Video Question Answering With Prior Knowledge and Object-Sensitive Learning

Pengpeng Zeng, Haonan Zhang, Lianli Gao, *Member, IEEE*, Jingkuan Song, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*— Video Question Answering (VideoQA), which explores spatial-temporal visual information of videos given a linguistic query, has received unprecedented attention over recent years. One of the main challenges lies in locating relevant visual and linguistic information, and therefore various attention-based approaches are proposed. Despite the impressive progress, two aspects are not fully explored by current methods to get proper attention. Firstly, prior knowledge, which in the human cognitive process plays an important role in assisting the reasoning process of VideoQA, is not fully utilized. Secondly, structured visual information (*e.g.,* object) instead of the raw video is underestimated. To address the above two issues, we propose a Prior Knowledge and Object-sensitive Learning (PKOL) by exploring the effect of prior knowledge and learning object-sensitive representations to boost the VideoQA task. Specifically, we first propose a Prior Knowledge Exploring (PKE) module that aims to acquire and integrate prior knowledge into a question feature for feature enriching, where an information retriever is constructed to retrieve related sentences as prior knowledge from the massive corpus. In addition, we propose an Object-sensitive Representation Learning (ORL) module to generate object-sensitive features by interacting object-level features with frame and clip-level features. Our proposed PKOL achieves consistent improvements on three competitive benchmarks (*i.e.*, MSVD-QA, MSRVTT-QA, and TGIF-QA) and gains state-of-the-art performance. The source code is available at https://github.com/zchoi/PKOL.**

***Index Terms*— Video question answering, prior knowledge, object learning.**

I. Introduction

THE task of Video Question Answering (VideoQA) refers to answer a natural language question based on the visual contents (*i.e.*, events, actions, and entities), which needs to understand the questions, recognize the visual content of different modalities, and reason about the accurate answers. Despite the significant improvement made in VideoQA in recent years [1], [2], [3], [4], [5], there are still challenges in VideoQA due to the complex reasoning procedure and the huge semantic discrepancies between videos and questions.

One of the promising and scalable solutions for learning VideoQA is the attention mechanism, which aims to discover the key visual content to improve answer generation, such as [4], [5], [6], [7], [8], [9]. Specifically, humans tend to ask questions about the salient objects, the correspondent actions, or the relationships between two objects in a video. Therefore, applying the attention mechanism to attend to key frames or video clips, to a certain extent, is beneficial for VideoQA. For instance, [8] proposes a positional self-attention mechanism to replace the classic recurrent structure and explores a parallel encoding of temporal and question features with a co-attention. [4] introduces a hierarchical attention structure to capture appearance-question and motion-question relations from frame-level to clip-level. [5] proposes an attention-based graph network, which is stacked in an iterative manner to perform multi-step reasoning. The performance of VideoQA has been significantly improved by various attention mechanisms to encode powerful features or capture accurate alignments.

However, attention-based approaches suffer from two problems. Firstly, they are restricted to exploring information of videos and questions without considering prior knowledge. This is not consistent with the nature of the human cognitive process, where prior knowledge plays an important role in assisting the reasoning process of VideoQA. More specifically, when being asked a video-related question, a human being habitually retrieves empirical information from his/her brain as a guide to support the reasoning process [10]. This is because prior knowledge associated with the video provides additional hints. Taking Fig. 1 (a) as an example, before answering a question “What did the man break the block of ice with?”, a human being is capable of capturing keywords (*e.g.,* “man” and “ice”) within the question and obtaining a preliminary impression of the video. Next, the impression helps to retrieve related prior knowledge. For instance, the returned sentences containing key words, *e.g.,* “pick axe”, “ice cube” and “handled hammer”, provide hints for reasoning about the accurate answer. Therefore, it is reasonable and necessary to integrate prior knowledge for VideoQA. Secondly, they are all object-insensitive and neglect structured visual information (*e.g.,* object). Specifically, attention-based mechanism focuses

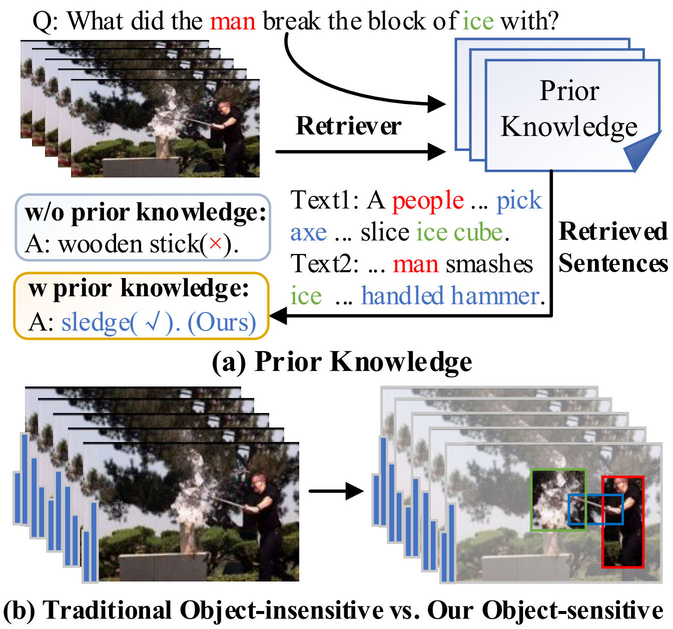
Manuscript received 25 January 2022; revised 18 June 2022 and 17 July 2022; accepted 17 August 2022. Date of publication 9 September 2022; date of current version 16 September 2022. This work was supported in part by the Chinese National Science & Technology Pillar Program under Grant 2022YFC2009900 and Grant 2022YFC2009903; and in part by the National Natural Science Foundation of China under Grant 62122018, Grant 62020106008, Grant 61772116, and Grant 61872064. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Chuang Gan. *(Corresponding author: Heng Tao Shen.)*

Pengpeng Zeng and Haonan Zhang are with the Future Media Center and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: is.pengpengzeng@gmail.com; zchiowal@gmail.com).

Lianli Gao is with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518028, China (e-mail: lianli.gao@uestc.edu.cn).

Jingkuan Song and Heng Tao Shen are with the Future Media Center and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: jingkuan.song@gmail.com; shenhengtao@hotmail.com).

Digital Object Identifier 10.1109/TIP.2022.3205212


Q: What did the man break the block of ice with?
Retriever
Prior Knowledge
w/o prior knowledge:
A: wooden stick(×).
Text1: A people ... pick axe ... slice ice cube.
Text2: ... man smashes ice ... handled hammer.
Retrieved Sentences
w prior knowledge:
A: sledge(✓). (Ours)

(a) Prior Knowledge

(b) Traditional Object-insensitive vs. Our Object-sensitive

Fig. 1. Existing attention-based approaches, to a certain extent, are not consistent with human cognition in VideoQA and suffer from two problems: a) lacking prior knowledge and b) being object-insensitive. In this paper, we explore the role of prior knowledge in the reasoning of VideoQA as well as propose an object-sensitive approach.

on how to attend key video frames or video clips guided by questions and fails to capture specific information about object regions. In reality, object-related questions are widely distributed in the VideoQA dataset. When answering an object-related question, a human being mainly pays attention to critical object regions and object-related motions. Therefore, an object-sensitive approach should be designed and explored (see Fig. 1 (b)).

Based on the above insights, we propose a novel **P**rior **K**nowledge and **O**bject-sensitive **L**earning (**PKOL**), termed as **PKOL**, for video question answering. Specifically, the pipeline of PKOL mainly consists of two major components: 1) **Prior Knowledge Exploring (PKE)** module, which incorporates the prior knowledge into the question feature in order to facilitate a more comprehensive understanding of the question and video. It provides additional evidence for VideoQA reasoning. In this module, we introduce a pluggable knowledge information retriever to search for the Top-K video-question-relevant sentences from a textual corpus as the prior knowledge based on queries of the input video and the given question. In the training phase, the corpus we adopted contains all descriptions about the videos in the training dataset, due to the laborious and time-consuming collection of the high quality data. In the inference stage, the corpus can be any text descriptions related to videos, such as the descriptions of the test set or the descriptions of the training set and validation set. 2) **Object-sensitive Representation Learning (ORL)** module, which learns object-sensitive features that are relevant to the content of the question, in particular the objects mentioned in the question. Essentially, our ORL model is not a substitute for attention but a revamp. It respectively integrates object features with appearance and motion features to extract object-sensitive appearance and motion features. Although the proposed co-attention, self-attention, or hierarchical attention [7], [8], [11] have the potential to boost the performance of VideoQA, their structures are complex and time-consuming. Therefore, we intend to combine both our PKE and ORL modules with vanilla attention. We extensively evaluate our PKOL on the three publicly benchmark datasets: MSVD-QA, MSRVTT-QA, and TGIF-QA, where experiments prove the effectiveness of our proposed model.

Our contributions can be summarized as follows:

- We propose a novel Prior Knowledge Exploring **(PKE)**, which introduces the prior knowledge in an information retrieval manner to facilitate the reasoning capability of a VideoQA model. Besides, this module can provide additional interpretability by treating the retrieved sentences as the evidence of reasoning.
- We propose a novel Object-sensitive Representation Learning **(ORL)**, which explores semantic rich object representations across both spatial and temporal domains. This module fully considers the structured visual information and it is more in line with the essence of the human attention mechanism.
- We apply PKE module and ORL module into a vanilla attention-based model to build our PKOL, which is more consistent with the process of human cognitive processes in VideoQA. Our proposed model achieves the new state-of-the-arts on three benchmark datasets.

The remainder of this paper is organized as follows. In Sec. II, we provide an overview of related works. Then we introduce our proposed PKOL in Sec. III, which consists of two main components: Prior Knowledge Exploring (PKE) and Object-sensitive Representation Learning (ORL). In Sec. IV, we present a quantitative and qualitative analysis for three benchmarks to evaluate the performance of the proposed method, as well as verify the effects of each component through ablation studies. We summarize our work in Sec. V.

II. Related Works

A. Video Question Answering

As VideoQA task requires fine-grained interaction between vision and language to understand the complex video scenario, it has attracted massive attention [4], [11], [12], [13]. Compared to the traditional visual question answering learning on a static image, VideoQA is a more sophisticated reasoning task including temporal exploration over-frame sequences, object action recognition, and causality in the temporal dimension.

1) Appearance-Motion Based Method: Typically, early progress on VideoQA has featured in learning visual representations by exploiting the interaction between video frames and questions, where one popular approach adopts an attention mechanism. [14], [15] propose a temporal attention mechanism to select the key information through questions as guidance. Next, [12] applies a co-attention to simultaneously localize important visual instances and focus on the relevant text. Moreover, [15] extends a single-path-based co-attention mechanism to a multi-path pyramid co-attention structure for explicit appearance-question learning. However, the above methods only use the appearance information without considering motion information [4], [11], [12], [16] proposes a dual channel to leverage appearance, motion, and question interactions to obtain better video representations. Besides, to capture higher-order semantic information, several methods [4], [17], [18] adopt hierarchical attention to accomplish multi-step reasoning to obtain contextual relationship representation.

2) Object-Centric Based Video Method: Due to the large number of questions with object information in the dataset, object-centric based methods slowly become prevalent. Usually, the object information is obtained by using object detection methods, such as Faster R-CNN [19] and YOLO [20]. The works in [21] and [22] send object features into a generalized relational graph for reasoning. [23] focuses on building an internal understanding of the dynamics and causal relations via detected objects. Although our method belongs to the object-centric based method, we aim to explore richer object-sensitive representation across both spatial and temporal domains, which can helpfully consider the structured visual information and be more in line with the human cognitive processes.

B. Video Reasoning

Different from video question answering, video reasoning aims to reason about temporal and causal events behind the interacting objects from videos. The task raises higher requirements for diverse visual scenes, such as learning action movement [11], focusing on temporal correlation [4], and language cues [24]. Since the existing datasets of VideoQA (*e.g.,* TGIF-QA [11], MSRVTT-QA [25], and MSVD-QA [26]) based on real-world video scenarios are not enough to satisfy the properties of visual reasoning, CLEVRER dataset is proposed to simulate causal relationships grounded on object dynamics and physical interactions. Due to large discrepancies in data domain and task formulation between video reasoning task and VideoQA task, recent researches have specifically designed multi-modal deep networks to address the problem of video inference [27], [28], [29]. Specifically, [27] formulates structure-aware interaction for semantic relation modeling between cross-modal information. [29] focuses on incorporating explicit reasoning in a knowledge graph with implicit reasoning in a multi-modal transformer for answer prediction. Besides, some methods [30], [31], [32], [33] focus on functional program reasoning or solve the problem with physical properties of visual content. For example, [30] proposes a program generator for reading the question and produces a plan and an execution engine for executing the resulting module network on the image to produce an answer. [32] aims to infer the hidden physical properties in videos and not directly observable them from visual appearances.

C. Video-to-Text Retrieval

Video-to-text retrieval is a fundamental and challenging task that requires performing semantic alignment among different modalities in a discriminative manner. Compared with image-text retrieval, video-text retrieval needs dynamically temporal consideration between multiple video frames. The existing methods can be summarized in two categories: Bi-encoder and Cross-encoder.

For Bi-encoder, the raw video frames and text are embedded via two extractors separately and then apply cross-modal interaction via a fusion network. [34] utilizes multi-modal features (different visual characteristics, motion, audio, and text) by a fusion strategy for video-text retrieval. [35] designs an efficient global-local model which aggregates multi-modal video sequences and text into a set of shared centers for semantic alignment and matching. Although they have suggested effective feature extraction based on different modalities, the separated encoding models still fail to bridge the semantic gap between the two modalities. [36], [37] tackle this issue with a Cross-encoder way to integrate rich interaction between query and candidates. [36] introduces a polysemous embedding network to compute multiple and diverse representations of an instance by combining global context with locally-guided features via self-attention and residual learning. To a certain extent, the cross-encoder has refined effective interaction between video and text modalities and achieved higher accuracy, but it also brings a tremendous amount of computational complexity.

III. Approach

In this work, we present a novel **P**rior **K**nowledge and **O**bject-sensitive **L**earning (PKOL) for video question answering (VideoQA). The overview of the proposed PKOL is depicted in Fig. 2.

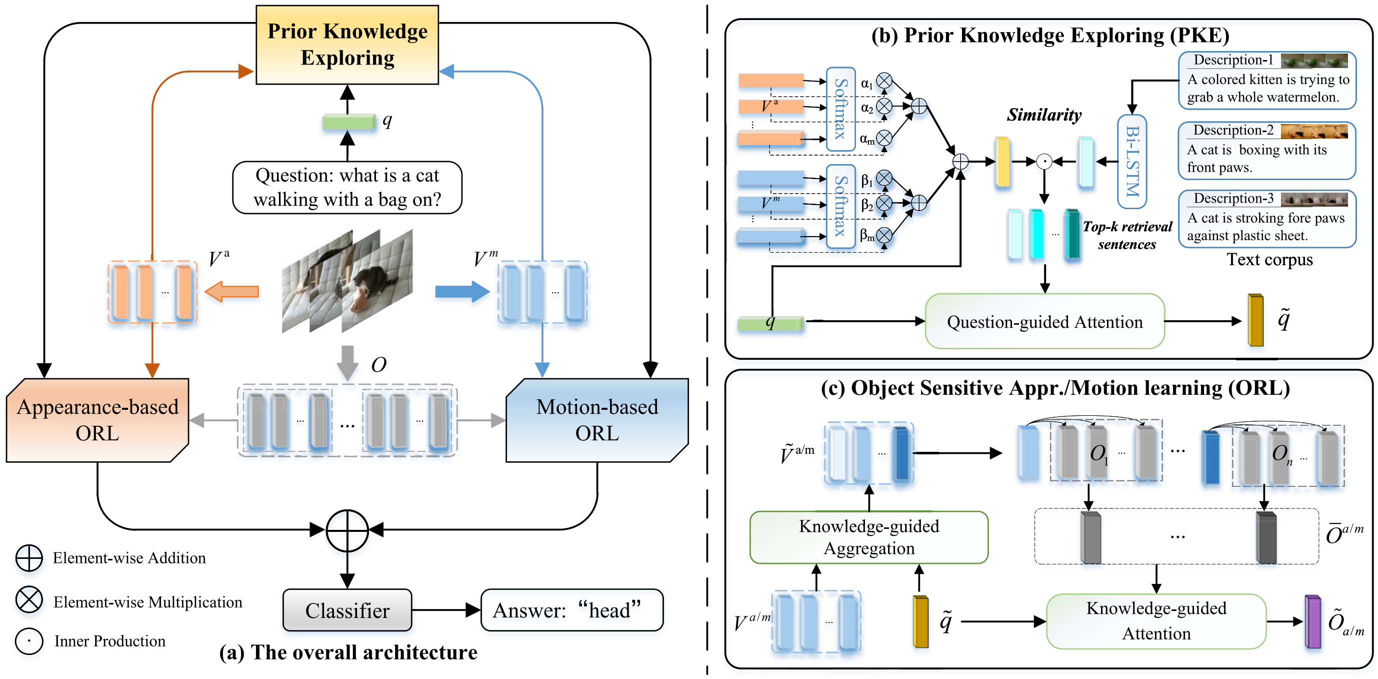
A. Overview

Given a question q and a corresponding video V, the task of VideoQA is to automatically generate an accurate answer a, which conventionally is formulated as follow:

$$\begin{aligned} \tilde{a} &= \underset{a \in A}{argmax}\, \mathcal{F}_{\theta}\left(a|q, V\right), \\ s.t.\ V &= \{V^{a}, V^{m}\}, \end{aligned} \tag{1}$$

where V^a and V^m denote the appearance feature and motion feature, respectively. A represents a pre-defined answer set and θ is the model parameter of the function $\mathcal{F}$, where $\mathcal{F}$ usually adopts the attention-based method.

However, existing methods suffer from two problems, including lacking prior knowledge and being object-insensitive. To address the above problems, we propose a **Prior Knowledge Exploring (PKE)** module and an **Object-sensitive Representation Learning (ORL)** module to explore the effect of prior knowledge and learn object-sensitive representations, respectively. Specifically, the PKE module first applies a knowledge information retriever to search for the video-question-relevant Top-K sentences Z from the prior knowledge corpus. It then utilizes a vanilla question-guided attention to merge the prior knowledge into the question feature to obtain a knowledge-aware question feature $\tilde{q}$. The ORL module aims to integrate object feature O with the appearance



(a) The overall architecture

(b) Prior Knowledge Exploring (PKE)

(c) Object Sensitive Appr./Motion learning (ORL)

Prior Knowledge Exploring
q
Question: what is a cat walking with a bag on?
V^a
V^m
O
Appearance-based ORL
Motion-based ORL
Element-wise Addition
Element-wise Multiplication
Inner Production
Classifier
Answer: "head"
Softmax
α_1 α_2 α_m
β_1 β_2 β_m
Similarity
Bi-LSTM
Top-k retrieval sentences
Description-1 A colored kitten is trying to grab a whole watermelon.
Description-2 A cat is boxing with its front paws.
Description-3 A cat is stroking fore paws against plastic sheet.
Text corpus
Question-guided Attention
$\tilde{q}$
$\tilde{V}^{a/m}$
O_1
O_n
Knowledge-guided Aggregation
$\bar{O}^{a/m}$
$V^{a/m}$
Knowledge-guided Attention
$\tilde{O}_{a/m}$

Fig. 2. Overview of the proposed PKOL architecture for video question answering. The left side is the overall pipeline of our method, which consists of two key components: Prior Knowledge Exploring (PKE) that aims to acquire and integrate prior knowledge into a question feature for feature enriching; Object-sensitive Representation Learning (ORL) that aims to generate object-sensitive feature on the appearance and motion across spatial-temporal. The right side illustrates the details of the PKE and ORL.

and motion features to extract object-sensitive appearance and motion features. Thus, the Eq. 2 can be redefined as:

$$\begin{aligned}\tilde{a} &= \underset{a\in A}{argmax}\, \mathcal{F}_\theta\left(a|\tilde{q},\tilde{O}\right),\\ s.t.\ V &= \{V^a, V^m\},\\ \tilde{q} &= PKE\{q, Z \xleftarrow[Z\in Corpus]{Retriever} (q,V)\},\\ \tilde{O} &= \{ORL_a(O,V^a,\tilde{q}), ORL_m(O,V^m,\tilde{q})\}, \qquad (2)\end{aligned}$$

where *Retriever* denotes the knowledge information retriever. In addition, ORL_a and ORL_m indicate the object-sensitive representation learning on appearance and motion features, respectively. Next, we introduce our PKOL in detail.

B. Feature Embedding

1) Visual Representation: For a T-frame video V, which is divided into N clips $C = \{c_n\}_{n=1}^N$, we adopt the multiple visual feature extractors to obtain visual representations. In this work, 2D-CNNs and 3D-CNNs are utilized to extract appearance features and motion features, respectively. In addition, object feature is extracted by the Faster-RCNN [38] to capture object regions from each frame. Since the dimensions of the extracted features are inconsistent, we use multiple linear transformation layers to map appearance feature, motion feature and object feature into a d-dimensional feature space to obtain $V^a = \{v_t^a\}_{t=1}^T, V^m = \{v_n^m\}_{n=1}^N$ and $O = \{o_t\}_{t=1}^T$, where $o_t = \{o_t^b\}_{b=1}^B$ and B denotes the number of objects in each frame.

2) Question Representation: Given a question Q, each word w_l is initialized with a pre-trained embedding GloVe model, denoted as $Q = \{w_l\}_{l=1}^L$. Next, each word embedding is fed into a Bi-LSTM as follows:

$$\begin{aligned}\vec{h}_l &= \overrightarrow{\text{LSTM}}(w_l, \vec{h}_{l-1}),\\ \overleftarrow{h}_l &= \overleftarrow{\text{LSTM}}(w_l, \overleftarrow{h}_{l+1}),\\ h_l &= (\vec{h}_l + \overleftarrow{h}_l)/2, \qquad (3)\end{aligned}$$

where $\vec{h}_l$ and $\overleftarrow{h}_l$ are the hidden states of forward and backward propagation, respectively. We choose the last output of Bi-LSTM (h_L) as the final question feature q.

3) Prior Knowledge Representation: In this paper, we adopt a massive text corpus related to the video as our prior knowledge, which contains multiple rich semantic sentences and provide useful external knowledge information to help model reasoning and explanation. Different from the question embedding, we replace pre-trained GloVe with pre-trained BERT to capture the better semantic representation. The remaining operations are consistent with the question encoding. Thus, we obtain the feature of each sentence, namely s.

C. Prior Knowledge Exploring

The purpose of this module (PKE) is to acquire and integrate the prior knowledge into a question feature to enrich its representation, including prior knowledge retriever and question-guided attention.

1) Prior Knowledge Retriever: This module aims to retrieve the Top-K relevant sentences from a large corpus as the prior knowledge based on the query information. There are many types of queries, such as question features, video features, and their combinations. Specifically, the appearance feature V^a

and motion feature V^m usually include redundant information, and their dimensions are not consistent with the dimension of question feature. Consequently, we first adopt an aggregation operation that gives higher weights to more representative features to form an enhanced global feature. Taking the appearance feature as an example, the process can be formulated as follows:

$$\begin{aligned} \bar{v}^a &= \sum_{\text{t=1}}^{T} \alpha_t v_t^a, \\ \alpha_t &= \text{softmax}(W_1 v_t^a), \end{aligned} \tag{4}$$

where W_1 is the learnable parameters and all formulas omit bias for the sake of simplicity. We define the above process as $\bar{v}^a = Agg(V^a)$. Thus, by respectively using the aggregation operation on V^a and V^m by Eq. 4, we can obtain two enhanced feature $\bar{v}^a$ and $\bar{v}^m$. The final query feature x combines two enhanced visual features and question feature:

$$x = (\bar{v}^a + \bar{v}^m + q)/3. \tag{5}$$

To measure the correlation between the query and sentence in the corpus, we calculate the similarity among them as follows:

$$\text{sim(s, x)} = s^T x. \tag{6}$$

In this manner, the retriever calculates the similarity of each query to all texts in the corpus to obtain a similarity vector, and sorts the similarity vector to get the Top-K closest sentences, namely $Z = \{z_1, \ldots, z_{topK}\}$.

2) Question-Guided Attention: To draw on the information of multiple retrieved sentences, we first utilize the multiplicative attention-mechanism to attend to the most relevant knowledge of retrieved sentences guided by the question feature:

$$\begin{aligned} \bar{z} &= \sum_{\text{i=1}}^{topK} \beta_i z_i, \\ \beta_i &= \text{softmax}(\mathrm{W_2 \tanh(W_3 z_i \otimes W_4 q)}), \end{aligned} \tag{7}$$

where $W_{2,3,4}$ are the trainable parameters and $\otimes$ is a Hadamard product. We define the above process as $\bar{z} = ATT(Z, q)$, where $\bar{z}$ means the attended knowledge feature. Subsequently, we combine the attended knowledge feature $\bar{z}$ and question feature q via a Fully-Connected (FC) network to obtain the final knowledge-aware question feature $\tilde{q}$ as follows:

$$\tilde{q} = \text{FC}([q; \bar{z}]), \tag{8}$$

where [;] means the operation of concatenation.

D. Object-Sensitive Representation Learning

The module (ORL) aims to learn object-sensitive representations by incorporating object-level region features with appearance or motion features across the space-time. It includes appearance-based object-sensitive representation learning (ORL_a) and motion-based object-sensitive representation learning (ORL_m). Due to the same structure of two sub-modules, we take ORL_a as an example for simplicity. Specifically, the ORL_a consists of a knowledge-guided aggregation module, an object-sensitive attention module and a knowledge-guided attention module.

1) Knowledge-Guided Aggregation: This module focuses on the important information on appearance feature guided by the knowledge-aware question feature. Specifically, we first compute a weight for appearance feature at frame-level and the weight values represent the importance for the question feature. Then, we assign these weights to the appearance features to obtain the weighted apperance feature $\tilde{V}^a$, defined as follows.

$$\begin{aligned} \tilde{V}^a &= \{\tilde{v}_t^a\}_{t=1}^T = \{\gamma_t^a v_t^a\}_{t=1}^T, \\ \gamma_t^a &= \text{softmax}(W_5 \tanh{(W_6 v_t^a \otimes W_7 \tilde{q})}), \end{aligned} \tag{9}$$

where $W_{5,6,7}$ are the trainable parameters.

2) Object-Sensitive Attention: This module aims to interact the object features with appearance features to learn the representative object feature. Specifically, we adopt the multiplicative attention-mechanism to aggregate object features into appearance feature at frame-level to obtain the attended object-appearance feature $\bar{O}^a$.

$$\bar{O}^a = \{ATT(o_t, \tilde{v}_t^a)\}_{t=1}^T. \tag{10}$$

3) Knowledge-Guided Attention: This module is to focus on the key information on appearance-object feature guided by a knowledge-aware question feature and then fuse the above features to generate a final object-sensitive features at the appearance-level ($\tilde{o}_a$), which can be formulated as follow.

$$\tilde{o}_a = FC([ATT(\bar{O}^a, \tilde{q}); \bar{v}^a]). \tag{11}$$

Similar to the operation of ORL_o, ORL_m can obtain the object-sensitive features at the motion-level ($\tilde{o}_m$).

E. Answer Classifier and Training Details

1) Answer Classifier: Similar to the previous works [4], [16], the answer classifiers are split two settings: Open-ended VideoQA setting and Multi-choice VideoQA setting, according to the type of question.

a) Open-ended VideoQA setting: Open-ended question setting is regarded as a multi-label classification task, where a ground-truth answer is selected from a predefined answer set Ω. Specifically, we first feed two object-sensitive visual features ($\tilde{o}_a$ and $\tilde{o}_m$) and knowledge-aware question feature $\tilde{q}$ into a classifier to obtain prediction probabilities about answer:

$$\begin{aligned} f &= \mathrm{W_8[\tilde{o}_a; \tilde{o}_m; \tilde{q}]} \\ p &= \text{softmax}(\mathrm{W_9 f})\ \mathrm{p \in R^{|\Omega|}}, \end{aligned} \tag{12}$$

where $W_{8,9}$ are learnable parameters. Then, we employ a cross-entropy loss to train the VideoQA model:

$$L_{ce} = -\sum_{i=1}^{|A|}[(1 - y_i)\log(1 - p_i) + y_i \log p_i], \tag{13}$$

where y_i is the one-hot encoded vector of the ground truth answer. Different from the above operation, for repetition count task, since the answer ranges from 1 to 10, we adopt a linear regression function to predict an integer-valued answer, which takes the feature f in Eq. 12:

$$\bar{c} = \varphi(W_{10} f), \tag{14}$$

where W_{10} is learnable parameter and φ means rounding. The model is optimized by Mean Square Error loss L_{mse} between the real value s and the predicted value $\bar{s}$:

$$L_{mse} = (c - \bar{c})^2. \tag{15}$$

b) Multi-choice VideoQA setting: In this setting, one correct answer is chosen from M candidate answers to a given question. As a result, the classifier needs to output the probability of each candidate. Specifically, we take two object-sensitive visual features ($\tilde{o}_a$ and $\tilde{o}_m$), knowledge-aware question feature $\tilde{q}$ and one answer feature a_m as inputs into classifier to compute the m-th answer score:

$$c_m = \mathrm{W}_{11}[\tilde{\mathrm{o}}_\mathrm{a}; \tilde{\mathrm{o}}_\mathrm{m}; \tilde{q}; a_\mathrm{m}], \tag{16}$$

where W_{11} is a learnable parameter. In order to train this model, we utilize a pairwise hinge loss L_{ph} between the positive score c^p and each negative score c^n_m:

$$L_{ph} = \sum\nolimits_{\mathrm{m}=1}^{M} \max(0, 1 - (c^p - c^n_m)). \tag{17}$$

2) Training Details: Except for optimizing the answer classifier, we should train our prior knowledge retriever in the PKE. Specifically, we adopt contrastive learning to optimize this retriever by using a triplet loss to measure the similarity between a query x and a sentence s:

$$\begin{aligned} L_{ret} = {} & max(0, \Delta - \mathrm{sim}(x, s) + \mathrm{sim}(x, s^-)) \\ & + max(0, \Delta - \mathrm{sim}(x, s) + \mathrm{sim}(x^-, s)), \end{aligned} \tag{18}$$

where Δ is a slack factor. The (x, s) denotes a positive pair, while (x, s^-) and (x^-, s) denote negative pairs.

For each setting, we separately optimize each model with the corresponding loss function mentioned above and the triplet loss L_{ret}:

$$L = L_{\{ce,mse,ph\}} + L_{ret}. \tag{19}$$

Thus, each model is evaluated separately.

IV. Experiments

In this section, we first describe three widely VideoQA dataset in Sec. IV-A and the detail of our experiments in Sec. IV-B. Then we verify the effectiveness of our proposed components (PKE and ORL) in Sec. IV-C, which is followed by the comparison with the state-of-the-art approaches in Sec. IV-D. Finally, we provide some visualization results to qualitatively analyze the benefit of our method in Sec. IV-E.

A. Dataset and Evaluation

1) Datasets: All the experiments are conducted on three VideoQA bechmark datasets: MSVD-QA, MSRVTT-QA, and TGIF-QA.

MSVD-QA [26] contains 50K QA pairs from 1,970 video clips. For fair comparison, we adopt the split provided by [26] that uses 61% of QA pairs for training, 13% QA pairs for validation and 26% QA pairs for testing. There are five types of questions: *What (62.6%), Who (34.1%), How (2.6%), When (0.5%)* and *Where (0.2%)*. Besides, the dataset provides the additional descriptions of the corresponding video clips, with an average of 30 human descriptions for each clip. Thus, we take these description corpus as prior knowledge set.

MSRVTT-QA [26] contains 243K QA pairs generated from 10K videos, which have more complex video scenarios with an average video length of around 10-15 seconds. Similar to MSVD-QA, each video is annotated with 20 descriptions. It also contains five types of questions: *What (68.5%), Who (27.7%), How (2.5%), When (1.0%)* and *Where (0.3%)*. Such descriptions are considered as the prior knowledge set for knowledge retriever. Following the prior work [25], we separate the dataset into 65% QA pairs for training, 5% for validation, and 30% for testing.

TGIF-QA [11] is a large-scale dataset containing 165K question answer pairs collected from 72K animated GIF files. Each GIF consists of an average of 3 relevant descriptions. TGIF-QA provides four types of tasks: Repeating action (Action), State transition (Trans.), FrameQA, and Repeating Count (Count). Repeating action is a task defined as a multiple-choice question about identifying repeated actions in a video. State transition (Trans.) is a multiple choice task to recognize an action that occurs before (or after) another action state. FrameQA aims to find a specific frame in a video that can answer the questions. Repeating Count (Count) is an open-ended task about counting the number of repetitions of an action in video.

2) Evaluation Metrics: Following the standard evaluation protocol, we adopt Mean Squre Error (MSE) for repetition count on TGIF-QA dataset and accuracy (Acc.) for all the others to validate our $PKOL$ and other VideoQA methods. For the task of retrieval, Recall@K (R@K) is typically applied to measure the performance, where K is set to 1,5, and 10.

B. Implementation Details

1) Feature Extractions: For the visual features, we first uniformly extract frames $L = 128, 64, 128$ for MSVD-QA, MSRVTT-QA, and TGIF-QA, respectively. The clip number is set as $N = 8$ for all. Then, the ResNet [39] and ResNeXt-101 [40], [41] are respectively employed to extract the appearance feature and the motion feature. We apply a Faster-RCNN [38] pre-trained on Visual Genome [42] to extract the object feature in each frame and the number of objects B is 10. For the textual features, sentences longer than 40 words will be truncated. We set all the Bi-LSTM, attention module and fully-connected layer in the model with 512 hidden states.

2) Training Details: We adopt Adam [43] to optimize model training, with an initial learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, then decay by 0.5 rate after every 10 iterations. The slack factor Δ is set as 0.2 in Eq. 18 and dropout probability is 0.15. We set batchsize as 32 for both MSVD-QA, MSRVTT-QA and TGIF-QA datasets. Our model is implemented by PyTorch [44]. The model will converge in around 25 epochs on 1 NVIDIA RTX3090 GPU.

C. Ablation Study

In this section, we conduct ablation study to verify the effectiveness of our proposed Prior Knowledge Exploring (PKE) module and Object-sensitive Representation Learning (ORL) module.

TABLE I

PERFORMANCE COMPARISON OF THE PROPOSED PRIOR KNOWLEDGE EXPLORING (PKE) AND OBJECT-SENSITIVE REPRESENTATION LEARNING (ORL) ON MSVD-QA AND MSRVTT-QA DATASETS. ALL VALUES ARE REPORTED AS ACCURACY (%)

PKE	ORL	MSVD-QA	MSRVTT-QA
✗	✗	37.3	35.5
✓	✗	39.3	36.6
✗	✓	40.0	36.3
✓	✓	**41.1**	**36.9**

TABLE II

PERFORMANCE COMPARISON OF THE PROPOSED PRIOR KNOWLEDGE EXPLORING (PKE) AND OBJECT-SENSITIVE REPRESENTATION LEARNING (ORL) ON TGIF-QA. ACCURACY (%) FOR ACTION, TRANS. AND FRAMEQA. MEAN ℓ_2 LOSS FOR COUNT, WHERE THE LOWER THE VALUE, THE BETTER

PKE	ORL	Count ↓	FrameQA ↑	Trans. ↑	Action ↑
✗	✗	4.08	58.3	79.2	72.3
✓	✗	3.99	59.1	82.0	74.1
✗	✓	3.76	60.8	81.2	74.3
✓	✓	3.67	61.8	82.3	74.6

1) Effect of the PKE and ORL: In Table I, we examine the impact of proposed PKE and ORL on MSVD-QA and MSRVTT-QA. We start from a baseline that only uses a vanilla attention-based model, where the result of the base model is shown in the first row of the table. Then, we incorporate PKE module and ORL module into the baseline, respectively. Finally, we combine both PKE and ORL into the baseline model to build our PKOL. From Table I, we can observe that PKE and ORL both bring an improvement with respect to the baseline model. Specifically, PKE and ORL increase the baseline from 37.3% to 39.3% and from 37.3% to 40.0% on the MSVD-QA, respectively. For the MSRVTT-QA, they also increase the baseline from 35.5% to 36.6% and from 35.5% to 36.3%, respectively. This proves that our proposed two modules are effective and beneficial. Moreover, the combination of the two modules further enhances the performance.

Additionally, we report the same ablation studies of each component on TGIF-QA. As shown in Table II, the performances are enhanced by individually adding PKE and ORL to the baseline on all sub-tasks. Specifically, after integrating PKE into the baseline, it drops the MSE loss by 0.09 for count task and increases the accuracy by 0.8%, 2.8% and 1.8% for FrameQA, Trans. and Action tasks, respectively. This illustrates the effectiveness of exploring prior knowledge. On the other hand, as we can see, integrating ORL into the baseline also drops the MSE loss by 0.32 for Count and increases the accuracy by 2.5%, 2.0% and 2.0% for FrameQA, Trans. and Action, respectively. This results imply that it is important to learn the relationship between objects for VideoQA task. Furthermore, the combination of two components can yield the highest performance.

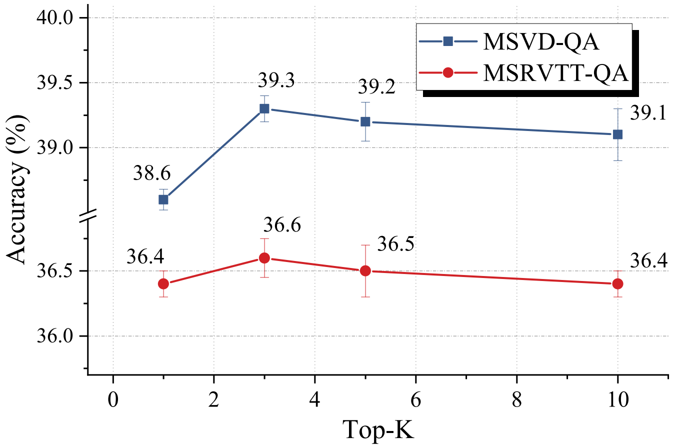


Fig. 3. Performance comparison of PKE with different numbers of retrieved sentences. Note that all models use only PKE without ORL for both two datasets.

For simplicity, to obtain the better configuration of each component, we only conduct the following experiments on MSVD-QA and MSRVTT-QA.

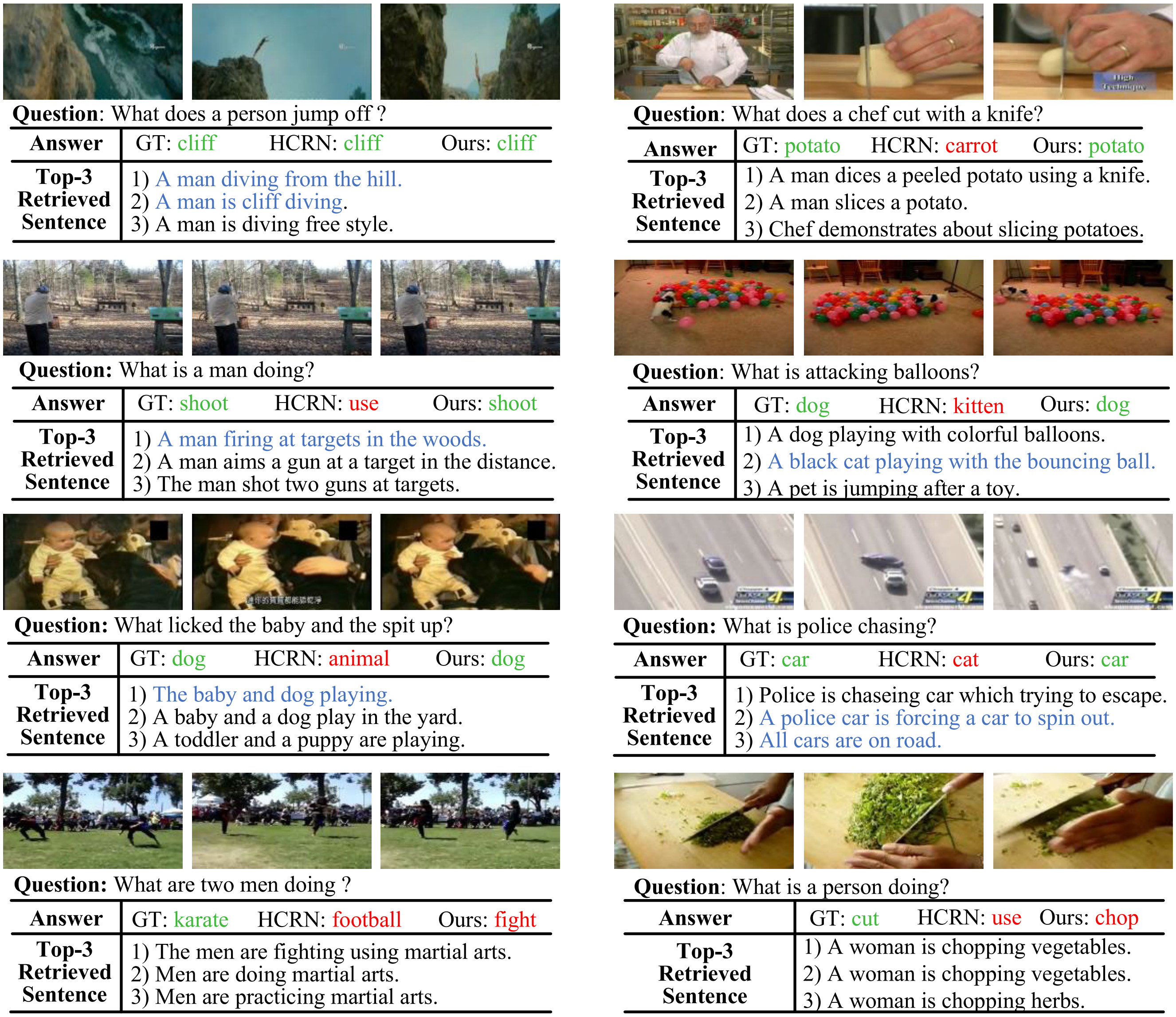
2) Effect of the Numbers of Retrieved Sentences in PKE: Empirically, the larger the value of k is, the more external information it will acquire and vice versa. However, an excess of external knowledge may contain irrelevant knowledge and noise, which may interfere the reasoning. A reasonable size of knowledge is important and should be guaranteed. Therefore, we conduct an ablation study to assess the effect of the number of retrieved sentences in PKE, shown in Fig. 3. Specifically, we take the baseline+PKE as our default model and set the k as 1, 3, 5, and 10. From Fig. 3, we can see that when $k = 3$, our method obtains the highest scores both on MSVD-QA and MSRVTT-QA. Further increasing the size of k may result in a slight decrease. Thus, we set $k = 3$ in the following experiments.

3) Effect of Query Types in PKE: In this paper, we attempt to integrate both video&question to obtain related prior knowledge. To fully explore their effects, we choose four variants: baseline (#1) without queries representing zero prior knowledge; with video query (#2); with question query (#3); and with video&question query (#4). All variants do not contain ORL. The experimental results are shown in Table III, which involves the results of two tasks: VideoQA and sentence retrieval. Compared with baseline #1, the other three variants (#2, #3 and #4) obtain significantly better answers for questions on both datasets. This demonstrates that prior knowledge is beneficial for VideoQA reasoning. Besides, we can also observe some interesting phenomena. Specifically, when using a single modality as a query (#2 and #3), #3 is better than #2 on the retrieval task (27.7% VS. 16.0% on R@1 in MSVD-QA and 9.5% VS. 8.3% on R@1 in MSRVTT-QA). This is because question queries share the same modality with the knowledge corpus, while video queries do not. Furthermore, #4 performs best on the VideoQA but goes second on the

TABLE III

PERFORMANCE COMPARISON OF PKE WITH DIFFERENT SETTINGS IN TERMS OF QUERY COMBINATIONS. VR AND QR RESPECTIVELY INDICATE THE VIDEO QUERY AND QUESTION QUERY FOR RETRIEVING SENTENCES FROM THE PRIOR KNOWLEDGE CORPUS. ALL VALUES ARE REPORTED AS ACCURACY (%). NOTE THAT ORL IS NOT INTEGRATED IN ALL MODELS

Model	VR	QR	MSVD-QA				MSRVTT-QA			
			Acc.	R@1	R@5	R@10	Acc.	R@1	R@5	R@10
1	✗	✗	37.3	-	-	-	35.5	-	-	-
2	✓	✗	38.6	16.0	36.6	49.2	36.4	8.3	23.9	34.1
3	✗	✓	39.1	27.7	52.0	62.7	36.4	9.5	26.6	30.2
4	✓	✓	39.3	19.4	42.0	52.0	36.6	9.2	25.7	36.7



Question: What does a person jump off ?

Answer	GT: cliff HCRN: cliff Ours: cliff
Top-3 Retrieved Sentence	1) A man diving from the hill. 2) A man is cliff diving. 3) A man is diving free style.

Question: What is a man doing?

Answer	GT: shoot HCRN: use Ours: shoot
Top-3 Retrieved Sentence	1) A man firing at targets in the woods. 2) A man aims a gun at a target in the distance. 3) The man shot two guns at targets.

Question: What licked the baby and the spit up?

Answer	GT: dog HCRN: animal Ours: dog
Top-3 Retrieved Sentence	1) The baby and dog playing. 2) A baby and a dog play in the yard. 3) A toddler and a puppy are playing.

Question: What are two men doing ?

Answer	GT: karate HCRN: football Ours: fight
Top-3 Retrieved Sentence	1) The men are fighting using martial arts. 2) Men are doing martial arts. 3) Men are practicing martial arts.

Question: What does a chef cut with a knife?

Answer	GT: potato HCRN: carrot Ours: potato
Top-3 Retrieved Sentence	1) A man dices a peeled potato using a knife. 2) A man slices a potato. 3) Chef demonstrates about slicing potatoes.

Question: What is attacking balloons?

Answer	GT: dog HCRN: kitten Ours: dog
Top-3 Retrieved Sentence	1) A dog playing with colorful balloons. 2) A black cat playing with the bouncing ball. 3) A pet is jumping after a toy.

Question: What is police chasing?

Answer	GT: car HCRN: cat Ours: car
Top-3 Retrieved Sentence	1) Police is chaseing car which trying to escape. 2) A police car is forcing a car to spin out. 3) All cars are on road.

Question: What is a person doing?

Answer	GT: cut HCRN: use Ours: chop
Top-3 Retrieved Sentence	1) A woman is chopping vegetables. 2) A woman is chopping vegetables. 3) A woman is chopping herbs.

Fig. 4. The qualitative VideQA results on MSVD-QA. Correct and wrong answers are marked in green and red, respectively. Besides, for each example, the Top-3 retrieved sentences are shown as an evidence to explain the reasoning process behind our model, where the ground-truth captions of the corresponding video are marked in blue.

retrieval task. This is because cross-modal retrieval always performs worse than same-modal retrieval due to the fact that the semantic gap between cross-modal data is larger than that between same-modal data. For the retrieval task, the model's performance measures the position of the ground-truth sentence in the retrieved lists. However, the purpose of PKE is

TABLE IV

PERFORMANCE COMPARISON OF ORL WITH DIFFERENT SETTINGS IN TERMS OF ACCURACY (%). NOTE THAT ORL_a AND ORL_m INDICATE THE OBJECT-SENSITIVE APPEARANCE LEARNING AND OBJECT-SENSITIVE MOTION LEARNING, RESPECTIVELY. NOTE THAT PKE IS NOT INTEGRATED IN ALL MODELS

ORL_a	ORL_m	MSVD-QA	MSRVTT-QA
✗	✗	37.3	35.5
✓	✗	39.8	36.1
✗	✓	38.7	35.9
✓	✓	**40.0**	**36.3**

TABLE V

PERFORMANCE OF PRIOR KNOWLEDGE WITH DIFFERENT CORPUS IN THE INFERENCE PHASE IN TERMS OF ACCURACY(%)

Text Corpus	MSVD-QA	MSRVTT-QA
-	37.3	35.5
train	40.9	36.7
val	40.8	36.5
test	**41.1**	**36.9**
train+val	39.7	36.7
train+val+test	41.0	36.9

TABLE VI

EXPERIMENT ON CROSS-DATASETS FOR VIDEOQA. THE MODEL IS TRAINED ON ITS OWN TRAINING SET AND TESTED WITH CORPUS FROM DIFFERENT TEST SETS

Text Corpus	MSVD-QA	Text Corpus	MSRVTT-QA
MSVD	41.1	MSRVTT	36.9
MSRVTT	38.6	MSVD	36.6
MSVD+MSRVTT	41.2	MSVD+MSRVTT	37.0

to retrieve video and question-related knowledge information to help the model predict accurate answers. In addition, one annotation problem with these datasets is that many sentences can be used to assist in describing the content and questions of other videos [45]. From the second example in the first row of Fig. 4, we can see that although the retrieved sentences are both not ground-truth sentences, they are relevant to the video and question, which can provide helpful information for model reasoning. To sum up, the results are reasonable, showing the effectiveness of our proposed PKE.

4) Effect of Motion and Appearance Based ORL: Table IV investigate the effectiveness of ORL on appearance feature and motion feature, namely Appearance-based Object-sensitive Learning (ORL_a) and Motion-based Object-sensitive Learning (ORL_m). All models in this section do not consider the PKE. From the table, we have the following observations. Firstly, compared with the baseline, baseline with ORL_a and ORL_m improves the performance on both datasets. Furthermore, the baseline with both ORL_a and ORL_m achieves the highest scores, reaching 40.0% on MSVD-QA and 36.3% on MSRVTT-QA. These results reveal the importance of object-sensitive representation learning, which captures high-level semantics features for object-related questions.

5) Effect of Different Prior Knowledge Corpus in the Inference Phase: In this paper, we take video descriptions as the prior knowledge corpus. The final model chooses the test video descriptions (#4) as the prior knowledge. However, test videos with description tags are hard to be obtained in practical scenarios, and thus we replace the original test descriptions corpus with other prior descriptions corpus from the dataset itself in the testing phase, including train (#2), val (#3), train+val (#5) and train+val+test (#6). Note that all models are trained using the corpus of the corresponding training set, and the best models are selected to perform best in the validation phase using the corpus of the corresponding val set. The experimental results are shown in Table V. Overall, all variants yield an improvement in a large margin compared with baseline #1. With other descriptions, the result only gains a slight drop in performance compared with test sets. Specifically, on MSVD-QA, the performance drops by 0.2% and 0.3% using train set' and val set' description as prior knowledge corpus, respectively. On MSRVTT-QA, when using train set' and val set' description as prior knowledge corpus, the performance drops by 0.2% and 0.4%, respectively. Besides, there exists an interesting phenomenon that the performance of utilizing more data as a corpus is comparable to that of only using the test as a corpus. The reason is that the model with train+val+test or train+val will be more inclined and overfitted to assign a higher rank for train (val) set' descriptions as they are used for training (validation). Overall, the above results prove that our model has a strong generalization ability, and it may only need some weak video-related descriptions as the prior knowledge corpus to obtain rich semantic information.

6) Generalization of the Model for Cross-Domain Prior Knowledge Corpus in the Inference Phase: In the practical scenario, it is hard to obtain the prior knowledge base that completely covers all the descriptions related to video content. To solve this problem, we conduct this experiment by training the model on the corresponding training set and measuring the performance on different cross-domain test sets, which puts forward higher requirements for the generalization ability of the model. As shown in Table VI, we can find that the configuration of cross-domain has a certain impact on the results (line2 VS. line 1), with a decrease of 1.5% on MSVD-QA and an increase of 0.8% on MSRVTT-QA. However, when we combine the MSVD and MSRVTT corpus as the final prior knowledge corpus, the performance can obtain a slight improvement, indicating that the model can retrieve and use more knowledge for inference from an additional corpus. Since the corpus combines descriptions from different test datasets, the descriptive information will contain more valuable semantic information. Different from the purpose of Table V, these results in Table VI show that our PKE has a strong generalization ability that does not require the same domain annotation description as corpus and is beneficial for the VideoQA task.

D. Performance Comparison

In this section, we compare our proposed $PKOL$ with the following main state-of-the-art (SOTA) methods.

- ST-VQA [11] is a model with an attention mechanism both in spatial and temporal dimensions. The goal of spatial attention is to figure out which regions in a frame to attend to for each word, whereas temporal attention is to figure out which frames in a video to attend to.
- Co-Mem [12] is a model with motion-appearance co-memory network. The co-memory attention mechanism generates attention by combining motion and appearance signals. A dynamic fact ensemble approach is meant to construct temporal facts dynamically during each cycle of fact encoding based on this attention.
- PSAC [8] is a model that substitutes the basic LSTM with positional self-attention blocks to capture data dependence. In addition, to anticipate more correct replies, a co-attention model is utilized to focus on textual and visual information.
- STA [46] is a model that adopts a structured two-stream attention to jointly reason across video and text's spatial and long-range temporal information.
- HME [47] is a model with three major components: 1) a novel heterogeneous memory that can learn global context information from appearance and motion features efficiently; 2) a revised question memory that aids in the comprehension of complicated question semantics and emphasizes questioned subjects; and 3) a novel multimodal fusion layer that conducts multi-step reasoning by self-updating attention to relevant visual and textual hints.
- L-GCN [21] is a model that combines an object's position characteristics into the graph construction to express richer visual content.
- QuesT [48] is a model that divides the semantic features generated from question into two separate parts: the spatial part and the temporal part, respectively guiding the process of constructing the contextual attention on spatial and temporal dimension.
- HCRN [4] is a model that designs a general-purpose reusable neural unit called Conditional Relation Network (CRN) that serves as a building block to construct more sophisticated structures for representation and reasoning over video.
- HGA [49] is a model with deep heterogeneous graph alignment network, which aligns and interacts information across inter- and intra-modality.
- DualVGR [5] is a model consisting of an explainable Query Punishment Module and a Video-based Multi-view Graph Attentio Network. The former can filter out unimportant visual features through multiple cycles of reasoning and the latter can record the relationships between appearance and motion features.
- BridgeToAns [27] is a model that infers right answer to questions regarding a correspondence video by leveraging sufficient network interactions of heterogeneous cross-modal graphs,.
- HOSTR [23] is a model that proposes an object-oriented reasoning approach. The model stands out with its authentic and explicit modeling of video objects leading to the effective and interpretable reasoning process.

TABLE VII

COMPARISON WITH THE STATE-OF-THE-ART ON THE MSVD-QA DATASET. ALL VALUES ARE REPORTED AS ACCURACY (%)

Model	MSVD-QA					
	ALL	What	Who	How	When	Where
ST-VQA	31.3	18.1	50.0	**83.8**	72.4	28.6
Co-Mem	31.7	19.6	48.7	81.6	**74.1**	31.7
HME	33.7	22.4	50.1	73.0	70.7	42.9
L-GCN	34.3	-	-	-	-	-
HGA	34.7	23.5	50.4	83.0	72.4	46.4
QuesT	36.1	24.5	52.9	79.1	72.4	**50.0**
HCRN	36.1	-	-	-	-	-
BridgeToAns	37.2	-	-	-	-	-
DualVGR	39.0	28.6	53.8	80.0	70.7	46.4
HOSTER	39.4	-	-	-	-	-
Baseline	37.3	27.4	53.0	73.7	68.9	46.4
PKE(ours)	39.3	29.2	54.1	76.0	72.4	50.0
ORL(ours)	40.0	29.1	**57.1**	80.5	72.4	39.3
PKOL(ours)	**41.1**	**30.4**	**56.8**	76.8	69.0	48.5

1) MSVD-QA: The comparison results on MSVD-QA are shown in Table VII, which displays the overall accuracy and the accuracy of each question type. From the Table, our proposed PKOL model exceeds all previous compared models with the best accuracy 41.1%. In particular, compared with the best baseline HOSTER, which designs an object-relational network similar to ours, PKOL obtains 0.7% improvement. In addition, we can also see that PKOL achieves the highest performance in the two question types ("What" and "Who"). Specifically, for the question of "What" and "who", PKOL increases the performance by approximately 1.8% and 3.0% compared with HOSTER, respectively.

Besides, to further verify the effectiveness of each proposed component, Table VII reports the performance of our baseline, PKE and ORL. From the table, we can see that compared to SOTA methods and the final model PKOL, each component achieves comparable results and even the best results in some categories (*e.g.,* ORL in "Who"). The possible reason is that since "What" asks more object-aware questions and the portion of such question (34.1%) is relatively more, ORL can perform better. However, the overall model PKOL performs best in "ALL", indicating that the final modal can trade-off between PKE and ORL. These results verify the effectiveness of prior knowledge exploring and object-sensitive representation learning.

2) MSRVTT-QA: The comparison results are summarized in Table VIII. We can find that our PKOL model maintains admirable performance, with the highest accuracy of 36.9%, which is 1.0% improvement over the second method HOSTER. Besides, PKOL also achieves relatively higher performance on most question types. Specifically, for the question of "Who" and "How", PKOL increases the performance by

TABLE VIII

COMPARISON WITH THE STATE-OF-THE-ART ON THE MSRVTT-QA DATASET. ALL VALUES ARE REPORTED AS ACCURACY (%)

Model	MSRVTT-QA					
	ALL	What	Who	How	When	Where
ST-VQA	30.9	24.5	41.2	78.0	76.5	34.9
Co-Mem	32.0	23.9	42.5	74.1	69.0	42.9
HME	33.0	26.5	43.6	82.4	76.0	28.6
L-GCN	-	-	-	-	-	-
HGA	35.5	29.2	45.7	83.5	75.2	34.0
QuesT	34.6	27.9	45.6	83.0	75.7	31.6
HCRN	35.6	-	-	-	-	-
BridgeToAns	36.9	-	-	-	-	-
DualVGR	35.5	29.4	45.6	79.8	76.7	**36.4**
HOSTER	35.9	-	-	-	-	-
Baseline	35.5	29.1	45.4	82.0	77.6	34.8
PKE(ours)	36.6	**30.8**	46.3	82.9	**79.0**	35.2
ORL(ours)	36.3	29.8	46.1	82.5	77.7	33.2
PKOL(ours)	**36.9**	30.2	**48.0**	**84.3**	78.0	35.6

TABLE IX

COMPARISON WITH THE STATE-OF-ARTS ON THE TGIF-QA DATASET. ACCURACY (%) FOR ACTION, TRANS. AND FRAMEQA. MEAN ℓ_2 LOSS FOR COUNT, WHERE THE LOWER THE VALUE, THE BETTER

Model	Count ↓	FrameQA ↑	Trans. ↑	Action ↑
ST-VQA	4.32	49.5	69.4	62.9
Co-Mem	4.10	51.5	74.3	68.2
PSAC	4.27	55.7	76.9	70.4
STA	4.25	56.6	79.0	72.3
HME	4.02	53.8	77.8	73.9
L-GCN	3.95	56.3	81.1	74.3
QueST	4.19	59.7	81.0	75.9
HCRN	3.82	55.9	81.4	75.0
BridgeToAns	3.71	57.5	82.6	**75.9**
HOSTR	4.13	58.2	82.1	75.6
Baseline	4.08	58.3	79.2	72.3
PKE (ours)	3.99	59.1	82.0	74.1
ORL (ours)	3.76	60.8	81.2	74.3
PKOL (ours)	**3.67**	**61.8**	**82.8**	74.6

approximately 2.3% and 0.8% compared with HGA, respectively. Compared with DualVGR, PKOL also obtains quite gain, especially 0.8% and 1.3% for the question of "What" and "When". Similarly, Table VIII also presents the performance of each component. From the table, we can conclude that PKE and ORL can complement each other and the final model PKOL obtains the best results.

3) TGIF-QA: To demonstrate the robustness of our method, we further provide quantitative results on TGIF-QA dataset in Table IX. From the table, we can see that our model PKOL achieves the best performance for Count, FrameQA, and Transition. PKOL also achieves relatively higher performance for Action. Specifically, PKOL significantly outperforms other compared methods since it obtains the lowest MSE for Count task. For FrameQA task, compared with the best counterpart HOSTR, our method achieves a higher gain, particularly improved by 3.6%. Besides, PKOL improves the second model Bridge2Ans by 0.2% for Transition task. This further improves the advantages of our method, which explores prior knowledge and mines the relationship across spatial and temporal domains simultaneously.

E. Qualitative Results

To better understand the contribution of our method PKOL, we provide the qualitative results on MSVD-QA, shown in Fig. 4. Each example consists of three images of a video, a question, a correct answer, two predicted answers by HCRN and our model, and Top-3 retrieved sentences. Six examples in the first three rows show that our method predicts an accurate answer while HCRN obtains a wrong answer, except for the first example. Besides, we also report two failure cases in the bottom examples in the figure. The failure cases show that our predicted answers are semantically related to the correct answer than HCRN. Furthermore, from the figure, we can observe that our PKOL can provide some evidence from Top-3 retrieved sentences to explain the reasoning process behind our model. It further proves the effectiveness and interpretability of our proposed component PKE.

V. CONCLUSION

In this work, we present PKOL, a novel prior knowledge exploring and object-sensitive learning for VideoQA, which explores the prior knowledge to the effect of reasoning and generates the object-sensitive representation based on the vanilla attention model. On the one hand, our PKOL incorporates prior knowledge into the linguistic question to enrich its representation in a way that mimics the human cognitive process. On the other hand, our PKOL interacts object feature with appearance and motion features across space-time to explore the structured visual information instead of the raw video. Extensive experiments on three benchmarks show the effectiveness of our proposed PKOL.

REFERENCES

[1] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *Proc. NeurIPS*, 2018, pp. 1571–1581.

[2] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. CVPR*, Jun. 2016, pp. 21–29.

[3] Z. Zhao *et al.*, "Long-form video question answering via dynamic hierarchical reinforced networks," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5939–5952, Jun. 2019.

[4] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proc. CVPR*, Jun. 2020, pp. 9972–9981.

[5] J. Wang, B. Bao, and C. Xu, "DualVGR: A dual-visual graph reasoning unit for video question answering," *IEEE Trans. Multimedia*, vol. 24, pp. 3369–3380, 2021.

[6] H. Xue, W. Chu, Z. Zhao, and D. Cai, "A better way to attend: Attention with trees for video question answering," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5563–5574, Nov. 2018.

[7] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. CVPR*, Jun. 2018, pp. 6087–6096.

[8] X. Li *et al.*, "Beyond RNNs: Positional self-attention with co-attention for video question answering," in *Proc. AAAI*, 2019, vol. 33, no. 1, pp. 8658–8665.

[9] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang, and H. T. Shen, "Hierarchical representation network with auxiliary tasks for video captioning and video question answering," *IEEE Trans. Image Process.*, vol. 31, pp. 202–215, 2022.

[10] Z. Zhang *et al.*, "Open-book video captioning with retrieve-copy-generate network," in *Proc. CVPR*, Jun. 2021, pp. 9837–9846.

[11] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *Proc. CVPR*, Jul. 2017, pp. 2758–2766.

[12] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proc. CVPR*, 2018, pp. 6576–6585.

[13] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *Proc. ICCV*, 2017, pp. 1811–1820.

[14] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, compositional video question answering," in *Proc. EMNLP*, 2018, pp. 1369–1379.

[15] X. Li *et al.*, "Learnable aggregating net with diversity learning for video question answering," in *Proc. ACM MM*, 2019, pp. 1166–1174.

[16] A. Seo, G.-C. Kang, J. Park, and B.-T. Zhang, "Attend what you need: Motion-appearance synergistic networks for video question answering," in *Proc. ACL*, 2021, pp. 6167–6177.

[17] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for visual question answering," in *Proc. CVPR*, 2018, pp. 6135–6143.

[18] 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 *Proc. IJCAI*, 2018, p. 27.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, vol. 28, 2015, pp. 91–99.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Jun. 2016, pp. 779–788.

[21] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 11021–11028.

[22] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "BERT representations for video question answering," in *Proc. WACV*, Mar. 2020, pp. 1556–1565.

[23] L. H. Dang, T. M. Le, V. Le, and T. Tran, "Hierarchical object-oriented spatio-temporal reasoning for video question answering," in *Proc. IJCAI*, Aug. 2021, pp. 636–642.

[24] K. Yi *et al.*, "CLEVRER: Collision events for video representation and reasoning," in *Proc. ICLR*, 2019, pp. 1–19.

[25] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. CVPR*, Jun. 2016, pp. 5288–5296.

[26] D. Xu *et al.*, "Video question answering via gradually refined attention over appearance and motion," in *Proc. ACM MM*, Oct. 2017, pp. 1645–1653.

[27] J. Park, J. Lee, and K. Sohn, "Bridge to answer: Structure-aware graph interaction network for video question answering," in *Proc. CVPR*, Jun. 2021, pp. 15526–15535.

[28] Z. Zhao, S. Xiao, Z. Song, C. Lu, J. Xiao, and Y. Zhuang, "Open-ended video question answering via multi-modal conditional adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 3859–3870, 2020.

[29] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA," in *Proc. CVPR*, Jun. 2021, pp. 14111–14121.

[30] J. Johnson *et al.*, "Inferring and executing programs for visual reasoning," in *Proc. ICCV*, Oct. 2017, pp. 2989–2998.

[31] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proc. CVPR*, 2018, pp. 4942–4950.

[32] Z. Chen *et al.*, "ComPhy: Compositional physical reasoning of objects and events from videos," 2022, *arXiv:2205.01089*.

[33] M. Ding, Z. Chen, T. Du, P. Luo, J. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," in *Proc. NeurIPS*, vol. 34, 2021, pp. 887–899.

[34] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proc. ACM MM*, Jun. 2018, pp. 19–27.

[35] X. Wang, L. Zhu, and Y. Yang, "T2VLAD: Global-local sequence alignment for text-video retrieval," in *Proc. CVPR*, 2021, pp. 5079–5088.

[36] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. CVPR*, Jun. 2019, pp. 1979–1988.

[37] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, "Learning modality interaction for temporal sentence localization and event captioning in videos," in *Proc. ECCV*. Springer, 2020, pp. 333–351.

[38] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, Jun. 2018, pp. 6077–6086.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[40] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *Proc. CVPR*, 2018, pp. 6546–6555.

[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. CVPR*, 2017, pp. 1492–1500.

[42] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[44] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019, pp. 8026–8037.

[45] H. Chen, J. Li, S. Frintrop, and X. Hu, "The MSR-video to text dataset with clean annotations," 2021, *arXiv:2102.06448*.

[46] L. Gao *et al.*, "Structured two-stream attention network for video question answering," in *Proc. AAAI*, 2019, vol. 33, no. 1, pp. 6391–6398.

[47] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proc. CVPR*, Jun. 2019, pp. 1999–2007.

[48] 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 *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 11101–11108.

[49] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 11109–11116.

Pengpeng Zeng received the B.Sc. degree in digital media technology from the Xi'an University of Technology in 2016 and the M.Sc. degree in computer technology from the University of Electronic Science and Technology of China in 2021, where he is currently pursuing the Ph.D. degree in computer science and technology.

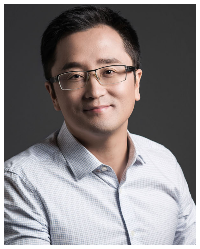
His current research interests include visual understanding, machine learning, and reinforcement learning.

Haonan Zhang received the B.Sc. degree in computer science and technology from Xidian University in 2020. He is currently pursuing the Ph.D. degree in computer science and technology with the University of Electronic Science and Technology of China.

His research interests include computer vision and natural language processing.

Lianli Gao (Member, IEEE) received the Ph.D. degree in information technology from The University of Queensland (UQ), Brisbane, QLD, Australia, in 2015. She is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. She is focusing on integrating natural language for visual content understanding. She was the Winner of the IEEE TRANSACTIONS ON MULTIMEDIA 2020 Prize Paper Award, the Best Student Paper Award in the Australian Database Conference, Australia, in 2017, the IEEE TCMC Rising Star Award in 2020, and the ALIBABA Academic Young Fellow.

Jingkuan Song (Senior Member, IEEE) is currently a Professor with the University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests include large-scale multimedia retrieval, image/video segmentation and image/video understanding using hashing, graph learning, and deep learning techniques. He was an AC Member/a SPC Member/a PC Member of IEEE Conference on Computer Vision and Pattern Recognition (2018–2021). He was the Winner of the Best Paper Award in International Conference on Pattern Recognition, Mexico, in 2016, the Best Student Paper Award in Australian Database Conference, Australia, in 2017, and the Best Paper Honorable Mention Award, Japan, in 2017.



Heng Tao Shen (Fellow, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively.

He is currently a Professor with the National Thousand Talents Plan and the Dean of the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, where he is also the Director of the Center for Future Media. He is also an Honorary Professor with The University of Queensland, Brisbane, QLD, Australia. His research interests include multimedia search, computer vision, artificial intelligence, and big data management. He is an ACM Distinguished Member and an Optical Society of America (OSA) Fellow.