Dynamic Spatio-Temporal Graph Reasoning for VideoQA with Self-Supervised Event Recognition

Jie Nie*, Xin Wang*, *Member, IEEE*, Runze Hou, Guohao Li, Hong Chen, and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Video question answering (VideoQA) requires the ability of comprehensively understanding visual contents in videos. Existing VideoQA models mainly focus on scenarios involving a single event with simple object interactions and leave event-centric scenarios involving multiple events with dynamically complex object interactions largely unexplored. These conventional VideoQA models are usually based on features extracted from the global visual signals, making it difficult to capture the object-level and event-level semantics. Although there exists a recent work utilizing a static spatio-temporal graph to explicitly model object interactions in videos, it ignores the dynamic impact of questions for graph construction and fails to exploit the implicit event-level semantic clues in questions. To overcome these limitations, we propose a Self-supervised Dynamic Graph Reasoning (SDGraphR) model for video question answering (VideoQA). Our SDGraphR model learns a question-guided spatio-temporal graph that dynamically encodes intra-frame spatial correlations and inter-frame correspondences between objects in the videos. Furthermore, the proposed SDGraphR model discovers event-level cues from questions to conduct self-supervised learning with an auxiliary event recognition task, which in turn helps to improve its VideoQA performances without using any extra annotations. We carry out extensive experiments to validate the substantial improvements of our proposed SDGraphR model over existing baselines.**

***Index Terms*—Vision and Language Model, Video Question Answering, Video Understanding, Spatio-temporal Graph**

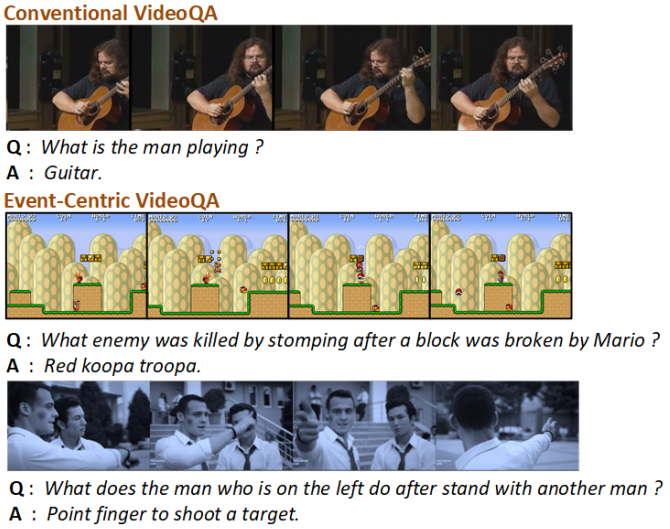


Fig. 1. An example demonstrating the differences between conventional VideoQA and event-centric VideoQA. Conventional VideoQA mainly focuses on simple scenarios where only one or very few events are involved, without the necessity of complex reasoning for answering the question. In contrast, event-centric VideoQA tends to contain multiple events with dynamic and complex interactions among objects in the videos.

I. Introduction

With the rapid development of applications in computer vision, the research community has encouraged more efforts towards a deeper understanding of video content. In recent years, machine learning algorithms have achieved impressive progress on visual perception tasks, including image classification, object detection and segmentation, etc. Nevertheless, it is still very challenging for machines to achieve human comparable performances for high-level visual understanding tasks, e.g., correctly answering questions based on video contents (VideoQA).

Conventional VideoQA works mainly focus on relatively simple scenarios where very few events are presented in the videos and correctly answering questions rarely requires complex reasoning. In this work, we investigate the more difficult problem of event-centric VideoQA, i.e., visual reasoning in event-centric videos containing multiple events with dynamically complex object interactions. Fig 1 presents an example demonstrating the differences between conventional and event-centric VideoQA scenarios.

On the one hand, substantial efforts have been devoted to the conventional VideoQA tasks that largely rely on low-level visual perceptions rather than high-level object-related reasoning. Existing methods for conventional VideoQA combine features of visual signals and textual questions together, using various mechanisms like attention, memory and graph to model the motions and appearances in videos [59], [62], [8], [6], [55]. On the other hand, reasoning in event-centric videos requires the abilities of well understanding the target task (question), perceiving principle parts from visual signals, modeling the complex spatio-temporal relationships in dynamic visual scenes across multiple events, and carrying out progressive operations for multi-step reasoning. Several visual reasoning models for visual question answering mainly focus on static images with recurrent approaches [40], [17] or modular approaches [1], [35], [31]. Other works are also proposed to extend image-based visual reasoning methods to the video domain through temporal reasoning [48], [63], or representing videos with spatio-temporal graphs to explicitly

Jie Nie is with Ocean University of China, China. Xin Wang, Runze Hou, Guohao Li, Hong Chen and Wenwu Zhu are with Department of Computer Science and Technology, BNRIST, Tsinghua University, China.

*Jie Nie and Xin Wang contribute equally. Corresponding authors: Xin Wang and Wenwu Zhu. E-mail: niejie@ouc.edu.cn, {xin_wang, wwzhu}@tsinghua.edu.cn, {hrz21,ligh16,h-chen20}@mails.tsinghua.edu.cn.

This work is supported by the National Key Research and Development Program of China No. 2023YFF1205001 and National Natural Science Foundation of China No. 62250008, 62222209, 62102222.

model the relationships between objects for VideoQA [15].

However, existing literature suffers from the following limitations when conducting reasoning in event-centric videos: i) Conventional VideoQA approaches are usually based on global visual features that lack sufficient semantic knowledge, failing to recognize different object instances and complex interactions between objects in space and time across multiple events. ii) The recent graph-based approach largely ignores the impact of questions when constructing the object relation graphs in videos. iii) In event-centric VideoQA, questions usually contain useful cues about what events have happened, which provides useful semantic information seldom exploited by the existing methods.

To overcome these limitations, we propose a Self-supervised Dynamic Graph Reasoning (SDGraphR) model for event-centric VideoQA. Based on the input question, our SDGraphR model first learns a question-guided spatio-temporal graph for the corresponding video, which dynamically encodes intra-frame spatial correlations and inter-frame correspondences between objects. The spatio-temporal graph is then exploited through a graph convolutional network (object-level modeling) and aggregated to capture the long-term relationships in the course of time (frame-level modeling). Further, the proposed SDGraphR model is capable of discovering event-related cues from questions by conducting self-supervised learning with an auxiliary event recognition task, which satisfies the event-level semantic constraints for providing a more accurate answer. This self-supervised procedure in our proposed SDGraphR model is designed to enable the consistent improvement of model performance for the VideoQA task, without using any extra annotations. We demonstrate the effectiveness of our approach in MarioQA dataset, a gameplay VideoQA dataset containing abundant events, and show that the proposed SDGraphR model with self-supervised training strategies can achieve significant improvement over existing state-of-the-art baselines. Then we carry out experiments on real-world datasets (MSVD-QA, MSRVTT-QA) and achieved competitive performance with existing baselines, which verify the generalization ability of our method on conventional VideoQA case.

To summarize, our work makes the following contributions:

- We propose a Self-supervised Dynamic Graph Reasoning (SDGraphR) model for video question answering in event-centric videos, which explicitly models the complex spatio-temporal intra-frame and inter-frame object interactions across multiple events by taking the impacts of questions on the dynamic graphs into consideration.
- We propose to exploit the event-related cues discovered from questions to conduct self-supervised learning with an auxiliary event recognition task, which can help to improve the model performance for visual reasoning without any extra annotations.
- We conduct extensive experiments and demonstrate the effectiveness of our proposed SDGraphR model against several state-of-the-art approaches.

The remainder of this paper is organized as follows. We review related works in Section II and present our proposed SDGraphR model in Section III. Section IV describes details about empirical evaluations over VideoQA datasets in terms of various metrics, followed by our detailed implementations introduced in Section V. Last but not least, we conclude the whole paper and point out research directions deserving further investigations in Section VI.

II. Related Work

In this section, we review existing works on video representation learning with graphs, video question answering and self-supervised learning.

Video Representation Learning. In the past decades, researchers have developed a series of approaches for modeling the appearances and dynamics of videos in the settings of video classification (video action recognition). Before the deep learning era, the hand-designed features were widely used, such as the SIFT-3D [45], HOG-3D [28] and Dense Trajectories [53], [54]. With the rise of convolutional neural networks, researchers have switched to focus on learning deep features from videos. One type of architecture is to first utilize spatial 2D ConvNets and then model temporal information [47], [64], [5]. Another type of architecture is the 3D ConvNets, such as C3D [20], [50], I3D [3], P3D [42] and R(2+1)D [51], which are shown to be more powerful over 2D ConvNets for videos.

However, the aforementioned approaches extract features from the whole visual scenes, without considering any explicit semantic meanings. Thus, they can hardly recognize different object instances and model object-object relationships in space and time, especially when the videos are event-centric and involve complex object interactions.

Representing Video as Graphs. In order to model semantically meaningful spatio-temporal interactions in videos, several works [56], [41], [13], [65], [37], [36] explore to represent videos as object-level spatio-temporal graphs in recent years.

The Object Relation Network (ORN) [2] conducts relational reasoning between pairwise semantic object instances through space and time. Wang et al. [56] propose to represent videos as space-time region graphs followed by graph convolutions for inference. Qi et al. [41] infer a graph from visual scenes within a message-passing framework for human-object interaction recognition. Later, Zhang et al. [65] employ a tracking module to aggregate long-term motion patterns and reason about interactions between actors and objects. Mavroudi et al. [37] propose to use a hybrid spatio-temporal visual graph and a symbolic attributed graph to capture rich visual and semantics.

These approaches are mainly proposed for human action recognition where very few actions are presented in these videos, while our work aims at answering complex questions in event-centric videos. In contrast to previous methods, we take the input textual question into consideration and develop a question-guided spatio-temporal graph reasoning method in the scenario of VideoQA.

Answering Visual Questions in Videos. There have been many works for multi-step reasoning on static visual scenes (e.g., CLEVR [23] and GQA [18] datasets). In general, researchers adopt either recurrent approaches or modular approaches. For recurrent approaches, each reasoning step is usually implemented with a general-purpose reasoning block [40],

[17], [16], while the modular approaches decompose the reasoning procedure into specialized modules [1], [24], [14], [35], [31]. Despite the impressive progresses that have been made on static images, visual reasoning on videos yet remains rarely explored, especially when the videos involve complicated events. Recently, Song et al. [48] construct a refined GRU (Gated Recurrent Unit) with temporal attention for VideoQA. Yi et al. [63] combine a neural video parser with a symbolic program executor to obtain an answer. Le et al. [30] build a hierarchical structure with Conditional Relation Network (CRN) to process input objects and conditions. The replication and stacking of reusable networks is beneficial for obtaining diverse modalities and contextual information. Recently, Gao et al. [7] proposed MIST, a multi-modal spatio-temporal transformer to answer the questions in long videos.

The above works are either feature-based methods that lack sufficient semantic knowledge, or rely on extra program annotations that are not usually available. Recently, Huang et al. [15] propose a location-aware graph convolutional network (L-GCN) by incorporating the location information of an object into the graph. There are also several methods [46], [21], [10] performing spatial-temporal reasoning by explicitly modeling the positional relationship between objects. Grunde-McLaughlin et al. [9] utilize dynamic spatio-temporal scene graphs to generate questions for QA dataset. There are also several methods [57], [61], [52], that introduce hyper-graph, a hierarchical structure that encodes scene sub-graphs with connections between objects, relations and actions for video frames and hyper-edges for connected sub-graphs, to solve the VideoQA problem. However, they require additional hyper-graph annotations, which is conducive to the solving the VideoQA problem, but limits its application in real-world systems. However, the existing VideoQA methods usually neglect the guidance information from the questions when operating on the spatio-temporal graphs. In comparison with the existing methods, our model explores a question-guided graph-based approach for spatio-temporal reasoning in videos. Meanwhile, we propose to exploit the question cues to gather self-supervision for an auxiliary event recognition task instead of modeling events as symbols for programs [57] and language representation in the quesiont [9] , which consistently improves model performances without using any extra annotations.

Self-Supervised Learning with Auxiliary Tasks. The idea of exploiting the question cues to form a self-supervised auxiliary event recognition task in this work is closely related to Self-Supervised Learning (SSL). As a subset of unsupervised learning methods, self-supervised learning leverages input data itself as supervisions [22], [33] without relying on expensive human annotations, where the model learns data representations through pre-designed auxiliary tasks and automatically generated pseudo-labels. Various auxiliary tasks have been proposed for self-supervised learning in computer vision (e.g., colorization [29], inpainting [39], etc.), neural language processing (e.g., next word prediction [43], masked language model [4], etc.) and visual-language understanding [49], [34]. For visual question answering, Zhu et al. [66] propose a self-supervised task which utilizes the generated balanced data to overcome the language prior problem recently. In our work, based on the observation that questions always contain information about what events have happened, we propose to exploit the implicit cues in questions through a self-supervised event recognition task.

III. Self-Supervised Video Question Answering with Dynamic Graphs

In this section, we elaborate on the details of our proposed model. Fig 2 presents an overview of our model based on a visualized example. We first describe the learning process of constructing the object-level spatio-temporal graphs for videos based on an inquired question in Section III-A, and then present the details of object-level and frame-level graph modeling in Section III-B, followed by discussions about the novel self-supervised auxiliary task and joint training strategies of the proposed SDGraphR model in Section III-C.

A. Dynamic Object-Level Graphs for Videos

For each video, we propose to build a question-guided spatio-temporal graph to explicitly model the object-object dependencies in videos. In the spatio-temporal graph, a graph node corresponds to an object snapshot at a specific time, which contains various information about the object, such as identity, position, visual appearance, etc. By linking the objects within the same frame and objects across different frames, the spatio-temporal graph is able to serve as an explicit video representation that reflects the underlying spatial and temporal dependencies in dynamic visual scenes. We adopt a two-step procedure to generate the spatio-temporal graphs for videos and in an alternative view, the graphs can be regarded as one type of prior knowledge.

Objects as Graph Nodes. Given a clip of the target video, we first sample a fixed number of frames from the clip. The sampled frames are then fed into a pre-trained Faster-RCNN [44] model, which predicts the bounding boxes and categories of objects that appear in the frames. In detail, the Faster-RCNN model is based on a ResNet-18 [12] (for synthetic dataset) and ResNet-101 (for real world dataset) backbone with Feature Pyramid Network (FPN) [32]. Based on the predicted bounding boxes, for synthetic dataset, we feed the frames again into the backbone and apply the ROIAlign [11] to extract visual features for each bounding box. And for real-world data, we feed the frames to another pretrained ResNeXt-101 [58] backbone to enhance the feature generalization in real-world data and then we directly crop corresponding feature map from overall feature map to get the object feature.

Formally, we sample a fixed number ($T = 12$) of frames for each video v where for the t_{th} frame there are $\boldsymbol{n}_t$ objects $\{o_t^k\}_{k=1,\cdots,\boldsymbol{n}_t}$ in the frame. Thus, there are in total $N = \sum_{t=1}^{T} \boldsymbol{n}_t$ objects, constituting the N nodes in the corresponding spatio-temporal graph of the video. For the k_{th} object in the t_{th} frame, i.e., noted as o_t^k, we denote its bounding box as $\boldsymbol{b}_t^k = [x_t^k, y_t^k, w_t^k, h_t^k]$ which represents the 2D coordinate of its center position, width and height of the

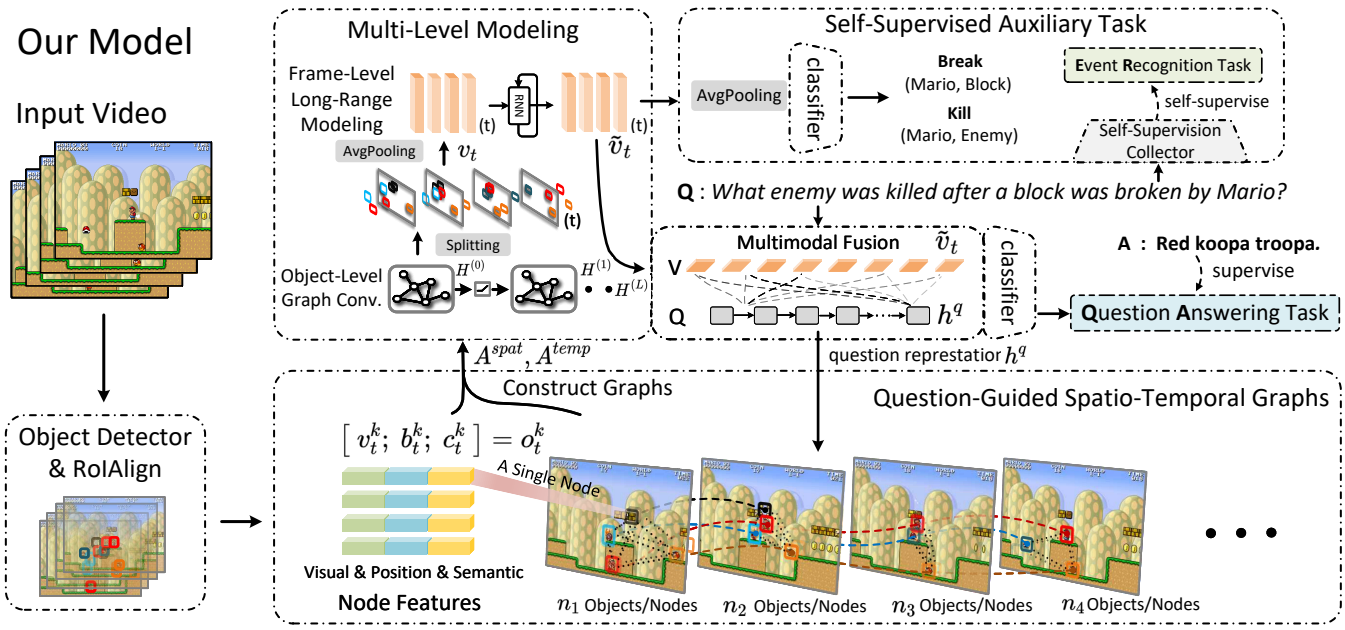


Fig. 2. **Overview of our proposed SDGraphR model.** Given a pair of video and question, our model first learns a question-guided spatio-temporal graph from the extracted objects for the video, encapsulating the object-object interactions in space and time. Subsequently, the graph and node features are fed into a graph convolutional network to exploit the neighborhood information, and then aggregated into context-aware frame-level features for modeling long-range relationships over time. We apply a general multi-modal fusion module to produce an adequate answer according to above representations for the given question. In addition to the primary Question Answering task, we propose a self-supervised auxiliary Event Recognition task, which can consistently boost the model performance without using any extra annotations.

bounding box. Based on $\boldsymbol{b}_t^k$, we obtain its visual feature, i.e., a feature map obtained from ROIAlign or direct crop and then squeeze it into a feature vector $\boldsymbol{v}_t^k \in \mathcal{R}^D$ by average pooling, where D is the channel number over feature map. Meanwhile, the Faster-RCNN model also generates a distribution $\boldsymbol{c}_t^k \in \mathcal{R}^C$ over all C possible object classes. Besides, we also take the global scene as an object in each frame to simultaneously consider the scene as a whole. Finally, we concatenate the visual feature, object bounding box and object category together to form the representation of a node, i.e., $\boldsymbol{o}_t^k = [\boldsymbol{v}_t^k; \boldsymbol{b}_t^k; \boldsymbol{c}_t^k] \in \mathcal{R}^{d_o}$, in a spatio-temporal graph.

Learning Question-Guided Graph Edges. For each video, we link the objects with weighted edges based on the given question q, where larger weights indicate higher possibilities of interactions between objects, e.g., objects of interest in the question.

Specifically, we build a spatial graph $\mathbf{A}^{spat}$ and a temporal graph $\mathbf{A}^{temp}$ for each video. For simplicity, we re-arrange the objects $\{\boldsymbol{o}_t^k\}_{t=1,\cdots,T;k=1,\cdots,\boldsymbol{n}_t}$ into $\{\boldsymbol{o}_i\}_{i=1,\ldots,N}$, and define a mapping function $\tau(\cdot)$, where $\tau(\boldsymbol{o}_i) = t$ means that $\boldsymbol{o}_i$ is an object in the t_{th} frame. Formally, the graph adjacency matrices $\mathbf{A}^{spat} \in \mathcal{R}^{N\times N}$ and $\mathbf{A}^{temp} \in \mathcal{R}^{N\times N}$ are calculated as follows, for $i = 1, \cdots, N$ and $j = 1, \cdots, N$:

$$\mathbf{A}_{i,j}^{spat} = \begin{cases} F^{(s)}(\boldsymbol{o}_i, \boldsymbol{o}_j|q) & \text{if } \tau(\boldsymbol{o}_i)=\tau(\boldsymbol{o}_j) \text{ and } i\neq j \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

$$\mathbf{A}_{i,j}^{temp} = \begin{cases} F^{(t)}(\boldsymbol{o}_i, \boldsymbol{o}_j|q) & \text{if } 0 < |\tau(\boldsymbol{o}_i) - \tau(\boldsymbol{o}_j)| \leq w \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $F^{(s)}$ and $F^{(t)}$ compute the interaction weights for intra-frame and inter-frame object pairs respectively, w is an integer indicating the window size for adjacent frames. The graphs of our method are completely dynamic. For the static graph [15], each frame in the video requires constructing a graph with a fixed number of nodes. Our dynamic method focuses on the whole video, and each graph created contains all the objects in the video. The number of graphs in the static method is equal to the number of frames sampled. However, the number of graphs in our proposed model is fixed at two, one for spatial modeling and the other for temporal modeling.

There are two ways to design the object interaction functions: i) heuristic and ii) learnable.

Heuristic Interaction Function. When building the graph edges heuristically, we follow several assumptions on those objects that tend to have more correlations:

1) **The objects tend to be closer in space** — due to that events usually happen to surround spatially nearby objects.
2) **The objects tend to be more similar in adjacent frames** — due to that the changes of object states over time can be captured through analyzing adjacent frames.

For intra-frame object pairs, the correlation is in inverse proportion to the Euclid Distance of the object pair within the frame, i.e.,

$$F^{(s)}(\boldsymbol{o}_i, \boldsymbol{o}_j|q) = F^{(s)}(\boldsymbol{o}_i, \boldsymbol{o}_j) = \frac{1}{\|(x_i, y_i) - (x_j, y_j)\|}. \quad (3)$$

For inter-frame objects, we compute the cosine similarity in terms of the categorical distributions between each possible object pair in adjacent frames and link the object pair whose similarity score is above zero, i.e.,

$$F^{(t)}(\boldsymbol{o}_i, \boldsymbol{o}_j|q) = F^{(t)}(\boldsymbol{o}_i, \boldsymbol{o}_j) = \max(0, \frac{\boldsymbol{c}_i \cdot \boldsymbol{c}_j}{\|\boldsymbol{c}_i\| \cdot \|\boldsymbol{c}_j\|}). \quad (4)$$

Learnable Interaction Function. For a question q with M words, the question words are first mapped into M 300-dimension randomly initialized vectors. We use a one-layer

Gated Recurrent Unit (GRU) with a hidden dimension of $d_q = 512$ as the question encoder, then feed the word vectors into GRU to get a sequence of hidden vectors. We take the final GRU hidden vector $\boldsymbol{h}^q \in \mathcal{R}^{d_q}$ as the question representation. We seek to model the pairwise interactions between two objects, e.g., $\boldsymbol{o}_i$ and $\boldsymbol{o}_j$, through the inner product of their joint representations with the given question representation $\boldsymbol{h}^q$ as follows:

$$F(\boldsymbol{o}_i, \boldsymbol{o}_j|q) = \mathrm{ReLU}(\mathbf{W}[\boldsymbol{o}_i; \boldsymbol{h}^q]) \cdot \mathrm{ReLU}(\mathbf{W}[\boldsymbol{o}_j; \boldsymbol{h}^q]), \quad (5)$$

where $\mathbf{W} \in \mathcal{R}^{(d_o+d_q)\times d_e}$ is a learnable parameter matrix that embeds the concatenated object and question features into a $d_e = 256$ dimensional joint embedding space.

In this way, we build question-guided spatio-temporal graphs that can encapsulate both spatial intra-frame and temporal inter-frame dependencies in videos simultaneously.

We would like to point out that both configurations of using heuristic interaction function and learnable interaction function to construct the spatio-temporal graphs in videos are compared in our experiments in Section IV-C.

B. Exploiting Spatio-temporal Graphs with Dynamic Multi-Level Modeling

To capture both the object-level spatio-temporal interactions and frame-level dynamic relationships, we propose to exploit the spatio-temporal graphs through object-level and frame-level modeling in a dynamic manner. For the object-level modeling, we develop a specific type of graph convolutional network to exploit the spatio-temporal interactions between objects. Afterwards, we aggregate the objects in each frame and directly model the long-term patterns in the course of time to capture the frame-level dynamic relationships. Besides, to fully utilize the question as guidance, we employ a multimodal fusion module to integrate the question and visual features before obtaining the final answer for the given question.

Spatio-Temporal Graph Convolutions. We employ a graph convolutional network inspired by [27]. Given the adjacency matrix $\mathbf{A}$, we first normalize the graph to introduce self-loop connections and balance the neighboring weights according to the node degrees:

$$\tilde{\mathbf{D}}_{i,i} = \sum_j (\mathbf{A}_{i,j} + \mathbf{I}_{i,j}), \ \tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\tilde{\mathbf{D}}^{-\frac{1}{2}}, \quad (6)$$

where $\mathbf{I}$ is the identity matrix representing self-connections, $\tilde{\mathbf{D}}$ is the diagonal degree matrix of $\mathbf{A} + \mathbf{I}$.

The GCN is a layer-wise network that takes the initial node features as input and updates the features in each layer. In our work, due to the intrinsic differences between spatial and temporal relationships, we separately conduct convolutional operations on spatial and temporal graphs in different layers and samples. Besides, we use residual connections between layers to improve the representation capacities of our model.

Formally, we stack the object features $\{o_i\}_{i=1,\ldots,N}$ into $\mathbf{O} \in \mathcal{R}^{N\times d_o}$ and perform the convolutions on spatio-temporal graphs as follows:

$$\mathbf{H}^{(0)} = \mathbf{O}, \quad (7)$$

$$\mathbf{H}^{(l+1)} = \mathrm{ReLU}(\tilde{\mathbf{A}}^{(l)}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \ l = 0, \cdots, L-1, \quad (8)$$

where $\mathbf{H}^{(l)} \in \mathcal{R}^{N\times d_h}$ for $l = 1, \cdots, L$ and $d_h = 512$, $\mathbf{W}^{(l)}$ denote the network parameters in the l_{th} layer. We use $L = 6$ convolutional layers, where adjacency matrix $\tilde{A}^{(l)}$ is one of the $\{A^{spat}, A^{temp}\}$. We discuss the effects of different choices in Section IV-C.

Long-term Patterns in Videos. Upon being processed through the graph convolutional networks, the object representations have aggregated the neighbor information from nearby objects in the spatio-temporal graphs. Furthermore, to capture the long-time correlations among objects in videos, we directly model their long-term dynamic patterns across frames with a recurrent neural network (RNN). In detail, we first aggregate the intra-frame objects with average pooling to obtain frame-level features $\{\boldsymbol{v}_t\}_{t=1,\cdots,T}$, then apply a recurrent neural network (RNN) to capture the inter-frame relationships. Formally, we have:

$$\{\boldsymbol{v}_t\} = \mathrm{avgpool}_{obj}(\mathrm{split}(\mathbf{H}^{(L)}; \tau(\cdot))), \quad (9)$$

$$\{\tilde{\boldsymbol{v}}_t\} = \mathrm{RNN}(\{\boldsymbol{v}_t\}), \quad (10)$$

where the split operator splits $\mathbf{H}^{(L)}$ into objects belonging to different frames according to the mapping function $\tau(\cdot)$. As for the RNN structure, we use a one-layer GRU with a hidden size of 512 in our model.

Applying Fusion Modules. Based on the visual features $\{\tilde{\boldsymbol{v}}_t\}_{t=1,\cdots,T}$ obtained through the dynamic modeling for the spatio-temporal graph, we get the final answer for the given question through applying our multimodal fusion framework. We note that a wide range of multimodal fusion approaches can be applicable in our framework, as long as it accepts sequential visual and question features as inputs before generating resultant features as output:

$$\boldsymbol{r} = \mathrm{FusionNet}(\{\tilde{\boldsymbol{v}}_t\}_{t=1,\cdots,T}; \{\boldsymbol{h}^q_m\}_{m=1,\cdots,M}), \quad (11)$$

where $\tilde{\boldsymbol{v}}_t$ is the visual feature of t_{th} frame and $\boldsymbol{h}^q_m$ is the question feature of m_{th} word.

After concatenating the fused feature $\boldsymbol{r}$ and the question representation $\boldsymbol{h}^q$, we feed them into a multi-layer perceptron to compute the possibilities for all candidate answers $\mathcal{A}$. Then, after $\mathrm{sigmoid}$ activation, the output is $p(y_{a_i} = 1|q, v)$, where $y_{a_i} \in \{0, 1\}$ indicates whether the i_{th} answer in candidate answer set $\mathcal{A}$, i.e., a_i can be the correct answer to the target question. The final answer will be selected as the one with the highest probability, i.e., $\hat{y} = \mathrm{argmax}_i \ p(y_{a_i} = 1|q, v)$.

C. Event Recognition as Self-supervision

When reasoning in event-centric videos, whether a specific event happens or not can serve as informative signals. These signals can potentially benefit the model learning process in better understanding the event-centric videos. The existing VideoQA approaches usually neglect this type of information implicitly carried in questions. To tackle this issue, in addition to the primary *Question Answering* task, we propose to construct a self-supervised auxiliary *Event Recognition* task that aims to boost the performance of the primary task through applying event-level semantic constraints.

Based on the observation that questions always contain information about what events have happened, we exploit

the cues in questions to gather self-supervision for the event recognition task without using any extra manual annotations. For example, the question "*Which enemy was stomped by Mario before a shell hit a goomba ?*" implies that at least two events *"Mario stomped an enemy" (Mario, Stomp, Enemy)* and *"A shell hit a goomba" (Shell, Hit, Goomba)* happen in the corresponding video. We also show another example in MarioQA dataset in Fig 3, which contains both positive and negative events.

Self-Supervison from Questions. Given that the names in the MarioQA dataset tend to be relatively rare in daily life (e.g., Mario, goomba, koopa troopa), we manually define a collection of concepts for the given dataset, including agents in the scenes (e.g., *Mario, Goobma, Enemy*, etc.), items (e.g., *Mushroom, Coin, Block*, etc.) and actions (e.g., *Jump, Stomp, Kick, Eat*, etc.).

We design an automatic procedure to parse each question into a list of concepts and assemble them into several event descriptors, i.e., tuples such as *(subject, action)* or *(subject, verb, object)*, which are considered as the positive events in the corresponding video. Our design benefits in the advantage of deriving more positive events through concept hierarchies. For example, considering the positive event *(Mario, Stomp, Goomba)* and the fact that *Goomba* is one type of *Enemy*, we can infer that *(Mario, Stomp, Enemy)* is also a valid positive event. In total, there are 116 events that have been observed from the questions in the MarioQA dataset. As for the negative events involved in each question-video pair, we refer to the object detection results for the video and identify those events that involve non-existing objects as negative events. In this way, the event recognition task can be formulated as a multi-label classification task over all candidate events.

We resort to the frame-level features $\{\tilde{\boldsymbol{v}}_t\}$ in videos for event recognition. Formally, we have:

$$p^{(er)}(y_{e_k}|v) = \text{sigmoid}\left(\text{MLP}_{er}(\text{avgpool}_t(\{\tilde{\boldsymbol{v}}_t\}))\right). \quad (12)$$

Here $p^{(er)}(y_{e_k}|v)$ is the predicted possibility of y_{e_k}, where $y_{e_k} = 1$ indicates event e_k is positive and $y_{e_k} = 0$ means negative. MLP_{er} introduces extra parameters for this task.

We use the binary cross-entropy (BCE) loss to train our model in an end-to-end fashion for the question answering task and event recognition task simultaneously. Formally, the loss of each data sample for the question answering task is formulated as follows:

$$\mathcal{L}^{(qa)} = -\sum_{a_i \in \mathcal{A}} \left(y_{a_i} \text{log} p(y_{a_i}) + (1 - y_{a_i}) \text{log}(1 - p(y_{a_i}))\right), \quad (13)$$

where i is the index of answer. Similarly, we compute the loss $\mathcal{L}^{(er)}$ for the event recognition task.

$$\mathcal{L}^{(er)} = -\sum_{e_k \in \mathcal{E}} \left(y_{e_k} \text{log} p(y_{e_k}) + (1 - y_{e_k}) \text{log}(1 - p(y_{e_k}))\right), \quad (14)$$

where $\mathcal{E}$ is the set of possible events. Note that we only consider the positive and negative events and ignore the uncertain events when computing the loss.

Joint Training Strategies. We consider two joint training strategies: *multi-task* and *pretrain-finetune*. For the *multi-task* training strategy, we apply a one-phase training procedure with a total loss $\mathcal{L} = \mathcal{L}^{(qa)} + \eta\mathcal{L}^{(er)}$, where η is a hyper-parameter to balance the multi-task losses. As for the *pretrain-finetune* training strategy, we apply two training phases sequentially:

1) Minimize the $\mathcal{L}^{(er)}$ loss for the event recognition task;
2) Transfer the pretrained model parameters and finetune on the primary question answering task, i.e., minimize the $\mathcal{L}^{(qa)}$ loss.

We close this section by pointing out that both joint training strategies can significantly improve the VideoQA performance, which is validated in Section IV-B.

IV. Empirical Experiments

We conduct extensive experiments on the MarioQA dataset [38], a synthetic VideoQA dataset collected from Super Mario Bros gameplay videos. There are 13 hours of gameplay videos totally in this dataset. These videos are divided into clips, and each clip contains 11.3 events on average.

Because of its event-intensive characteristic and excessive requirements for event-centric spatio-temporal modeling, the MarioQA dataset serves as an excellent benchmark for evaluating the ability of high-level reasoning in videos. Furthermore, the semantics in gameplay videos are clear, unambiguous, and relatively easy to learn compared with real-world videos, which makes it a good test-bed for this challenging research field. Thus, we think MarioQA is the most suitable dataset in our target multi-agent multi-event VideoQA scenario so far.

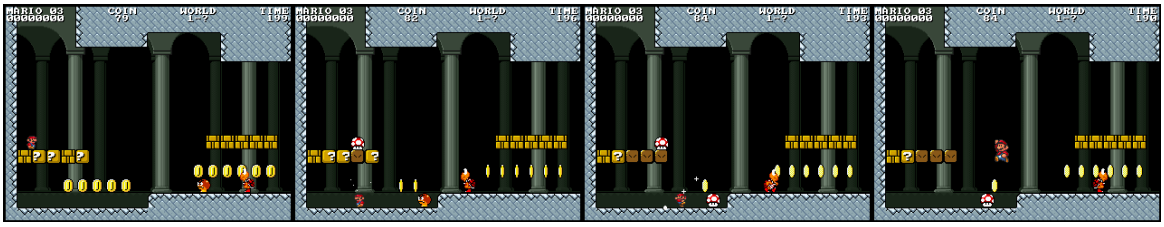
Besides, to prove the generalization of our method, we also conduct experiments on real-world VideoQA datasets. We choose MSVD-QA [59] and MSRVTT-QA [59], two of the most commonly used datasets for evaluation. These two datasets are generated from two video description datasets, MSVD and MSR-VTT [60], respectively with NLP tools. The MSVD-QA dataset contains 1970 videos and over 50k *Question-Answer* pairs (QA pairs). The MSRVTT-QA dataset is larger, containing over 10k videos and 243k QA pairs. More detailed statistics are shown in Table I. The questions in both datasets can be divided into 5 types, i.e., *what, who, how, when*, and *where*. The detailed statistical results are shown in Table II and Table III.

We remark that compared with MarioQA which involves multiple events per question for high-level reasoning, the two real-world VideoQA datasets, MSVD-QA and MSRVTT-QA, mostly contain questions involving only one event and thus fail to utilize our proposed event recognition self-supervised learning module. Therefore, we will test our proposed SDGraphR model without event recognition self-supervised learning module on the two real-world datasets.

We first describe the comparative baselines and several variants of our model in Section IV-A, then analyze their performance in Section IV-B. To gain more insights into our proposed method and demonstrate its effectiveness, we carry out several ablation studies and provide visualization examples in Section IV-C.

A. Baselines and Model Variations

In this section, we briefly describe the baseline approaches for comparison as well as several variants of our proposed SDGraphR model. The baseline approaches include conventional



Q : *What item was released after an enemy was killed ?*
A : *Mushroom.*
PE : *'Appear', 'Appear,Item', 'Kill,Enemy', 'Kill'*
NE : *'Kill,GreenKoopa', 'Appear,GreenKoopa', 'Shoot,Fireball', 'Appear,BulletBill', 'Kick,Shell' ...*

Fig. 3. An example illustrating our definition of positive events (PE) and a subset of negative events (NE). Given the words *'released'* and *'kill'*, it is intuitive to select *'Appear'*, *'Kill'* and their associated concepts as positive events. Additionally, negative events can be derived from the interplay between the question and visual concepts. For instance, the event *'Kill, GreenKoopa'* involves killing, but it doesn't qualify as a positive event in this context. The reason is that *'GreenKoopa'* is not present in the video clip, rendering this specific event a negative one despite the thematic similarity to a positive event.

TABLE I
Statistics on selected datasets. WE SHOW THE BASIC INFORMATION OF SELECTED DATASETS IN SEVERAL ASPECTS SUCH AS THE NUMBER OF VIDEOS, THE NUMBER OF CLIPS, THE NUMBER OF QA PAIRS, THE MEAN LENGTH OF QUESTIONS AND THE NUMBER OF UNIQUE CANDIDATE ANSWERS.

Dataset	Videos	Clips	QA pairs	Mean question length	Unique Answers
MarioQA	12	167,036	187,757	11.14	57
MSVD-QA	1,970	1,970	50,505	7.62	1,852
MSRVTT-QA	10,000	10,000	243,680	8.35	6,211

TABLE II
Detailed statistics on MSVD-QA dataset.

Split	Clips	QA pairs	Quation Type				
			What	Who	How	When	Where
Train	1,200	30,933	19,485	10,469	736	161	72
Val	250	6,415	3,995	2,168	185	51	16
Test	520	13,157	8,149	4,552	370	58	28
All	1,970	50,505	31,629	17,199	1,291	270	116

TABLE III
Detailed statistics on MSRVTT-QA dataset.

Split	Clips	QA pairs	Quation Type				
			What	Who	How	When	Where
Train	6,513	158,581	108,792	43,592	4,067	1,626	504
Val	497	12,278	8,337	3,439	344	106	52
Test	2,990	72,821	49,869	20,385	1,640	677	250
All	10,000	243,680	166,998	67,416	6,051	2,409	806

VideoQA models and state-of-the-art reasoning models. Besides, our model variations include different variants equipped with different fusion modules and self-supervised training strategies. The following methods including our proposed approach are compared.

- **V-Only** The V-Only model predicts the answer without knowing the questions, and the video features are extracted from a 3D fully convolution network (3DFCN).
- **Q-Only** The Q-Only model predicts the answer without referring to the videos, and the questions are embedded using a GRU pre-trained on a large corpus.
- **1-Step Temporal Att.** The single-step temporal attention model applies a soft attention mechanism for each frame based on the question.
- **Spatio-Temporal Att.** As for the spatio-temporal attention model, the soft attention mechanism is applied throughout the spatio-temporal space.
- **Global Context Embedding** This is the state-of-the-art results reported on the MarioQA dataset, which flattens the video features throughout the spatio-temporal space and uses a multi-layer perceptron for video embedding to capture the global context. The video features and question features are jointly embedded into a common space for final classification.

The above five baselines are all described in the MarioQA dataset [38], readers can refer to Mun et al.'s work [38] for more detailed information.

- **HCRN** The Hierarchical Conditional Relation Network (HCRN) [30] is a video question answering methods based on global-level visual feature. HCRN consists of mutiple Conditional Relation Network (CRN), in which CRN performs reasoning on objects (i.e. the visual features of multiple clips) under conditioning feature (i.e. the text feature and high level visual semantic). HCRN makes the reasoning process a replication, rearrangement and stacking of basic blocks to conduct high-order and multi-step reasoning.
- **L-GCN** L-GCN is a video question answering method based on object-level visual feature. L-GCN takes a fixed number of objects from a single frames of the video clip. Then for every frame in the video, we can obtain graph at

the same size. L-GCN apply graph convolutional network on these object graph for extracting visual features that containing the interactions between different objects.

- **ORN + MAC** The Object Relation Network (ORN) [2] is an object-level model for action recognition, which models pairwise interactions between objects in videos to capture spatio-temporal relationships. The Memory Attention and Composition (MAC) network [17] is a state-of-the-art recurrent reasoning module for visual reasoning in static images, where each MAC cell serves as a general-purpose reasoning step. We combine the two models together to form an *ORN + MAC* baseline for object-level reasoning in videos.
- **SDGraphR.BAN** and **SDGraphR.MAC** Our proposed *SDGraphR.X* model takes *X* as the multimodal fusion module. In our experiments, we employ two representative modules, i.e., the Bilinear Attention Network (BAN) [25] as well as the Memory Attention and Composition (MAC) network [17], as the multimodal fusion module in our proposed SDGraphR model.
- **SDGraphR.MAC (qa+er)** This variant further improves the *SDGraphR.MAC* model with our proposed self-supervised auxiliary event recognition task. The *qa+er* represents training with the primary *question answering* task and the auxiliary *event recognition* task simultaneously. We employ the *pretrain-finetune* training strategy described in Section III-C and will discuss other *multi-task* joint training strategies in ablation studies later in Section IV-C.

For the two real world datasets that are not specifically designed for *event-centric* VideoQA (i.e., one question only refers to one single event), we directly employ recent methods as baselines, and compare them with the proposed *SDGraphR.MAC* (which consistently outperforms *SDGraphR.BAN*) model without self-supervision signals for event recognition task.

B. Performance Analysis

Results on the MarioQA dataset. As shown in Table IV, we provide the QA accuracy of our model on MarioQA dataset versus the baselines. In comparison with the conventional VideoQA model *Global Context Embedding* baseline, our best model *SDGraphR.MAC(qa+er)* improves the overall accuracy by 7.31%, achieving the best performance among all models. In addition, our model also exceeds the baseline methods. Even without event supervision, our SDGraphR model can still surpass these methods, which shows that the dynamic graph construction and text guided optimization method can better represent the video features containing events than the global visual feature and static graph methods. Our model obtains various accuracy boosts for different question types, where the most significant improvement comes from the *event-centric* questions, indicating that our model can better capture the dynamics of events in the videos. We observe that the *Global Context Embedding* baseline outperforms the object-level reasoning model *ORN + MAC* by 1.22%, demonstrating that the combination of object-level action recognition with reasoning model is not effective. When equipped with the self-supervised auxiliary event recognition task, the *qa+er* augmented variants consistently improve the performance of various base models, e.g., $+4.73\%$ for *ORN+MAC* and $+1.66\%$ for *SDGraphR.MAC*.

Results on the real world datasets. As shown in Table V, we provide the results of comparisons between our *SDGraphR.MAC* model and several recent baselines. Here we do not use the *SDGraphR.MAC(qa+er)* model because in most conventional VideoQA datasets there is only one event in a video clip, which does not match the event-centric setting that our event recognition module is designed for. In addition, we provide the comparisons of different question types in Fig 4 and Fig 5 for baselines that take question types into account. Comparing with these recent baseline methods, our *SDGraphR.MAC* model surpasses all the baselines in terms of overall QA accuracy while using the fewest video frames. Given that many of the baseline methods also need spatio-temporal graph to conduct reasoning, the experimental results show that our proposed model can utilize the fewest frames to build a more effective spatio-temporal graph, demonstrating the superiority of our question-guided spatio-temporal graph construction process.

To further explore the impact of different numbers of sampled frames on the prediction accuracy, we gradually increase the number of sampled frames from 1 to 32 and obtain the corresponding accuracies of our proposed *SDGraphR.MAC* model. The comparisons between different sampling numbers of the *SDGraphR.MAC* model and other baseline methods are illustrated in Fig 6. Here we can see that our *SDGraphR.MAC* model is able to maintain good stability when the number of sampled frames is between 2 and 24. When the sampling number is 1, the accuracy decreases, which is reasonable because a single frame can hardly provide any dynamic information in video reasoning. Besides, there is another drop when the number of sampled frames approaches 32, where one possible reason may be that it is necessary to correspondingly increase the complexities of temporal reasoning graph in order to handle the increasing number of sampled frame.

We note that the accuracy of our *SDGraphR.MAC* model is able to exceed 40% when we only sample two frames from each video. On the one hand, this shows that *SDGraphR.MAC* can build adequate spatial-temporal relationships between the two frames. On the other hand, this also validates the stability of the temporal reasoning step in *SDGraphR.MAC* which demonstrates excellent performances when it starts to obtain temporal information from two sequential frames.

More Analysis of the self-supervised event recognition. As shown in Fig 7, when jointly training the primary Question Answering task with the auxiliary Event Recognition task, both *SDGraphR.MAC(qa+er, multi-task)* model and *SDGraphR.MAC(qa+er, pretrain-finetune)* model significantly outperform the *SDGraphR.MAC* baseline at all percentage levels of training data. Specifically, the *multi-task* strategy obtains a larger accuracy boost when using less training data, while the *pretrain-finetune* strategy demonstrates more consistent and substantial improvement with an increas-

TABLE IV

Results on the MarioQA dataset. WE SHOW THE OVERALL QUESTION ANSWERING ACCURACY AND THE ACCURACIES FOR DIFFERENT QUESTION TYPES. THE NT, ET, HT ARE THE DEGREES OF TEMPORAL RELATIONSHIPS INVOLVED IN THE QUESTION, REPRESENTING NO TEMPORAL, EASY TEMPORAL, HARD TEMPORAL, RESPECTIVELY. THE QUESTIONS ARE ALSO CATEGORIZED INTO DIFFERENT TYPES BASED ON THE QUERY PREFERENCES, I.E., EVENT-CENTRIC, STATE AND COUNTING.

Method	Overall	NT	ET	HT	Event	State	Count
V-Only	29.10	21.16	35.32	34.00	-	-	-
Q-Only	38.34	39.79	35.67	39.65	-	-	-
1-Step Temporal Att.	66.82	64.28	69.64	67.21	-	-	-
Spatio-Temporal Att.	69.26	66.38	72.73	69.27	-	-	-
Global Context Embedding	70.02	66.47	75.10	68.89	-	-	-
HCRN [30]	72.54	68.86	77.96	71.15	80.53	84.13	52.85
L-GCN [15]	72.00	68.36	76.90	71.30	79.97	97.75	51.57
ORN + MAC	68.80	66.49	71.88	68.37	75.76	86.30	51.11
ORN + MAC (qa+er)	73.53	71.02	77.38	72.37	81.92	91.90	52.40
SDGraphR.BAN	73.81	72.31	77.36	71.29	81.68	95.58	53.68
SDGraphR.MAC	75.67	75.01	79.02	71.98	84.25	**98.23**	**53.81**
SDGraphR.MAC (qa+er)	**77.33**	**75.63**	**81.72**	**73.94**	**86.97**	92.93	53.36

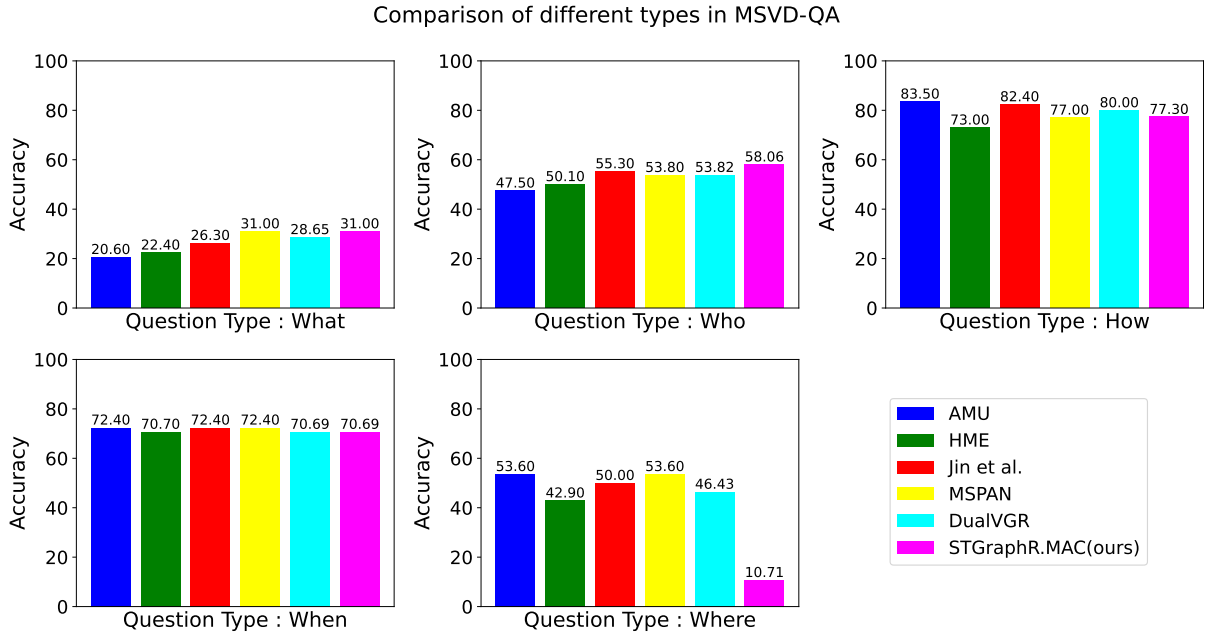


Fig. 4. **Comparisons of different types in MSVD-QA**. Our method achieves the best performances in question type *what* and *who* where question type *what* and *who* accounts for the largest proportion among all question types (account for 96.53% in the test set). The excellent performances in question type *what* and *who* makes our *SDGraphR.MAC* model achieve the best overall performances compared with other baseline methods. It must be mentioned that the lagging 42.89% under question type *where* only means that we have answered 12 more questions incorrectly, while the leading 2.76% under question type *who* means that we have answered 1256 more questions correctly.

ing percentage of training examples. The comparisons of *SDGraphR.MAC(qa+er, multi-task)*, *SDGraphR.MAC(qa+er, pretrain-finetune)* as well as *SDGraphR.MAC* verify that our proposed self-supervised auxiliary event recognition task indeed helps the model to learn the correlations between events and visual contents, thus reducing the requirement for extra training data upon achieving the same accuracy. We would like to point out that since our proposed models exploit the implicit cues in each question to gather the self-supervision for the event recognition task, there will no need for any extra manual annotations in each QA pair.

C. Ablation Studies and Visualizations

In order to gain more insights into our method, we conduct several ablation studies on the MarioQA dataset and provide several visualized examples in Fig 8.

Ablation Studies #1: Effectiveness of several components. We measure the effects of each component of our SDGraphR model by evaluating the accuracy when this part gets removed or replaced with a simpler design.

In order to evaluate the effects of spatio-temporal graph convolutions, we replace it with a fully-connected layer following a ReLU activation. As shown in Table VI (a), replacing the spatio-temporal graph convolutions leads to an accuracy decrease of 2.47%, and removing the long-term temporal

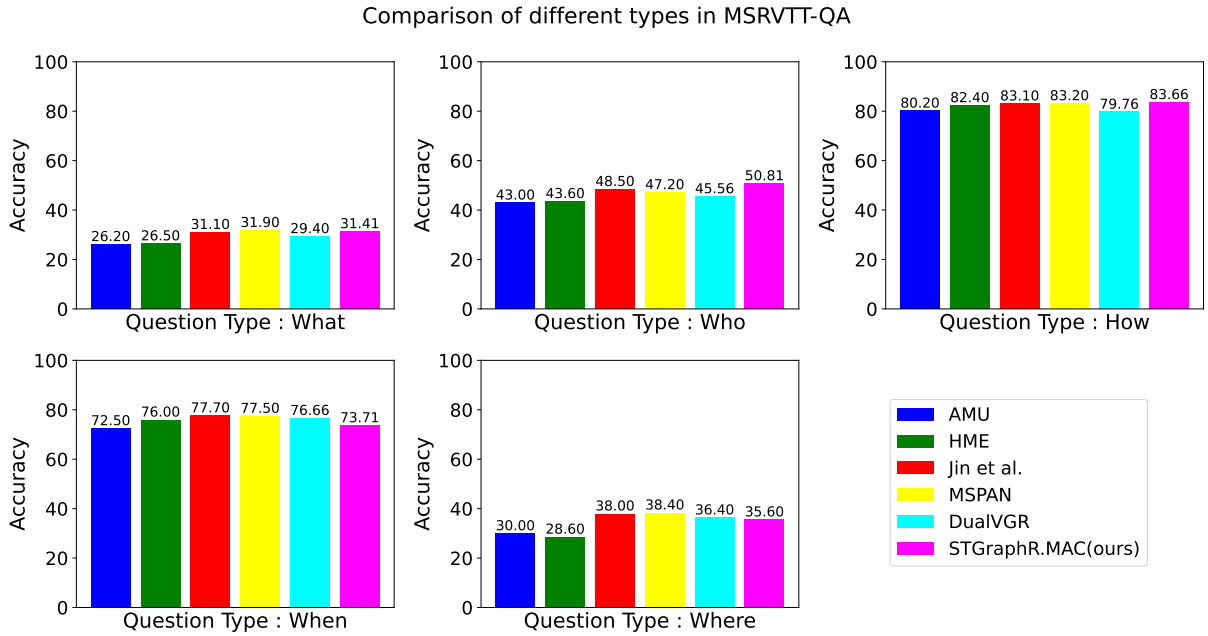


Fig. 5. **Comparisons of different types in MSRVTT-QA**. Our method achieves the best performance in question type*who* and competitive performances to other baseline methods in other question types such as *what* and *how*.

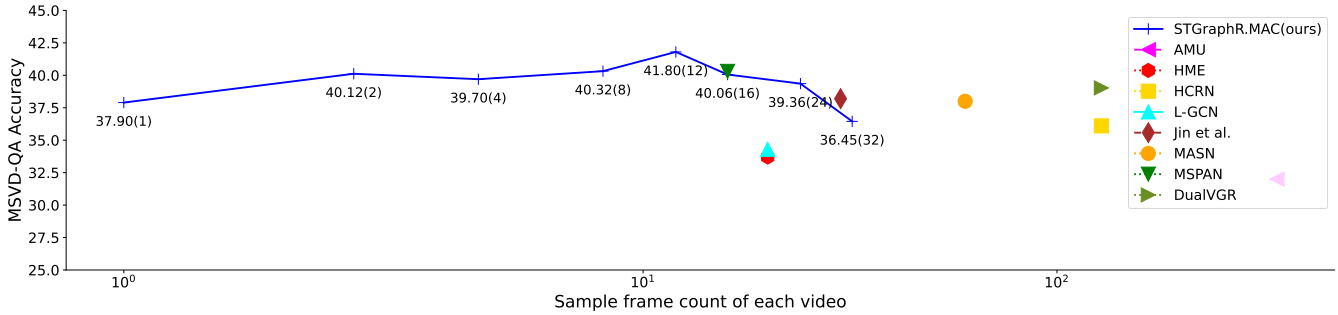


Fig. 6. **Comparisons on different numbers of sampled frames between our method and state-of-the-art baselines.** The horizontal axis indicates the number of sampled frames, while the vertical axis represents the accuracy on the MSVD-QA dataset. For MASN, we assume that the length of each clip is 10 seconds. In the case of DualVGR, we adhere to the standard MSVD-QA setting by sampling 8 clips per video. This results in a total of 128 frames, with each individual clip comprising 16 frames. Besides, we show the detailed results under different numbers of sampled frames for our method, where the accuracy for each configuration is annotated beneath its respective data point, and the number of sampled frames presented in parentheses.

TABLE V
Results on real world dataset. WE SHOW THE RESULTS OF DIFFERENT METHODS WITH RESPECT TO THE ACCURACY AND THE NUMBER OF SAMPLED FRAMES NEEDED TO ACHIEVE THE REPORTED PERFORMANCES IN EACH DATASET.

Method	Sampled frames	MSVD-QA	MSRVTT-QA
AMU [59]	20 + 20 × 16	32.00	32.50
HME [6]	20	33.70	33.00
HCRN [30]	8 × 16	36.10	35.60
L-GCN [15]	20	34.30	/
Jin et al. [21]	30	38.20	37.60
MASN [46]	6 × clip time	38.00	35.20
MSPAN [10]	16	40.30	37.80
DualVGR [55]	(8 or 16) × 16	39.03	35.52
SDGraphR.MAC	**12**	**41.80**	**38.42**

TABLE VI
ABLATION RESULTS ON THE MARIOQA DATASET.

Ablation	Accuracy	Acc.Diff
- Base model *(qa)*	75.67	0.00
a) - model components		
- w/o st-graph convolution	73.20	-2.47
- w/o long-term modeling	72.53	-3.14
b) - spatio-temporal graph		
- with heuristic s-graph	75.28	-0.39
- with only s-graph (learnable)	75.17	-0.50
- with only s-graph (heuristic)	73.94	-1.73
- with only t-graph	75.45	-0.22
- mixing s-graph and t-graph	73.88	-1.79
- w/o residual connections	73.37	-2.30
c) - self-supervised training		
- with multi-task	76.46	+0.79
- with pretrain-finetune	77.33	+1.66

modeling (i.e., without the RNN) causes a decrease of 3.14%.

Ablation Studies #2: Design Choices for the Spatio-

TABLE VII
EXAMPLES OF SELF-SUPERVISION FOR THE AUXILIARY EVENT RECOGNITION SELF-SUPERVISED LEARNING TASK. WE SHOW THE POSITIVE EVENTS EXPLOITED FROM THE QUESTION ITSELF. WE LIST THE POSITIVE EVENTS THAT CAN BE DIRECTLY IMPLIED FROM THE QUESTION, AS WELL AS NEW POSITIVE EVENTS IN TURN DERIVED FROM THE EXISTING POSITIVE EVENTS.

Questions	Positive Events
How many goombas were killed by fireballs ?	*(Kill,Goomba,Fireball)*; *(Kill)*; *(Kill,Goomba)*;
How many times did Mario eat coins before a fireflower appears ?	*(Eat,Coin)*; *(Appear,FireFlower)*; *(Eat)*; *(Eat,Item)*; *(Appear)*; *(Appear,Item)*;
Where did a red koopa troopa appear after Mario held a shell ?	*(Appear,RedKoopa)*; *(Hold,Shell)*; *(Appear,Enemy)*; *(Appear)*;
Where did Mario eat an item after killing a green koopa troopa ?	*(Eat,Item)*; *(Kill,GreenKoopa)*; *(Eat)*; *(Kill)*; *(Kill,Enemy)*;
Where did Mario punch a mushroom block before hitting a coin block ?	*(Hit,Mushroom Block)*; *(Hit,Coin Block)*; *(Hit,Block)*
Where did Mario stomp on an enemy before kicking a shell ?	*(Stomp,Enemy)*; *(Kick,Shell)*; *(Stomp)*
Who did Mario kill after the appearance of a bullet bill	*(Kill)*; *(Appear,BulletBill)*; *(Appear)*; *(Appear,Enemy)*;

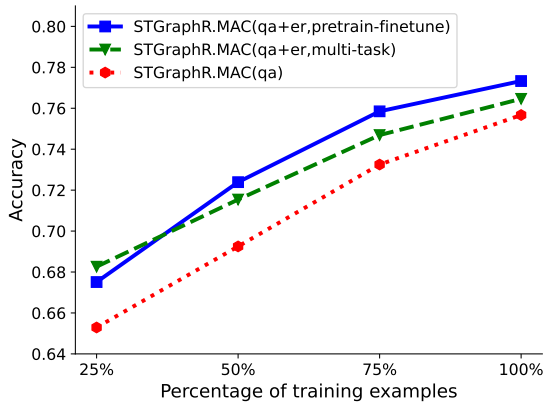


Fig. 7. **Comparisons of QA accuracies on the MarioQA dataset with/without self-supervised auxiliary event recognition task.** The y-axis denotes the QA accuracy, while the x-axis shows the percentage of training examples.

Temporal Graphs.

In our full model, we stack the learnable spatial graph and heuristic temporal graph in the order of $\tilde{S}\tilde{S}STTT$ when performing convolutions, where we use S/T for "heuristic spatial/temporal graph" and $\tilde{S}/\tilde{T}$ for "learnable spatial/temporal graph". As shown in Table VI (b), replacing the learned s-graphs with heuristic ones ($SSTTT$) leads to an accuracy decrease of 0.39%, meanwhile, using only the heuristic s-graphs performs much worse than the learnable ones (-1.73% v.s. -0.50%). The accuracies drop by different degrees when using only the s-graphs or the t-graphs (i.e., -0.22% and -0.50%). It is also possible to mix the s-graph with t-graph to form a unified graph, but it performs 1.79% worse than to process them individually. As for the graph convolutions, removing the residual connections between layers harms the performance by 2.30%. These results validate the effectiveness of our design choices for the spatio-temporal graphs.

Visualizations. Fig 8 illustrates a couple of examples with videos and the questions, where we compare our proposed SDGraphR.MAC(qa+er) model with the ORN+MAC approach. It is quite clear that the proposed SDGraphR.MAC(qa+er) model demonstrates its superiority over existing method through correctly answering questions involving more complex events with dynamic patterns.

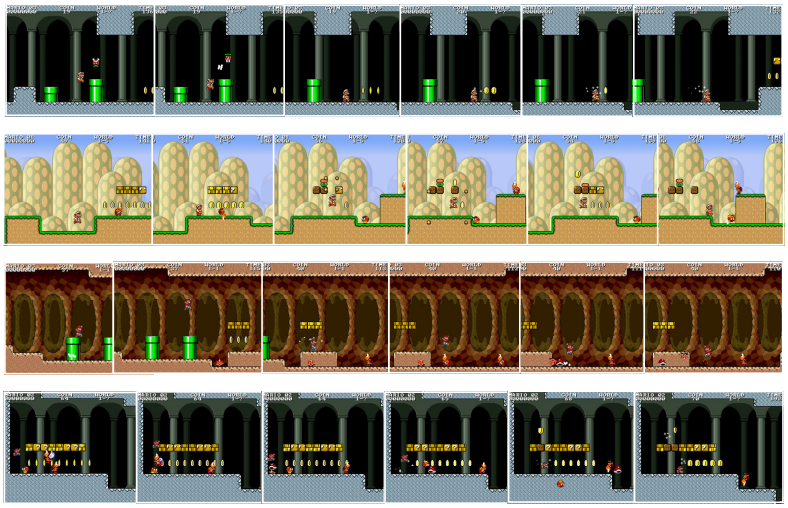
V. IMPLEMENTATION DETAILS

We sample 12 frames from each video with a *linspace* method, i.e., the same intervals between two adjacent sampled frames. Before being fed into the object detector, the video frames are normalized by extracting the mean RGB values of the dataset. We apply a batch normalization layer [19] on the object's visual features $\boldsymbol{v}$ before running the graph convolutions. In our model variants, we adopt two representative modules (the Bilinear Attention Network (BAN) [25] and the Memory Attention and Composition (MAC) network [17]) as the multimodal fusion module to reach an answer for the given question. In detail, the BAN module uses 2 glimpse operations and sets the hidden size equal to 512, while the MAC module performs 4 recurrent steps.

In the experiments, we follow the official train/validation splits of the MarioQA, MSVD-QA and MSRVTT-QA dataset. We constitute a set of candidate answers that appear at least 10 times in the training set. The probability of being the correct answer is predicted by a multi-layer perceptron (MLP) classifier with the dimensions of $(1024, 2048, classes)$.

We train our model using the Adam optimizer [26] with a batch size of 8 and an initial learning rate as 0.00005 for MarioQA dataset, and with an initial learning rate as 0.0005 for real world datasets, setting batch size to 16 for MSVD-QA dataset and 64 for MSRVTT-QA dataset. The initial learning rate gradually warms up to 1x, 2x, 3x, 4x times larger during the first 4 epochs, and decays by 50% for every 2 to 5 (depending on the size of dataset) epochs after the 10_{th} epoch. For the multi-task training strategy, we set the hyper-parameter $\eta = 0.5$ to balance the two multi-task losses between Question Answering and Event Recognition.

In Table VII, we provide more examples showing that our SDGraphR model is able to exploit the cues in questions to gather self-supervision for the auxiliary event recognition self-supervised learning task. We list the positive events that can be directly implied from the question, as well as new positive events in turn derived from the existing positive events. Take the last example in Table VII as an example, three positive events, i.e., *(Kill)*, *(Appear, BulletBill)* and *(Appear)* can be discovered directly from the given question, and another new positive event, e.g., *(Appear, Enemy)*, can then be derived based on the three discovered positive events.



Q : *What type of stage is Mario in?*
GT Answer: *castle.*
Base: *castle.* ✔ **Ours**: *castle.* ✔

Q : *What enemy was killed by a fireball after a fireflower block was hit by Mario?*
GT Answer: *goomba.*
Base: *red koopa troopa.* X **Ours**: *goomba.* ✔

Q : *By what means did Mario attack a spiky after Mario killed a red koopa troopa?*
GT Answer: *by a shell.*
Base: *by stomping.* X **Ours**: *by a shell.* ✔

Q : *How many enemies were killed by Mario?*
GT Answer: *4.*
Base: 2. X **Ours**: 3. X

Fig. 8. **Visualized examples on MarioQA.** We show the results of the *SDGraphR.MAC(qa+er)* model and compare it with the *ORN + MAC* baseline.

VI. Conclusions

In this paper, we study the problem of video question answering (VideoQA) and propose the Self-supervised Dynamic Graph Reasoning (SDGraphR) model for VideoQA in event-centric scenarios. Our SDGraphR model learns question-guided object-level spatio-temporal graphs, which dynamically encapsulate the intra-frame spatial correlations and inter-frame temporal correspondences among objects in the video. The self-supervised learning framework exploits the implicit cues hidden in questions to gather self-supervision for an auxiliary event recognition task. By utilizing object-level representations and imposing event-level semantic constraints, our model consistently boosts the performances in various experimental settings. We validate the effectiveness of our approach in an event-centric gameplay VideoQA dataset and two conventional real-world VideoQA datasets, showing that our proposed SDGraphR model exhibits superior performances and achieves substantial improvement over several state-of-the-art baselines.

The event-centric approach is suitable for scenes that contain a large number of interactions in a short time like game videos or scenes that contain a large number of objects like sport videos, however, when the interactions and objects become more and more complex, well-designed event auxiliary also needs to be further improved. In the future, we will explore how to maintain the effectiveness and efficiency of graph modeling in longer videos or more sampling frames. At the same time, complex hierarchical relationships rather than simple inclusion relationships of real-world events also need to be further explored.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[2] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[6] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007, 2019.

[7] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023.

[8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018.

[9] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021.

[10] Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Lingling Li. Multi-scale progressive attention network for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 973–978, 2021.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. Spatio-temporal action graph networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[14] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017.

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

[16] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems*, pages 5903–5916, 2019.

[17] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for compositional question answering over real-world images. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[21] Weike Jin, Zhou Zhao, Xiaochun Cao, Jieming Zhu, Xiuqiang He, and Yueting Zhuang. Adaptive spatio-temporal graph enhanced vision-language representation for video qa. *IEEE Transactions on Image Processing*, 30:5477–5489, 2021.

[22] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[24] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.

[25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[28] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008.

[29] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[30] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020.

[31] Guohao Li, Xin Wang, and Wenwu Zhu. Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 530–538, 2019.

[32] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[33] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[35] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4942–4950, 2018.

[36] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[37] Effrosyni Mavroudi, Benjamín Béjar Haro, and René Vidal. Representation learning on visual-symbolic graphs for video understanding. In *European Conference on Computer Vision*, pages 71–90. Springer, 2020.

[38] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[39] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[40] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[42] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[43] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[45] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360, 2007.

[46] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6167–6177, Online, August 2021. Association for Computational Linguistics.

[47] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[48] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247, 2018.

[49] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[52] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hypergraphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14879–14889, 2023.

[53] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR 2011*, pages 3169–3176. IEEE, 2011.

[54] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[55] Jianyu Wang, Bingkun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 2021.

[56] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[57] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[58] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.

[59] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[60] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296, 2016.

[61] Zenan Xu, Wanjun Zhong, Qinliang Su, Zijing Ou, and Fuwei Zhang. Modeling semantic composition with syntactic hypergraph for video question answering. *arXiv preprint arXiv:2205.06530*, 2022.

[62] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832, 2017.

[63] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[64] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[65] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019.

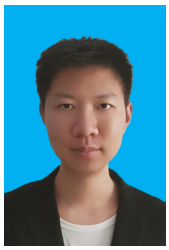
[66] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. IJCAI, 2020.

Jie Nie is currently an Associate Professor at Ocean University of China where she received the B.S.and Ph.D. degrees in 2002 and 2011, respectively, all in computer science. She was a visiting scholar at University of Pittsburgh, USA from Sept. 2009 to Sept. 2010. She was a postdoctoral fellow with Tsinghua University from 2015 to 2017. Her current research interests lie in the fields of artificial intelligence, multimedia analysis, and visual analysis of spatio-temporal big data. She has published high-quality papers in top journals and conferences in the area of artificial intelligence including IEEE Transactions on Multimedia, ACM Multimedia, ACM Transactions on Multimedia Computing, Communications and Applications etc. She received IEEE CCIS 2016 Best Paper Award.

Xin Wang is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications. He has published over 150 high-quality research papers in ICML, NeurIPS, IEEE TPAMI, IEEE TKDE, ACM KDD, WWW, ACM SIGIR, ACM Multimedia etc., winning three best paper awards including ACM Multimedia Asia. He is the recipient of ACM China Rising Star Award, IEEE TCMC Rising Star Award and DAMO Academy Young Fellow.

Runze Hou recived his B.E. degree from the School of Automation, Southeast University, Nanjing, China in 2021. He is currently working toward his M.S. degree in the Tsinghua-Berkerly Shenzhen Institute of Tsinghua University. His research interests include deep learning, multi-modal learning.

Guohao Li received his Ph.D. degree in computer science and technology from Tsinghua University in 2021. His main research interests include multimedia analysis, computer vision and deep learning. He is currently a researcher in Baidu, Beijing, China.



Hong Chen received B.E. from the Department of Electronic Engineering, Tsinghua University, Beijing, China in 2020. He is currently a Ph.D. candidate in the Department of Computer Science and Technology of Tsinghua University. His main research interests include machine learning, multi-modal information processing.

Wenwu Zhu is currently a Professor in the Department of Computer Science and Technology at Tsinghua University. He also serves as the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs, New Jersey as Member of Technical Staff during 1996-1999. He received his Ph.D. degree from New York University in 1996. His research interests are in the area of data-driven multimedia networking and Cross-media big data computing. He has published over 380 referred papers and is the inventor or co-inventor of over 80 patents. He received eight Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019.

He served as EiC for IEEE Transactions on Multimedia from 2017-2019. He served in the steering committee for IEEE Transactions on Multimedia (2015-2016) and IEEE Transactions on Mobile Computing (2007-2010), respectively. He serves as General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019, respectively. He is an AAAS Fellow, IEEE Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).