the TGIF-QA dataset but only reaches inferior accuracies on the MSVD-QA and MSRVTT-QA datasets. Similarly, HQGA [46] can achieve satisfactory performance on MSVD-QA and MSRVTT-QA but not on TGIF-QA (especially for the action task). In contrast, experiments show the superiority of our method on the three datasets, which implies that our method can effectively capture complex object interactions in space and time for VideoQA.

HGA and B2A are graph-based methods that model the temporal relations between different frames. In these methods, multiple graphs are constructed to exploit the correlation between question words and video segments. However, they ignore the fine-grained object-level information important for question answering, leading to inferior performance.

GMN and HOSTR build a holistic spatio-temporal graph to model object relations by taking object trajectories as input. They assume objects exist throughout the video, which may be invalid for videos with multiple objects and heavy occlusions. This assumption may bring extra noise in object relation modeling and result in inferior performance. MASN and HQGA are similar to our work in taking independent object proposals as input. However, both MASN and HQGA perform relation reasoning ignoring the keywords from questions, and they do not consider the relative relation between objects. By filling these gaps, our model demonstrates a clear superiority over these SOTA methods on all the three datasets.

Graphs	Pooling			Action↑	Transition↑	Frame↑	Count↓
	Max	Mean	Sum				
spatial	✓			**85.0**	**88.8**	**60.9**	**3.62**
		✓		84.1	88.4	60.4	3.73
			✓	84.0	88.0	60.3	3.72
aggregation	✓			**85.0**	**88.8**	**60.9**	**3.62**
		✓		84.7	88.6	60.8	3.65
			✓	84.9	88.7	60.7	3.66
temporal	✓			84.7	88.7	60.7	3.67
		✓		84.5	88.6	60.6	3.70
			✓	**85.0**	**88.8**	**60.9**	**3.62**

Fig. 4. Impact of the number of graph layers H. The lower the better for *Count*.

D. Ablation Study

In this section, we conduct a set of ablation studies on TGIF-QA dataset. First, we study the impact of important hyper-parameters (number of neighbors in graphs and number of graph layers). Then, we compare different pooling methods in graph construction. Lastly, we verify the effectiveness of each proposed component.

a) Number of neighbors in graphs: The hyper-parameter $\alpha \in [0, 1]$ is a ratio to control the number of neighboring nodes in graphs. A higher value of α means information from more neighboring nodes is aggregated to compute the center node. To study the impact of α, we train the model using different values of α. Note that we fix α as 0.6 (0.8) in spatial (temporal) graphs when changing α in the temporal (spatial) graphs. As shown in Fig. 3, the hyper-parameter α has significant impact on model performance, and both a lower and a higher α may lead to a performance drop. The reason could be that a higher α may bring much noise from irrelevant objects, while a lower α may miss necessary relation information. Furthermore, there may exist redundant object proposals with poor quality, so α in spatial graphs should be relatively lower than that in temporal graphs.

b) Number of graph layers H: We train our model using different values of H and compare the performance in Fig. 4. It is observed that model performance on different tasks reacts

Question: What does the dog do before place a cat on the table ?   **Ground truth**: grab food and a can
Prediction: grab food and a can

Fig. 5. Visualization of (a) word attention $\boldsymbol{A}_w$, (b) object attention $\boldsymbol{A}_o$ and (c) the k-nearest ($k = 4$) relevant neighbors of the dog according to the question. In (a) and (b), our method pays high attention to the keywords and the relevant objects, such as "dog", "cat", *etc*. In (c), our method selects the most relevant objects as the neighbors in the representation space for graph reasoning.

TABLE IV
ABLATION STUDY FOR EACH PROPOSED COMPONENT.

Model	Action↑	Transition↑	Frame↑	Count↓
Keyword Attention				
w/o Word Attention	84.2	87.9	60.3	3.70
w/o Object Attention	83.6	87.0	59.7	3.81
Relative Reasoning				
w/o Relative Relation	83.8	87.2	59.8	3.70
w/o Absolute Relation	84.0	87.1	60.1	3.72
Disentangled Graphs				
w/o Disentangling	83.3	86.7	60.2	3.76
KRST full model	**85.0**	**88.8**	**60.9**	**3.62**

differently towards the increase of H. On *FrameQA* task, the model achieves the best performance when $H = 1$ and starts to drop as H continues increasing. This suggests that a one-layer graph is sufficient for *FrameQA* task, because questions in this task focus on spatial relations and generally do not require complex relation reasoning. On the other tasks, the model achieves the best performance when $H = 2$, starting to drop as H continues increasing. This suggests that one-layer graph may be insufficient to perform complex object relation reasoning, while a deep graph model may lead to the over-smoothing problem, resulting in inferior performance.

c) Impact of pooling methods: We explore different pooling methods in Eq. 8 and summarize the results in Ta-

ble III. It is observed that maximum pooling generates the best performance for spatial graphs while sum pooling generates the best performance for temporal graphs. As for aggregation, the three pooling methods have similar performance, with mean pooling showing the best result.

d) Effectiveness of each proposed component: To verify the effectiveness of each component in the proposed method, we train ablation models under different settings and summarize the results in Table IV. Overall, we find that removing any component in KRST would reduce the model performance. The impact of each component is detailed as follows.

Keyword attention mainly includes word attention $\boldsymbol{A}_w$ and object attention $\boldsymbol{A}_o$. $\boldsymbol{A}_w$ is used to generate the attended question feature $\hat{E}$, and then $\hat{E}$ is used to compute $\boldsymbol{A}_o$. By removing word attention, we replace $\hat{E}$ using the global question embedding to compute $\boldsymbol{A}_o$. We find that removing word attention will lead to certain performance drops. This is because the model may not effectively identify all the relevant objects when the keywords are not highlighted. By dropping object attention (*i.e.*, the Keyword Attention module), we observe a more significant performance drop. This is as expected since treating all the objects equally may reduce the reasoning efficiency.

Relative reasoning aims to augment absolute relation with relative relation graph reasoning. It is observed that when removing either the relative relation or the absolute relation in graph modeling, the model performance will drop by

(a) Action

Question: What does the blue ball do 4 times? **Ground truth**: Bounce
Predictions: HGA: Kick boy   MASN: Shake hand   Ours: Bounce

(b) Transition

Question: What does the man do after look at pool? **Ground truth**: Pick up bottle
Predictions: HGA: Raise arm   MASN: Smile   Ours: Pick up bottle

(c) FrameQA

Question: How many people inside a room yell into a telephone? **Ground truth**: Three
Predictions: HGA: One   MASN: Two   Ours: Three

(d) Count

Question: How many times does the man with a black shirt stroke a guitar? **Ground truth**: 3
Predictions: HGA: 2   MASN: 4   Ours: 3

(e) Failure Case

Question: How many men are break dancing in synchronization ? **Ground truth**: Three
Predictions: HGA: One   MASN: One   Ours: Two

Fig. 6. Examples from the TGIF-QA dataset. The examples cover four different tasks: Action, Transition, FrameQA and Count. Correct answers (Ground truths) are shown in green and wrong predictions are in red. Rectangles denote the detected object bounding boxes.

around 1% on *Action*, *Transition* and *FrameQA* tasks. These results demonstrate that both relative and absolute relations are important in modeling object interactions.

Disentangled graphs model the spatial relation and temporal relations using separate graphs. We study the impact of disentangled graphs by constructing a holistic graph over all the detected objects, where the spatial and temporal relations are modeled equally. We observe that without disentanglement, the model performance drops by only 0.7% on *FrameQA* task, but by around 2% on *Action* and *Transition* task. This

is because the latter two tasks are more challenging as they require reasoning about both spatial and temporal relations, while a holistic graph lacks the capability to handle such scenarios. This confirms the effectiveness of disentangled graphs in performing complex relation reasoning.

E. Qualitative Results

We visualize the learned attention maps (word attention A_w and object attention A_o) and the k-nearest neighbors of the object. In Fig. 5 (a), we can observe that the keywords such

as "dog", "cat" and "food" are highlighted by higher attention weights. These highlighted keywords can help to identify the question-relevant objects for the subsequent relation reasoning. In Fig. 5 (b), we present the detected objects with their learned attention weights (red numbers) which are also indicated by the brightness of the object regions. It is shown that the keyword-relevant objects (*e.g.*, "dog", "cat" and "food") have higher attention weights and therefore are much more brighter than the keyword-irrelevant objects (*e.g.*, "closet", "oven" and "jar"). This implies that the most keyword-relevant objects are effectively identified from the video frames by object attention. In Fig. 5 (c), we visualize the k-nearest neighbors of the object "dog". It is observed that our model can correctly select the most relevant objects as neighbors, which help improve the efficacy of relation reasoning.

We also compare our method with the existing SOTA methods on different tasks of the TGIF-QA dataset in Fig. 6. The comparing methods include two graph-based methods: 1) HGA [7] on frame relations and 2) MASN [11] on object relations. It is observed that our model can correctly answer the questions based on given videos, which demonstrates the efficacy of our model in object relation reasoning. In Fig. 6 (a), our model can attend to the relevant object (*i.e.*, "blue ball") from other distracting objects (*e.g.*, "green ball") and generate the correct answer. This validates the effectiveness of the keyword attention module in filtering out the question-irrelevant features from videos. In Fig. 6 (b), the example shows that our model can correctly predict the answer to the question, which requires the modeling of both spatial relations (*e.g.*, "pick up bottle") and temporal dynamics (*e.g.*, "after"). In Fig. 6 (c), our model correctly answers the question by aggregating the information in multiple frames. The reason could be that we apply max pooling to aggregate information from neighboring frames in the temporal graph to maintain critical spatial relations in each frame. In Fig. 6 (d), our model generates the correct answer by attending to the question-relevant object (*i.e.*, "man with a black shirt"). This again demonstrates the effectiveness of the proposed keyword attention module. In Fig. 6 (e), all the three models generate incorrect predictions for this sample. The primary cause could be the object detector's inability to identify all the individuals presenting in the video, due to the heavy occlusions and poor lighting conditions.

V. Conclusion

In this paper, we present a Keyword-aware Relative Spatio-Temporal (KRST) graph network for VideoQA. Specifically, we apply attention mechanism to generate a keyword-aware question embedding for the construction of video graphs. To better capture the spatio-temporal relation among object nodes, we introduce relative relation modeling into graph networks. Furthermore, to explicitly model the spatial and temporal relations, we disentangle the holistic spatio-temporal graph into a spatial graph over objects and a temporal graph over frames. Extensive experiments on three datasets demonstrate the effectiveness of our proposed method in performing complex object relation reasoning for VideoQA. In future work,

we plan to exploit the hierarchical structure of questions and videos, which may improve the relation reasoning by building fine-grained correspondences between linguistic and visual elements.

References

[1] W. Zhu, X. Wang, and W. Gao, "Multimedia intelligence: When multimedia meets artificial intelligence," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1823–1835, 2020.
[2] X. Yang, F. Liu, and G. Lin, "Effective end-to-end vision language pre-training with semantic visual loss," *IEEE Transactions on Multimedia*, pp. 1–10, 2023.
[3] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Video storytelling: Textual summaries for events," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 554–565, 2020.
[4] W. Zhang, S. Tang, Y. Cao, S. Pu, F. Wu, and Y. Zhuang, "Frame augmented alternating attention network for video question answering," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1032–1041, 2020.
[5] X. Zhang, F. Zhang, and C. Xu, "Explicit cross-modal representation learning for visual commonsense reasoning," *IEEE Transactions on Multimedia*, vol. 24, pp. 2986–2997, 2022.
[6] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
[7] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
[8] J. Park, J. Lee, and K. Sohn, "Bridge to answer: Structure-aware graph interaction network for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
[9] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[10] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE Transactions on Image Processing*, vol. 30, pp. 2758–2770, 2021.
[11] A. Seo, G.-C. Kang, J. Park, and B.-T. Zhang, "Attend what you need: Motion-appearance synergistic networks for video question answering," in *ACL*, 2021, pp. 6167–6177.
[12] J. Li, X. He, L. Wei, L. Qian, L. Zhu, L. Xie, Y. Zhuang, Q. Tian, and S. Tang, "Fine-grained semantically aligned vision-language pre-training," *arXiv preprint arXiv:2208.02515*, 2022.
[13] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, "Dense video captioning using graph-based sentence summarization," *IEEE Transactions on Multimedia*, vol. 23, pp. 1799–1810, 2021.
[14] H. Fan, L. Zhu, Y. Yang, and F. Wu, "Recurrent attention network with reinforced generator for visual dialog," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–16, 2020.
[15] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8746–8755.
[16] X. Wang, L. Zhu, and Y. Yang, "T2vlad: global-local sequence alignment for text-video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5079–5088.
[17] S. Zhao, L. Zhu, X. Wang, and Y. Yang, "Centerclip: Token clustering for efficient text-video retrieval," *arXiv preprint arXiv:2205.00823*, 2022.
[18] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, 2017.
[19] X. Yang, S. Wang, J. Dong, J. Dong, M. Wang, and T. Chua, "Video moment retrieval with cross-modal neural architecture search," *IEEE Trans. Image Process.*, vol. 31, pp. 1204–1216, 2022.
[20] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, and T.-S. Chua, "Video question answering: datasets, algorithms and challenges," *arXiv preprint arXiv:2203.01225*, 2022.
[21] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[22] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 634–23 651, 2021.

[23] W. Yu, H. Zheng, M. Li, L. Ji, L. Wu, N. Xiao, and N. Duan, "Learning from inside: Self-driven siamese sampling and reasoning for video question answering," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 462–26 474, 2021.

[24] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4953–4963.

[25] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[26] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *Conference on Empirical Methods in Natural Language Processing*, 2018.

[27] X. Li, L. Gao, X. Wang, W. Liu, X. Xu, H. T. Shen, and J. Song, "Learnable aggregating net with diversity learning for video question answering," *ACMMM*, pp. 1166–1174, 2019.

[28] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, and Y. Zhuang, "Multi-interaction network with object relation for video question answering," *ACMMM*, pp. 1193–1201, 2019.

[29] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu, "Multi-turn video question answering via multi-stream hierarchical attention context network." in *IJCAI*, vol. 2018, 2018, p. 27th.

[31] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang, "Multimodal dual attention memory for video story question answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[32] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

[33] J. Li, J. Xie, L. Qian, L. Zhu, S. Tang, F. Wu, Y. Yang, Y. Zhuang, and X. E. Wang, "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3032–3041.

[34] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2765–2775.

[35] J. Li, S. Tang, L. Zhu, H. Shi, X. Huang, F. Wu, Y. Yang, and Y. Zhuang, "Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1867–1877.

[36] L. H. Dang, T. M. Le, V. Le, and T. Tran, "Hierarchical object-oriented spatio-temporal reasoning for video question answering," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, August 2021.

[37] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[40] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[42] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018.

[43] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2758–2766.

[44] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[45] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao, "Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[46] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T.-S. Chua, "Video as conditional graph hierarchy for multi-granular question answering." AAAI, 2022.

[47] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[49] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.