Event Graph Guided Compositional Spatial–Temporal Reasoning for Video Question Answering

Ziyi Bai, Ruiping Wang, *Senior Member, IEEE*, Difei Gao, *Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*— Video question answering (VideoQA) is challenging since it requires the model to extract and combine multi-level visual concepts from local objects to global actions from complex events for compositional reasoning. Existing works represent the video with fixed-duration clip features that make the model struggle in capturing the crucial concepts in multiple granularities. To overcome this shortcoming, we propose to represent the video with an *Event Graph* in a hierarchical structure whose nodes correspond to visual concepts of different levels (object, relation, scene and action) and edges indicate their spatial-temporal relationships. We further propose a *H*ierarchical *S*patial-*T*emporal *T*ransformer (HSTT) which takes nodes from the graph as visual input to realize compositional reasoning guided by the event graph. To fully exploit the spatial-temporal context delivered from the graph structure, on the one hand, we encode the nodes in the order of their semantic hierarchy (depth) and occurrence time (breadth) with our improved graph search algorithm; On the other hand, we introduce edge-guided attention to combine the spatial-temporal context among nodes according to their edge connections. HSTT then performs QA by cross-modal interactions guaranteed by the hierarchical correspondence between the multi-level event graph and the cross-level question. Experiments on the recent challenging AGQA and STAR datasets show that the proposed method clearly outperforms the existing VideoQA models by a large margin, including those pre-trained with large-scale external data. Our code is available at https://github.com/ByZ0e/HSTT.**

***Index Terms*— VideoQA, video representation, transformer, spatial-temporal reasoning, compositional reasoning.**

I. Introduction

In THE past decade, AI systems have been required to reason about the static scenes (images) from basic objects to more complex scenes [1], [2], [3], [4], [5]. Toward real-world applications, what they need to face is a dynamic world with constant events happening around them. Compared to the static scene simply composed of objects with spatial interaction among them, the visual events in videos comprise complicated spatial-temporal dependencies of multi-granularity visual clues from local objects to global actions. Accordingly, Video Question Answering (VideoQA), is much more challenging for it requires the system to fulfill multi-step reasoning across different understanding levels as shown in Figure 1. Sometimes, the system needs to sort the actions according to the changing process of local human-object relation (like **Q1**), while sometimes it has to locate the specific object that a person is interacting with during a certain period (like **Q2**). To equip the model with such cross-level reasoning ability, a hierarchical representation with clear spatial-temporal relations for characterizing video content becomes the prerequisite.

However, most of the existing VideoQA models follow the tradition of broadly representing the video as single-level (frame-level or clip-level) features [6], [7], [8], [9]. While they have achieved promising results on relatively simple questions, these methods often struggle when encountering complex cross-level reasoning questions [10]. This limitation is most likely caused by a lack of explicitly characterizing the multi-level visual concepts in the video, which incurs considerable difficulty for the model to uncover the visual clues in different scales from such single-level unstructured video representation.

Targeting this issue, we propose a hierarchical event graph to clearly represent the spatial-temporal structure of the videos. Motivated by the prior scene graphs [11], [12] and action graphs [13], [14], [15], we construct a uniform four-level hierarchical graph, including `Object`, `Relation`, `Scene` and `Action` to roughly mimic the bottom-up visual understanding process of our human-beings [16]. That is, we first determine the interaction relationship between salient objects in a single frame (i.e., object and relation level), then combine multiple relationships to get an overall understanding of the static scene (i.e., scene level), and finally recognize the dynamic actions that the person is performing by continuously observing frames during a time period (i.e., action level). We use graph nodes to represent the above four levels of visual concepts extracted from the video. Each node has its specific semantic level, timestamp, and confidence score. Then we link them with edges to indicate their spatial-temporal relationships.

Manuscript received 23 February 2023; revised 20 October 2023; accepted 9 January 2024. Date of publication 31 January 2024; date of current version 2 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0111901 and in part by the Natural Science Foundation of China under Contract U21B2025 and Contract U19B2036. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hichem Sahbi. *(Corresponding author: Ruiping Wang.)*

The authors are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ziyi.bai@vipl.ict.ac.cn; wangruiping@ict.ac.cn; difei.gao@vipl.ict.ac.cn; xlchen@ict.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2024.3358726, provided by the authors.

Digital Object Identifier 10.1109/TIP.2024.3358726

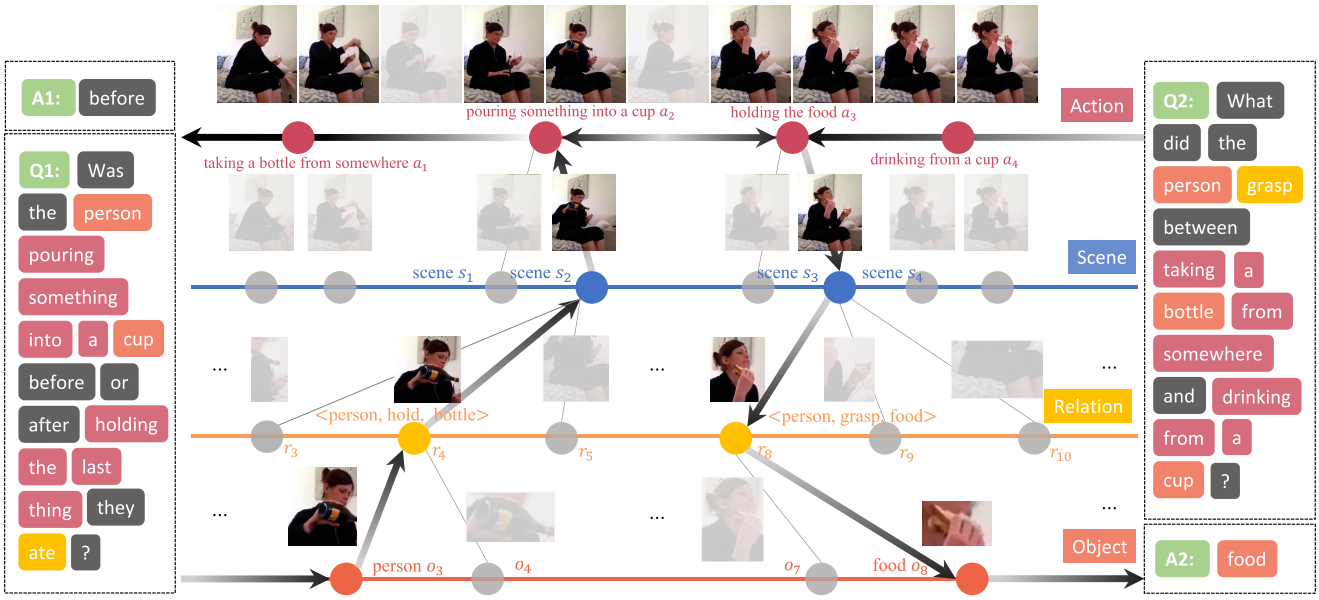


Fig. 1. VideoQA involves multi-step reasoning across four understanding levels (Object, Relation, Scene and Action) of the videos. The Scene here denotes a static frame at time t serving as the connecting link between local and global visual elements. In the example on the left (**Q1**), the reasoning route (black arrow) starts from the person in the Object level and infers upward to obtain the sequence of Action; The right case (**Q2**) begins from the top Action level and performs reasoning downward to obtain the bottom Object level information.

We further propose a structure-aware Hierarchical Spatial-Temporal Transformer (HSTT) for compositional reasoning. Different from most existing multi-modal Transformers, HSTT takes the nodes from the above event graph as visual input. What's more, it combines the spatial-temporal context of them to achieve the answer guided by the graph structure prior. On the one hand, we encode nodes in the unique graph traversal order to preserve their hierarchical order and temporal order. On the other hand, we further follow the event graph structure to prune the fully self-attention interactions among graph nodes. Namely, we prevent those interactions between nodes that are not linked with spatial-temporal edges. In this way, redundant spatial-temporal context aggregation is avoided in the model. Then the HSTT performs QA by jointly encoding the nodes in the event graph and words from the question. For cross-modal concepts from the same semantic layer, *e.g.*, object nodes & object words, we add consistent object level embeddings to guide the model to better establish hierarchical semantic correspondence. Therefore, the model can efficiently ground the visual concepts required by the cross-level question and combine them for multi-step compositional reasoning.

Experiments on the challenging AGQA [10] and STAR [17] datasets show the proposed method outperforms existing VideoQA models by a large margin, including those models pre-trained with millions of external data. It sufficiently indicates the effectiveness and superiority of our method. In summary, our main contributions are: 1) parsing the complex video content into a structured event graph, which is composed of multi-level visual concepts being well organized with their spatial-temporal relationships; 2) Exploiting the event graph structure prior to guide the model to precisely encode the spatial-temporal context information among the nodes; 3) Proposing a VideoQA model aiming to address multi-step compositional reasoning based on the hierarchical semantic alignment between the graph and the question.

II. Related Works

In general, the challenges of VideoQA lie in two aspects. 1) The agent needs to reason about the complex video content which is composed of various visual concepts from temporal action to spatial human-object relationship [10]. Therefore prior VideoQA works have introduced different structures to represent the video, which are shown in Figure 2. 2) The agent has to ground the visual clues from these video representations and combine them for compositional reasoning. We introduce recent three types of VideoQA models in the following.

A. Holistic Spatial-Temporal Models

A couple of early attempts [7], [8], [18], [19], [20], [21], [22], [23] have been made to jointly model the spatial content and temporal dynamics in the video. Some works like [23] take the temporal difference to encode motion information of video. Others usually extract the appearance and motion features of video units (clip or frame) separately which is shown as Figure2(a) and let them interact with the question to find the answer. MACN [18] and HME [7] propose co-memory and co-attention mechanisms respectively to capture appearance-question and motion-question relations. B2A [8] further uses the question as a bridge for appearance and motion interaction. HCRN [19] explores a temporal hierarchy of video representation from frame level to video level by stacking a reusable unit conditioned by motion and question features. While both appearance and motion information are involved, this type of works all take holistic video features for dynamic reasoning, which as a kind of coarse representation is hard to depict the multi-granularity visual elements of the

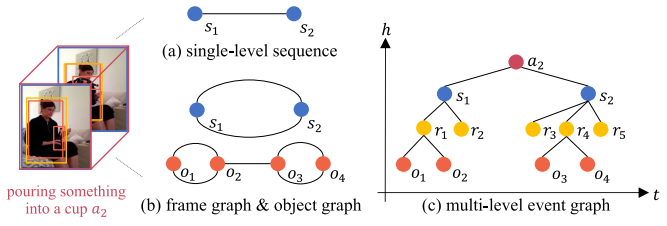


Fig. 2. Different video representation structures. (a) Single frame- or clip-level feature sequence. (b) Graph with frame nodes or object nodes. (c) Our multi-level event graph whose nodes are visual concepts arranged in their semantic hierarchies and occurrence time.

complex event from local object interactions to global human actions.

B. Graph-Structured Models

Recently, many works [8], [9], [24], [25], [26], [27], [28], [29], [30], [31], [32] try to exploit graph structure to model the complex relations in VideoQA task including temporal relations, spatial relations and visual-linguistic relations. Some works [8], [9], [27] take video units as nodes and use the graph to model their temporal relations, namely, the frame graph shown as Figure 2(b). Among them, [8] and [27] further parse the question into a graph to let it interact with the video unit graph for visual-linguistic relations modeling. Since nodes in such graphs contain unparsed mixed semantics, the model is unable to explicitly establish the semantic correspondence between the video and the question.

Apart from them, [24], [25], [26], [28], [29], [30], [31], and [33] introduce an object detection module to obtain object nodes of each frame/segment to construct object graph shown as Figure 2(b). Most of them only apply static intra-frame object interactions to enrich the frame representation with local information, which cannot depict the temporal dynamics of objects. References [28], [29], and [31] further apply object interactions across video sequences for dynamics modeling. VGT [30] further uses object edges to model the relations. VISA [34] connects object-action edge for dynamic context modeling. Compared to monolithic frame and object graphs, our proposed hierarchical event graph in Figure 2(c) explicitly depicts four-level visual concepts of an event from local to global and their spatial-temporal relationship.

C. Multi-Modal Transformer

Recently, the Transformer architecture has been widely used in video-language pre-training models [35], [36], [37], [38], [39], [40], [41] to learn joint cross-modal representation and has shown great efficiency in many downstream tasks including VideoQA. Most of these works take frame-level or clip-level features as visual inputs. However, the video frames have mixed semantics that differ from the specific tokens in the questions. Thus, it causes difficulties for the model to learn the cross-modal correspondence between visual features and textual features. A group of latest works [36], [42], [43], [44] start to study hierarchical learning of vision-language representation. ActBERT [36] involves objects and actions as visual inputs into Transformer to learn both local and global joint representations. SHA [43] proposes to use the Transformer Decoder to implicitly identify relations and actions for each video frame and concatenate them as abstract video representation. All these models take the unstructured visual elements as input aiming to learn better cross-modal correspondence. Compared with them, HSTT can more efficiently build the cross-modal correspondence thanks to the rich structure prior from the event graph.

There also exist recent efforts [45], [46], [47] in other fields that explore the possibility of combining graph and Transformer, which are designed to replace the positional encoding in Transformer with graph Laplacian eigenvectors to encode structure information. Reference [47] further proposes to jointly encode local to cluster node information in the graph. However, the nodes in the graph are all identical elements. Different from them, our HSTT can encode hierarchical graphs whose nodes are from multiple semantic levels. Besides, the latest works [17], [30], [31] share similar motivation as ours to solve the VideoQA task with graph reasoning under the Transformer pipeline. However, most of them simply apply the self-attention mechanism of the Transformer to aggregate the graph context. Reference [31] further designs a kernel attention mechanism to aggregate spatial-temporal context of the graph nodes. These works can hardly guide the model to learn the multi-level semantic correspondence between the graph and the question. In contrast, HSTT explicitly applies the structure prior of the event graph at the input and encoding stages separately to enhance the cross-modal reasoning ability of the Transformer.

III. Method

Given a video $\mathbf{V}$ and question $\mathbf{q}$ as input, we first construct a hierarchical event graph that gives a structured representation of the visual event (Section III-A). Then we introduce our Hierarchical Spatial-Temporal Transformer (HSTT) in the order of its input embeddings, jointly cross-modal modeling, edge-guided attention mechanism and answer prediction (Section III-B).

A. Event Graph Construction

Figure 2(c) shows the structure of our proposed multi-level event graph. Let an undirected graph $\mathcal{G}_{evt} = (\mathcal{V}, \mathcal{E})$ represent the event graph of a video with T frames. The node set $\mathcal{V}$ defines different visual elements contained in the event and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ describes the spatial-temporal context correlation among the visual elements. As introduced in Section I, the visual elements can be divided into four types, `Object`, `Relation`, `Scene` and `Action`, corresponding to the different levels of a video, *i.e.*, $\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_r \cup \mathcal{V}_s \cup \mathcal{V}_a$. Each node has its timestamp t and confidence score c. There are two types of edges $\mathcal{E} = \mathcal{E}_{intra} \cup \mathcal{E}_{inter}$. The intra-level edges $\mathcal{E}_{intra}$ indicate those ones that connect the nodes from the same level to describe their long-form temporal dependencies. The inter-edges $\mathcal{E}_{inter}$ can be divided into three subsets $\mathcal{E}_{inter} = \mathcal{E}_{r,o} \cup \mathcal{E}_{s,r} \cup \mathcal{E}_{a,s}$, where each subset denotes the spatial-temporal inclusion relationship between two adjacent levels. With such definitions, we construct the complete event graph in a bottom-up manner.

1) Object and Relation Level: Given a video with T frames, we extract a set of object nodes $\mathcal{V}_o^t = \{\mathbf{o}_i\}_{i=1}^M$ and a series of relation nodes $\mathcal{V}_r^t = \{\mathbf{r}_k\}_{k=1}^N$ for each frame at t, where M, N indicate the number of objects and relations respectively. Similar to the scene graph generation task [48], firstly, an object detection network is used to predict the object proposals. The region of interest (RoI) pooling is applied to extract the object features of each object node $\mathbf{o}_i \in \mathbb{R}^{d_o}$. Then a relation prediction head is concatenated to the object detection network to predict the relation triplets $\langle$`Subject`, `Predicate`, `Object`$\rangle$ for each pair of objects. The conjunction vector $\mathbf{r}_k = [\mathbf{o}_i, \mathbf{p}_{ij}, \mathbf{o}_j] \in \mathbb{R}^{d_r}$, $1 \le i, j \le M$ is used to represent the relation node, where $\mathbf{p}_{ij}$ is the union spatial feature of the object pair $\mathbf{o}_i$ and $\mathbf{o}_j$ characterizing their relation. The relation prediction also outputs the confidence scores of the object and relation nodes. Naturally, each $\mathbf{r}_k$ is connected to its associated subject node $\mathbf{o}_i$ and object node $\mathbf{o}_j$ with relation-object edge $e_k \in \mathcal{E}_{r,o}$ to indicate their correspondence.

2) Scene and Action Level: Scene level is the most commonly used video representation in previous studies. We simply use Convolutional Neural Network (CNN) to extract 2D feature of each frame to denote the scene node $\mathcal{V}_s = \{\mathbf{s}_t\}_{t=1}^T \in \mathbb{R}^{d_s}$ as [38]. For the action level, we first apply 3D-CNN model to extract video feature $\{\mathbf{f}_t\}_{t=1}^T \in \mathbb{R}^{d_f}$, and then feed them into an action proposal generation network to generate a series of action proposals with their start and end timestamps $[t_l^s, t_l^e]_{l=1}^P$, $1 \le t_l^s \le t_l^e \le T$, where P is the number of action proposals. The action node representation denoted as $\mathcal{V}_a = \{\mathbf{a}_l\}_{l=1}^P \in \mathbb{R}^{d_a}$, is finally obtained by pooling the involved features corresponding to each proposal as $\mathbf{a}_l = pool([\mathbf{f}_{t_l^s}, \ldots, \mathbf{f}_{t_l^e}])$. Then, each action node $\mathbf{a}_l$ is linked to all the scene nodes within its time span by action-scene edge $e_l \in \mathcal{E}_{a,s}$ to indicate the scene changing process contained in the action. At last, we associate the scene node $\mathbf{s}_t$ through all relation nodes $\mathcal{V}_r^t$ extracted at time t with scene-relation edge $e_t \in \mathcal{E}_{s,r}$ to depict the spatial layout of each static scene.

B. Hierarchical Spatial-Temporal Transformer

With the structured representation of a visual event, our Hierarchical Spatial-Temporal Transformer (HSTT) jointly encodes the extracted visual concepts in the event graph and semantic words in the questions by a multi-layer Transformer. Figure 3 shows the overall framework of HSTT.

Input Embeddings: Different from most of the current multi-modal Transformers, the visual input of our model is graph nodes rather than image patches. Therefore, there are four types of input elements to be embedded in HSTT: in addition to graph nodes and question words, we also inject their hierarchical information and the node position information. We embed each of them with a corresponding embedder to get different input embeddings for the model.

1) Node & Word Embeddings: To embed graph nodes, we project all node features in each level into a common D-dimensional embedding space through corresponding linear transformations $\mathbf{W}_* \in \mathbb{R}^{D \times d_*}$ (we use $*$ to simply represent any level). Then we can get their corresponding embedding $\mathbf{X}_* = \mathcal{V}_* \mathbf{W}_*^T$, where $\mathcal{V}_*$ is the extracted node feature. In consideration that the original Transformer [49] can only encode the sequence input, we arrange the embeddings into a node sequence in the order of `Object`, `Relation`, `Scene` and `Action` as $\mathbf{X}_N = [\mathbf{X}_o, \mathbf{X}_r, \mathbf{X}_s, \mathbf{X}_a]$.

For question (together with answer candidates for multi-choice questions), we follow BERT [50] to embed the question into a token sequence $\mathbf{X}_q$ with the length $|\mathbf{q}|$. Each word in the question is embedded with Word-Piece embeddings using a 30,000-token vocabulary that has the same dimension D as node embedding. And special tokens `[CLS]` and `[SEP]` are added at the beginning and the end of the question. For multi-choice questions, we follow [26], [38] to concatenate each answer option to the question to form a set of query representations $\{\mathbf{X}_q^n\}_{n=1}^{|\mathcal{N}|}$, where $|\mathcal{N}|$ is the number of answer options.

2) Level Embedding: To answer the complex compositional question, the model needs to find the corresponding visual concepts in the structured event graph with respect to the particular question. For each question, we tokenize it and tag each word with Part-of-Speech (POS), *i.e.*, categorize them into `Object`, `Relation`, `Action` and others. Since both graph nodes and question words consistently carry hierarchical semantics, the model can learn multi-level cross-modal alignment between them. To distinguish they are from different levels, we add corresponding level embeddings $\mathbf{X}_H \in \{\phi_h(h), 0 \le h \le 4\}$, to each of the graph nodes and question words, where $\phi_h(.)$ is a learnable function. The values of h ranging from 1 to 4 correspond to level from bottom `Object` to top `Action` respectively. For those words not belonging to any level, we set $h = 0$.

3) Order Embedding: As shown in Figure 4(a), tokens are treated equally and encoded in parallel in Transformer for long dependency modeling. However, this will shuffle the original context relationships among visual tokens, which are important for video understanding. Most of the current video Transformers [36], [51] exploit segment or frame encoding as a flexible way to inject sequential order of video tokens. Different from them, our node tokens are well organized in a hierarchical graph structure rather than a simple node sequence. As Section III-A introduced, the nodes are arranged according to their semantic hierarchy h from bottom to top. Meanwhile, nodes in the same hierarchy are arranged in the temporal order t.

To preserve the hierarchical and temporal order of nodes at the same time, we turn to traditional graph search algorithms. By taking different travel strategies, *e.g.*, depth-first and breadth-first, they can adapt the search direction along the edges at each step and finally provide an overall graph search order. In the event graph, the depth search direction is the semantic hierarchy descending direction, *i.e.*, $-h$, and the breadth search direction is the occurrence time ascending direction, *i.e.*, t. As shown in Figure 4(b), when taking the event graph as input, Breadth-First Search (BFS) algorithm travels the graph firstly according to temporal order then hierarchical order while Depth-First Search (DFS) algorithm takes hierarchical-order-first strategy on the contrary. As shown in the example, DFS gives an order

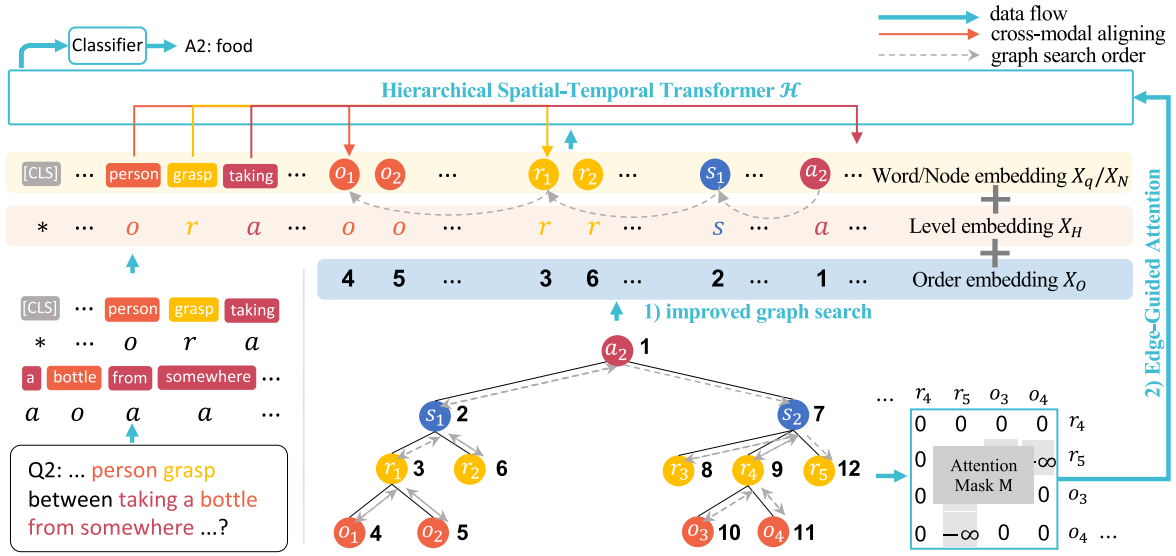
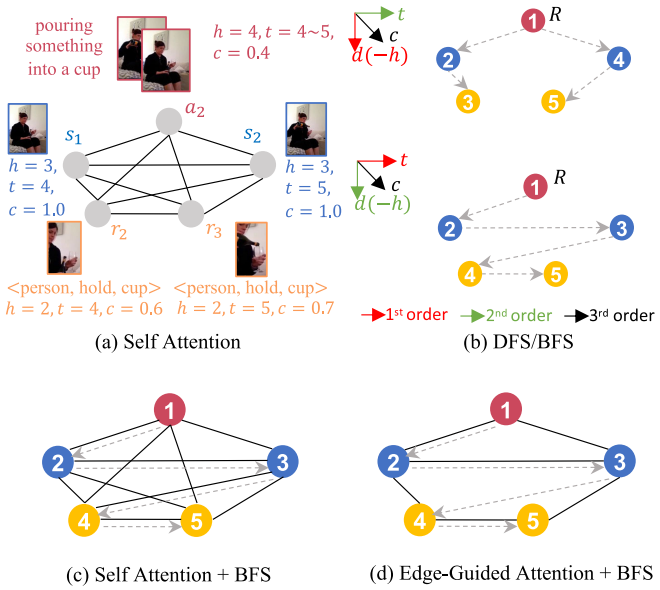


Fig. 3. The framework of the proposed hierarchical spatial-temporal transformer. The given question is embedded into word embeddings X_q and tagged with level embeddings X_H. HSTT encodes the structure prior from the event graph from two aspects: 1) generate reasoning order embeddings X_O with the improved graph search algorithm; 2) Generate the attention mask M for edge-guided attention. HSTT takes the sum of word/node embedding, level embedding, order embedding as input, and jointly encodes them based on the graph structure prior.



(a) Self Attention (b) DFS/BFS

(c) Self Attention + BFS (d) Edge-Guided Attention + BFS

Fig. 4. An example to show how our HSTT exploits the graph structure prior. (a) Vanilla transformer exploits fully connected self-attention interactions (black solid edges) to encode each token in parallel (unordered). Each token actually has different semantic hierarchies h, occurrence time t and confidence scores c. (b) Our improved DFS and BFS can provide reasonable search order (gray dashed arrows). (c) With BFS, HSTT can encode nodes in order. (d) Edge-guided attention only preserves the interactions among nodes with useful spatial-temporal context.

like "a2→s1→r2→s2→r3" while BFS gives an order like "a2→s1→s2→r2→r3". Formally, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a root node R, the traditional graph search algorithm $\psi(.)$ output the traversal order sequence $\mathbf{U} = [u_1, \ldots, u_L]$, $1 \leq u_i \leq L$, where L is the number of all accessible nodes in the graph, formulated as,

$$\psi(R, \mathcal{G}) = [u_1, \ldots, u_L]. \tag{1}$$

Due to some nodes having the same timestamp, *e.g.*, object nodes extracted in the same frame, the traversal order could be uncertain. When encountering uncertainty, the algorithm will choose the node with maximal confidence score c to let the model focus on more salient visual concepts. We apply this improved graph search algorithms on the event graph to obtain the search order sequence $\mathbf{U}$. Then a learned order embedder $\phi_o(.)$ is used to get the graph structure encoding $\mathbf{X}_O$ from the graph search result

$$\mathbf{X}_O = [\phi_o(u_1), \ldots, \phi_o(u_L)]. \tag{2}$$

At last, we add the order embedding together with the level embedding on the node embedding to get the overall input visual embeddings, $\mathbf{X}_V = [\mathbf{X}_N + \mathbf{X}_H + \mathbf{X}_O]$. Similarly, the textual embeddings are $\mathbf{X}_T = [\mathbf{X}_q + \mathbf{X}_H]$.

a) Joint encoding: To fully exploit the inter- and intra-modality interaction for more thorough context information fusion, HSTT jointly encodes the visual embeddings and textual embeddings at the same time. Namely, the cross-modal embeddings are concatenated as an overall input, and the word encoding and node encoding are computed in parallel. HSTT applies a multi-layer Transformer [49] of stacked multi-head attention building blocks as backbones, which is defined as:

$$\begin{aligned} MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= Concat(\mathbf{head}_1, \ldots, \mathbf{head}_h)\mathbf{W}_O, \\ \mathbf{head}_i &= Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \end{aligned} \tag{3}$$

where W are different parameter matrices. Specifically, there are two types of multi-head attention in HSTT: the intra-modality attention shares the same queries, keys, and values, i.e., $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{X}, \mathbf{X} \in \{\mathbf{X}_V, \mathbf{X}_T\}$ to learn the pairwise relationship between the paired sample $\langle x_i, x_j \rangle$ within the single modality. The inter-modality attention takes two groups of input embeddings $\mathbf{X}_V$ and $\mathbf{X}_T$ from different

modalities and learns the cross-modal correspondence between the paired sample $\langle x_i, x_j \rangle$ from $\mathbf{X}_V$ and $\mathbf{X}_T$, respectively.

As shown in Figure 4(a), the self-attention organizes the embeddings into a fully connected "graph" among all embeddings. All embeddings are packed together as a set of queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ simultaneously, then the attention function are applied to them:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}, \quad (4)$$

in which d_k is the dimension of the key vector. Compared to such a fully connected interaction graph, our event graph, which is shown in Figure 4(d), has a much sparse structure, where only nodes with spatial-temporal context relationships are linked together. Such structure reflects the natural spatial-temporal relationship among all visual elements.

b) Edge-guided attention: Different from the original Transformer which employs fully multi-head attention, we adopt sparse edge-guided attention to encode node embeddings, which allows for the precise aggregation of spatial-temporal context based on the constructed event graph. To achieve this, we maintain an attention mask $\mathbf{M} \in \{0, -\infty\}^{|X_V|}$ to indicate whether to allow or prevent the attention interaction among embeddings. The attention mask is generated based on the adjacency matrix $\mathbf{A} \in \{0, 1\}^{|\mathbf{X}_V|}$ of the event graph. Specifically, for each pair of nodes that have no edge linkage, *i.e.,* $A_{ij} = 0$, the interaction between them will be prevented by a mask $M_{ij} = -\infty$. The edge-guided attention function can be calculated as:

$$Attention_{mask}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}. \quad (5)$$

According to the definition of the event graphs in Section III-A, the edges in the graphs are divided into two types, i.e., $\mathcal{E} = \mathcal{E}_{intra} \cup \mathcal{E}_{inter}$. So, we will introduce the intra-level interaction and inter-level interaction separately.

4) Intra-Level Interaction: The intra-level edges $\mathcal{E}_{intra}$ indicate nodes linkage from the same level to describe their long-form temporal dependencies. As shown in Figure 5, through intra-level interaction, nodes' representations are enhanced by aggregating spatial-temporal context from the other nodes at the same level. *Spatial context*: "`Person`" and "`Bottle`" can aggregate the spatial context from each other; *Temporal context*: "`Person`" in the first frame can aggregate temporal context from other frames for dynamic modeling. The intra-level nodes are fully connected with each other, *i.e.*, the adjacency matrix $\mathbf{A}^{intra} = \mathbf{1}$. Therefore the intra-level masks on the diagonal are all zero matrices. Namely, all intra-level interactions are preserved:

$$\mathbf{M}_{ij}^{intra} = \mathbf{0}, h(i) = h(j), \quad (6)$$

where $h(.)$ denotes the level index of node.

5) Inter-Level Interaction: The inter-edges $\mathcal{E}_{inter}$ describe the spatial-temporal inclusion relationship between two adjacent levels. As shown in Figure 6, through inter-level interaction, nodes' representations are updated by aggregating the information from parent/child nodes. To encode the relation "$\langle$`Person, Hold, Bottle`$\rangle$",

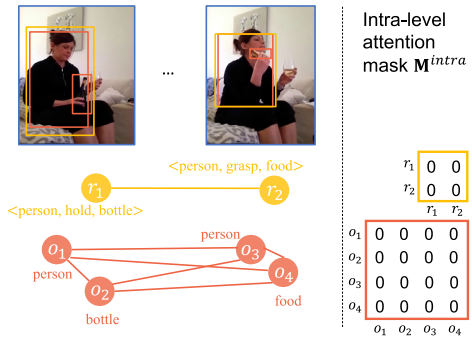


Fig. 5. An example of intra-level interaction within objects and relations and the corresponding attention mask.

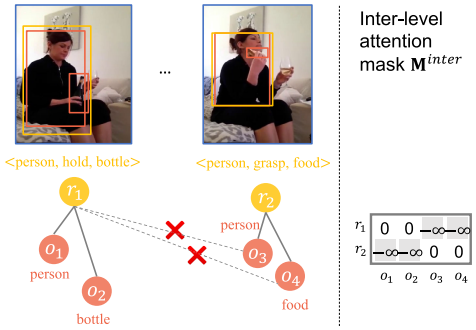


Fig. 6. Inter-level interaction. The redundant interactions between nodes without edge linkage are avoided (red crosses) by applying the inter-level attention mask $\mathbf{M}^{inter}$.

we aggregate information from its child nodes, "`Person`" and "`Bottle`". On the opposite, we may also recognize the unclear object "`Food`" by reasoning from its parent node "$\langle$`Person, Grasp, Food`$\rangle$". Notably, the redundant interaction like "$\langle$`Person, Hold, Bottle`$\rangle$" with "`Person`" and "`Food`" from different scenes should be prevented from the original self-attention in Equation 4. Therefore, we apply the mask $\mathbf{M}^{inter}$ corresponding to the adjacency matrix of inter-level edges to let the nodes interact following the event graph:

$$\mathbf{M}_{ij}^{inter} = -\infty(1 - \mathbf{A}_{ij}), h(i) \neq h(j). \quad (7)$$

Overall, adding order embedding as an input element has injected the node's position in the graph into the model. The designed edge-guided attention mechanism can further encode the graph structure information by applying precise interactions among nodes strictly following the graph edges.

a) Answer prediction: Through joint encoding, the HSTT constructs multi-level correspondence between two modalities by aligning the semantic concepts in the questions with the visual concepts in the videos from different levels. Therefore, HSTT can predict the answer by precisely grounding all the visual clues concerned in the question. Specifically, in each multi-head attention layer, the output attended embeddings will be passed into a Feed-Forward and an Add & Norm layer [49] to compute the next layer's hidden-layer representations. After L layers enoding, the output of the Transformer is a list of D-dimensional feature vectors Z which is consistent with the input embeddings. The vector $Z_{\texttt{[CLS]}}$ corresponding to the `[CLS]` is used for predicting the

probabilities over all candidate answers $\boldsymbol{y}$ through a two-layer MLP with Softmax, formulated as,

$$P(\boldsymbol{y}|\mathbf{V}, \mathbf{q}) = Softmax(\mathbf{Z}_{[\mathtt{CLS}]}\mathbf{W}_Z), \tag{8}$$

where $\mathbf{W}_Z \in \mathbb{R}^{d_z * D}$ is learnable weight. The cross-entropy loss function is applied to calculate the loss value.

IV. Experiment

A. Datasets

We make a full investigation of our proposed model using recently proposed AGQA [10] and STAR [17] datasets. They all involve various common multi-step reasoning tasks from real-world daily life video clips. AGQA is currently the largest VideoQA dataset containing 192M open-ended questions about 9.6K videos of real-world daily human activities collected in Charades [52] dataset. All questions are automatically generated. We utilize its balanced subset (Balanced) of 3.9M question-answer pairs for evaluation. It contains various question types aiming to comprehensively evaluate the model's compositional spatial-temporal reasoning ability. We also use its two additional settings, novel composition (Novel) and more reasoning steps (Step) to evaluate the model's generalization ability.

STAR is a multi-choice type VideoQA benchmark covering four types of daily-life situated reasoning questions, including present situation reasoning question type: Interaction and Sequence, and future situation reasoning question type: Prediction and Feasibility. It consists of 60K questions and 240K candidate choices about 22K trimmed situation video clips with ground-truth programs and answers for model diagnosis.

B. Implementation Details

1) Event Graph Construction: Following [53], the Faster R-CNN model [54] with a ResNeXt-101-FPN backbone [55], [56] pre-trained on the Visual Genome [57] is used as object detector. Then it is concatenated with a predefined Transformer with Total Direct Effect (TDE) analysis [48] as the relation predictor. We follow [14] to extract dumped frames from the video, then uniformly sample 10 frames in the videos and collect $M = 10$ objects and $N = 20$ relations with top confidence scores on each frame. For scene and action levels, the ResNet-50 [58] is applied to extract 2D features of each dumped frame for scene nodes. For actions, we first get the video features by a 3D-ResNet model from RGB frames of the video sampled at 8 FPS. The network's weights are initialized from a model pre-trained on Kinetics [59]. Then based on the extracted video features, BMN [60] model is trained to generate action proposals, and those with confidence scores larger than 0.3 are saved.

2) HSTT: We simply use a shallow 2-layer encoder with 8 attention heads Transformer as a backbone in our experiments, considering the inputs visual tokens already have strong semantics. The hidden state size is set to 768. For comparison, our baseline employs the Vanilla Transformer of the same architecture with the input of 2D frame features as in [38]. AdamW [61] is utilized to optimize end-to-end model training, using an initial learning rate of 5e-5 with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and use learning rate warmup over the first 10% training steps followed by linear decay to 0. Our model is trained for 15 epochs using 4 NVIDIA TITAN RTX GPUs and implemented in PyTorch [62].

TABLE I
AVERAGE RECALL (%) OF VARIOUS MODELS ON SCENE GRAPH GENERATION TASK OF ACTION GENOME DATASET FOLLOWING THE SETTING IN [14]

Model	R@20	R@50
RelDN [14]	24.45	25.63
MVSGG [63]	**32.43**	**43.13**
Transformer+TDE [48]	31.76	35.94

TABLE II
AVERAGE RECALL (%) OF VARIOUS MODELS ON ACTION LOCALIZATION TASK TESTING ON THE CHARADES DATASET

Model	R@1	R@5	R@10
BSN [64]	18.20	39.57	53.98
BMN [60]	**20.73**	**41.96**	**54.60**

C. Evaluation of the Event Graph

The correctness of the event graph directly affects the quality of video structure representation. Since the AGQA dataset is constructed based on the action annotations from Charades [52] and scene graph annotations from Action Genome [14]. Thus, we use them as benchmarks to evaluate all involved intermediate tasks, including scene graph generation and action localization to evaluate the quality of each level of our constructed event graph.

Action Genome [14] is a scene graph generation benchmark. We use it to evaluate the correctness of the object and relation levels. As shown in Table I, the scene graph predictor achieves 31.76% and 35.94% mean recall at the top 20 and 50 respectively, comparable with state-of-the-art performances [51].

Charades [52] is an action recognition benchmark. We use it to evaluate the quality of generated action proposals. As shown in Table I, the action localizer also shows promising results. In addition, we visualize some constructed event graph examples in Figure 11 at the end of the article.

D. Comparison With State-of-the-Art

1) Evaluation on the AGQA Dataset: In Table III, we compare our model with the state-of-the-art (SOTA) VideoQA works on the Balanced setting of the AGQA dataset. HCRN [19], HME [7] and PSAC [6] are all listed as baselines to benchmark the AGQA dataset in [10]. VQA-T [39] is the latest Transformer-based VideoQA framework. As shown in the last line, our HSTT model acquires the best 59.72% result on all questions. It improves the strongest SOTA method, VQA-T, by 5.27%, though the Transformer-based methods commonly show stronger modeling ability over others. HSTT also outperforms HCRN by a clear margin, which uses a frame-to-clip video modeling way. Such facts verify that through explicitly modeling the multi-granularity semantic

TABLE III

PERFORMANCE COMPARISON ON ALL KINDS OF QUESTION TYPES ON THE AGQA DATASET. WE COUNTED THE PROPORTION OF THE QUESTION NUMBER OF EACH TYPE. THE TABLE IS DIVIDED INTO FOUR PARTS: ABLATION EXPERIMENTS, SOTA COMPARISON (TRANSFORMER-BASED & OTHERS), AND LOWER AND UPPER BOUNDS. FOR ABLATION EXPERIMENTS, RESULTS FROM MODELS WITHOUT CERTAIN LEVEL VISUAL INPUTS ARE HIGHLIGHTED IN DIFFERENT COLORS IF THE QUESTIONS ARE CLOSELY RELATED TO THE CORRESPONDING LEVELS

	Question Types (Proportion)	HSTT	HSTT (w/o act.)	HSTT (w/o rel.)	HSTT (w/o obj.)	VQA-T [39]	Vanilla Transformer	HME [7]	HCRN [19]	PSAC [6]	Most Likely	Human
Reasoning	obj-rel (77.69%)	**56.40**	55.80	55.47	53.91	49.11	44.82	43.91	43.00	34.75	8.82	80.65
	obj-act (0.01%)	60.60	55.80	62.25	55.98	**80.30**	54.71	50.00	63.33	58.33	50.00	93.75
	rel-act (4.7%)	69.01	63.66	67.47	59.46	**73.45**	58.76	57.84	56.75	56.84	50.00	90.20
	obj-act (0.01%)	60.60	55.80	62.25	55.98	**80.30**	54.71	50.00	63.33	58.33	50.00	93.75
	super (24.82%)	**43.79**	43.48	42.65	41.98	38.58	34.72	41.10	37.48	30.51	10.29	81.25
	sequence (7.7%)	65.66	62.51	63.98	58.32	**71.00**	56.32	59.60	61.28	59.95	49.15	90.77
	exists (13.78%)	**77.98**	77.39	77.30	76.95	77.10	72.42	70.01	72.22	69.94	50.00	79.80
	duration (1.29%)	**48.28**	48.43	43.45	48.47	48.10	45.47	44.19	45.10	29.75	23.70	92.00
	act recog (0.14%)	20.79	16.29	19.00	11.55	**26.40**	6.95	3.23	11.21	3.78	4.72	78.00
Semant.	obj (75.12%)	**56.33**	55.79	55.46	53.69	48.62	44.05	42.48	41.55	32.79	9.38	87.97
	rel (13.18%)	72.70	70.37	70.87	69.12	**74.26**	66.90	66.10	66.71	65.51	50.00	83.58
	act (11.71%)	66.50	64.09	64.22	60.93	**69.51**	57.76	58.12	60.41	58.24	32.91	86.45
Structure	query (52.05%)	**58.24**	57.39	57.66	54.91	49.49	44.30	36.23	37.18	27.20	11.76	83.53
	compare (13.59%)	65.51	61.91	63.59	58.10	**70.16**	56.48	58.06	58.77	56.68	50.00	92.53
	choose (15.53%)	43.84	**44.12**	41.80	42.36	36.14	32.98	49.32	40.60	33.41	50.00	83.02
	logic (2.97%)	71.59	**72.58**	73.18	72.20	71.29	71.37	69.75	69.90	67.48	50.00	70.69
	verify (15.86%)	**72.78**	71.83	71.18	71.67	72.03	67.41	68.40	71.09	68.34	50.00	88.26
Overall	binary (47.95%)	**61.27**	60.09	59.64	58.37	59.83	53.41	59.77	58.11	54.19	50.00	86.65
	open (52.05%)	**58.24**	57.39	57.66	54.91	49.49	44.30	36.23	37.18	27.20	11.76	83.53
	all (100%)	**59.72**	58.68	58.61	56.57	54.45	48.67	47.74	47.42	40.40	10.35	86.02

hierarchy of the video content, our HSTT can effectively address the complex cross-level reasoning task.

Besides, AGQA uses various dimensions to evaluate the model's reasoning ability, including the reasoning types (Reasoning), the semantic types of the question's target (Semantic), the organizational structure of the question (Structure), and the forms of the answers (Overall). As shown in the table, our HSTT performs best on most question types. Specifically, it achieves significant performance gain with 8.75% at query Structure at most (HSTT *vs.* VQA-T: 58.24% *vs.* 49.49%). It is worth mentioning that our model achieves relatively higher promotion on the open-answer type questions, which are more challenging compared to binary-answer type questions with 50% accuracy purely by blind guessing. It indicates that our model is good at answering the question by grounding the specific visual clues based on the multi-level cross-modal alignment rather than turning to exploit linguistic biases [10]. Although VQA-T achieves better performance on some question types, *e.g.*, the obj-act type, the question number proportion is quite small (only 0.01%) with great model randomness. Thus, in general, we conclude that our HSTT acquires better generalization on various question types.

2) Evaluation on STAR Dataset: We also evaluate our HSTT on the STAR dataset. The graph-based method (L-GCN [25]) which combines the object-level and scene-level features and more Transformer-based methods are added as SOTA methods to be compared. The results are shown in Table IV. Although dose not pretrained with large-scale datasets like most of the SOTAs, our HSTT still achieves the best performance and obtains a clear advantage of 4.41% over the best of them on average. Among different question types, MIST [65] achieves a minor advantage on the "Interact" type. However, HSTT outperforms it by a large margin on all other types which challenges the temporal reasoning ability of models. What's more, HSTT also shows strong future reasoning capability on "Predict" and "Feasible" question types owing to its ability to analyze high-level actions from complex events.

TABLE IV

QA ACCURACY (%) OF FOUR QUESTION TYPES ON STAR DATASET

Model	Question Type				
	Interact	Sequence	Predict	Feasible	Mean
L-GCN [25]	39.01	37.97	28.81	26.98	33.26
HCRN [19]	39.10	37.97	28.81	26.98	33.28
ClipBERT [38]	39.81	43.59	32.34	31.42	36.85
SHA [43]	47.98	42.03	35.34	32.52	39.47
RESERVE [66]	44.75	42.40	38.83	36.17	40.54
Flamingo [40]	42.15	44.56	40.64	41.57	42.23
NS-SR [17]	42.61	46.26	43.44	43.88	44.08
Temp[ATP] [67]	50.63	52.87	49.36	40.61	48.37
MIST-CLIP [65]	**55.59**	54.23	54.24	44.48	51.13
VQA-T [39]	51.05	60.61	61.17	59.83	58.17
HSTT(w/o)	53.07	62.98	63.69	62.61	60.59
HSTT(rand-inter)	53.33	64.57	66.48	62.43	61.70
HSTT(rand-intra)	54.78	63.46	64.80	**63.65**	61.67
HSTT(BFS+EGA)	54.16	**65.08**	**68.30**	62.78	**62.58**

E. Model Analysis

1) Analysis on the Importance of Each Level: We first analyze what role each level of our event graph plays on different question types. In Table III, firstly, compared to Vanilla Transformer (baseline) with only scene-level features as input, our HSTT with the complete four levels of visual

TABLE V
PERFORMANCE COMPARISON ON HSTT WITH OR WITHOUT GRAPH STRUCTURE ENCODING EVALUATED ON THE AGQA BALANCED SUBSET. ONLY RESULTS WITH REPRESENTATIVE QUESTION TYPES UNDER REASONING AND OVERALL DIMENSION ARE SHOWN

	Method	HSTT (w/o)	HSTT (DFS)	HSTT (BFS)	HSTT (EGA)	HSTT (BFS+EGA)
Reasoning	obj-rel	54.89	53.95	55.88	55.99	**56.40**
	superlative	41.32	41.22	43.62	43.13	**43.79**
	sequencing	57.26	61.36	61.29	63.40	**65.66**
	duration	48.06	46.37	47.24	47.78	**48.28**
Overall	binary	57.84	60.47	59.81	60.02	**61.27**
	open	56.96	56.29	57.26	58.10	**58.24**
	all	57.38	58.29	58.48	59.02	**59.72**

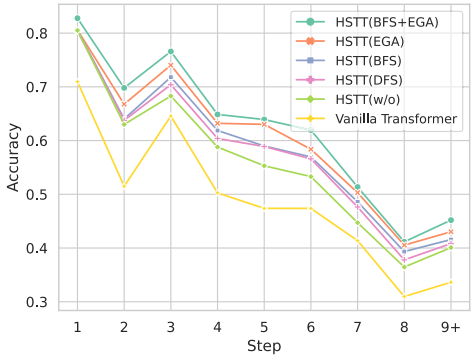


Fig. 7. QA accuracy of various models on questions with different reasoning steps. The more steps of reasoning, the greater the difficulty of the questions. HSTT models with different graph structure encoding strategies outperform HSTT(w/o) and vanilla transformer on even extremely complex questions of steps over 9.

concepts modeling achieves substantial improvement with 11.05% on the overall accuracy. Then, we try to remove all nodes from a certain level to evaluate their importance. The results show that without any level of visual concepts will all lead to a performance drop, which indicates the necessity of each semantic level in event reasoning. Specifically, the model without a certain level of visual concepts will struggle on answering the related questions (colorful cells), *e.g.*, comparing HSTT and HSTT (w/o act.), the performance drops from 20.79% to 16.29% on action recognition Reasoning types (act recog.). Meanwhile, the performances do not change obviously on irrelevant questions. But, there is an exception, that is, removing the object-level nodes causes obvious performance drops on almost all question types. This is because the objects are the basic visual elements for other visual concepts. In conclusion, by completely parsing the low-level objects to high-level actions in the complex events, our HSTT can not only deal with local detailed reasoning but also global event reasoning.

2) Analysis on the Graph Structure Encoding: In HSTT, we encode the graph structure prior by 1) encoding the nodes in the graph search order, *i.e.*, HSTT(BFS) and HSTT(DFS), and 2) using edge-guided attention (EGA) to encode the graph nodes, *i.e.*, HSTT(EGA). We make ablation studies on the effect of each of them on Balanced settings with HSTT(w/o), without using any kind of graph structure prior. As shown in Table V, both of them gain better performances than HSTT(w/o) in different types of questions. It indicates that the natural spatial-temporal context provided by the graph structure can boost the reasoning ability of HSTT. Specifically, comparing the two different search strategies, BFS performs better than DFS, which shows that temporal-order-first is a more appropriate encoding strategy for Transformer than hierarchical-order-first. Therefore we combine the BFS with the EGA in graph encoding, the performance obtains further gain from 58.48% to 59.72% overall. We also visualize the performances on various reasoning step questions in Figure 7. Models with graph structure encoding achieve better results on almost all reasoning steps even for extremely hard questions with more than 9 steps.

We additionally try to analyze our Edge-Guided Attention mechanism on the STAR dataset. Since EGA uses mask attention to control the interactions among embeddings guided by the inter- and intra- edges, we respectively apply a random mask on the inter- and intra- attention map with the same masking number to prove its effectiveness. Namely, $|\mathcal{E}|$ interactions will be blocked randomly for fairness. Randomly masking the inter-level (rand-inter) and intra-level (rand-intra) attention map all suffer performance degradation. The comparisons are shown in Table IV. Both random inter- and intra-attention mask strategies show no effect. It indicates that our EGA can effectively prevent the model from aggregating redundant even wrong spatial-temporal context information.

3) Analysis on Generalization: We then evaluate the generalization of our HSTT model. Firstly, we compare HSTT with VQA-T pre-trained on a large-scale automatically generated VideoQA dataset, HowToVQA69M [39]. From Table VI, our HSTT without any pre-training beats VQA-T in all kinds of answer forms of questions. Besides, the pretraining helps little when addressing complex multi-step compositional reasoning. On the contrary, with graph structure encoding, our HSTT can further improve the overall performance from 57.38% to 59.72%. Such results demonstrate that explicitly constructing multi-level cross-modal alignment is more effective than implicitly obtaining it through pre-training on large-scale data. And when data is limited, our method can take advantage of the good graph structure prior introduced by graph structure encoding. What's more, compared to VQA-T, our model is more light-weighted and computational efficiency with the same multi-head attention layer. When we apply a deeper Transformer, the model still has some performance improvement space (4 layers HSTT *vs.* HSTT: 60.34% *vs.* 59.72%). However, it is not necessary to adopt a deeper model to consume more computing resources.

We further evaluate the generalization of our model using novel composition (Novel) and more steps (Step) settings in the AGQA dataset. In the Novel setting, the model will meet unseen compositions of visual concepts in the testing stage from different reasoning types. And in the Step settings, the model needs to handle complex questions with more compositional steps. The performance comparison is reported in Table VII. HSTT achieves SOTA performance on open-ended questions. For the binary-answer questions, our model

TABLE VI

GENERALIZATION EVALUATION BETWEEN HSTT AND VQA-T ON AGQA BALANCED SUBSET

Method	Pretrain	Size	FLOPs	Overall		
				binary	open	all
VQA-T	-	156M	1.79T	59.83	49.49	54.45
VQA-T	How2VQA69M	156M	1.79T	60.23	50.02	54.92
HSTT(w/o)	-	**108M**	**1.72T**	57.84	56.96	57.38
HSTT	-	**108M**	**1.72T**	61.27	58.24	59.72
HSTT(4 layers)	-	122M	1.74T	**62.08**	**58.74**	**60.34**

TABLE VII

PERFORMANCE COMPARISON WITH OTHER VIDEOQA METHODS ON NOVEL COMPOSITIONS (NOVEL) AND MORE COMPOSITIONAL STEPS (STEP) SETTINGS

	Method	PSAC	HME	HCRN	VQA-T	HSTT
Novel	binary	43.00	**52.39**	43.40	46.20	47.99
	open	14.80	19.46	23.72	38.93	**40.61**
	all	32.49	40.11	36.06	43.48	**45.24**
Step	binary	35.39	**48.09**	42.46	43.82	43.55
	open	28.00	33.47	34.81	53.60	**55.01**
	all	31.13	39.70	38.00	49.37	**50.18**

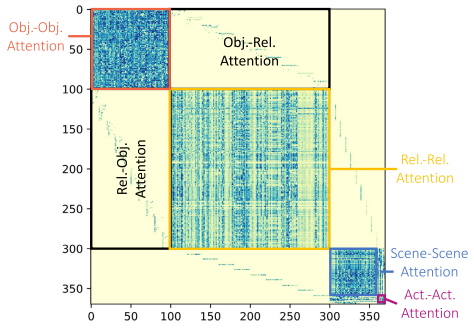


Fig. 8. Attention map among all node embeddings in the event graph of HSTT model with EGA. EGA guides the model learn dense attention among intra-level nodes (diagonal) and extremely sparse attention among inter-level nodes (off-diagonal).

meets the 4.40% and 4.54% gap between HME. However, HME has severe performance degradation on more challenging open-answer questions, which shows the limited generalization of its memory network. Our HSTT model has a stronger generalization ability on overall questions with 1.76% and 0.81% advantages on the Novel and Step setting respectively compared to the VQA-T model. In fact, the novel compositions or the complex multi-step questions can be decomposed into basic concepts from different semantic levels. Thanks to taking a structured video representation as input, HSTT can explicitly model the cross-modal hierarchical correspondence of these basic concepts, therefore generalizes well on such Novel and Step settings by compositional reasoning of all grounded visual concepts.

F. Qualitative Results

We visualize the attention map of a QA example in the AGQA dataset in Figure 8, Figure 9, and 10. The attention map is an obvious partitioned matrix with four dense intra-level attention maps on the diagonal while the inter-level attention maps are extremely sparse, with only about **1%** interactions

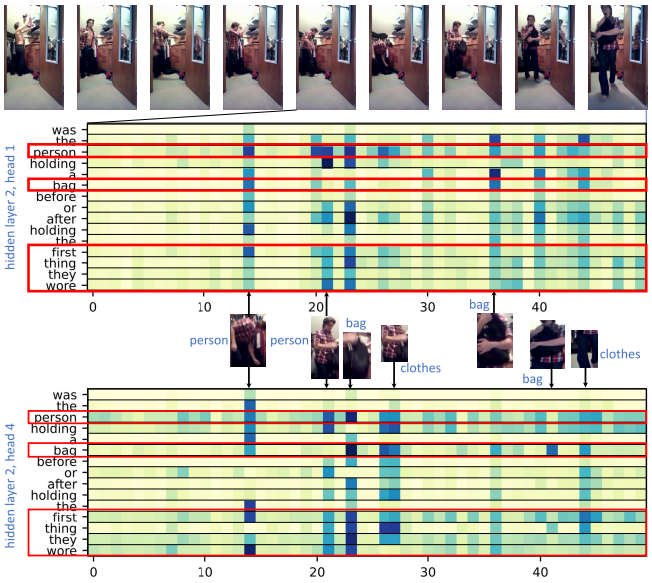


Fig. 9. The language-object part of the Co-attention map in HSTT from head 1 and head 4 respectively. The color reflects the cross-modal correlation. We only show 5 keyframes of the video for better visualization. The activated objects ("person", "bag" and "clothes") are highlighted and their corresponding visual regions are listed.

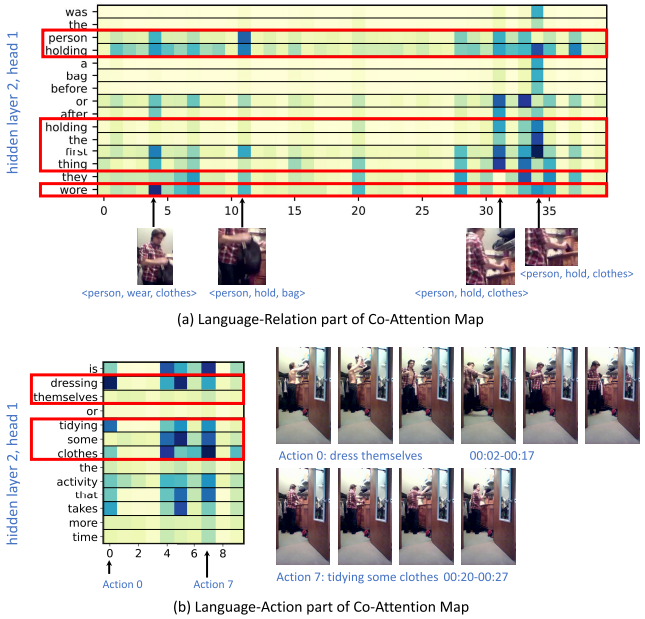


Fig. 10. The language-relation and language-action part of the Co-attention map in HSTT. In the language-relation part (a), we only show the 2 key frames of the video for better visualization. Activated concepts are highlighted. In the language-action part (b), the actions with their corresponding keyframes and timestamps are shown. "Action 0" is about "dress themselves" while "Action 7" is about "tiding some clothes".

preserved. It indicates that intra-level interaction commonly exists especially at the object level and scene level. On the contrary, the interactions across different levels are very sparse. With such sparse interaction, our HSTT can still achieve great performance thanks to the precise spatial-temporal relationship delivered by the event graph.

Figure 9 and Figure 10 show the correspondence between question words and detected objects, relations and actions

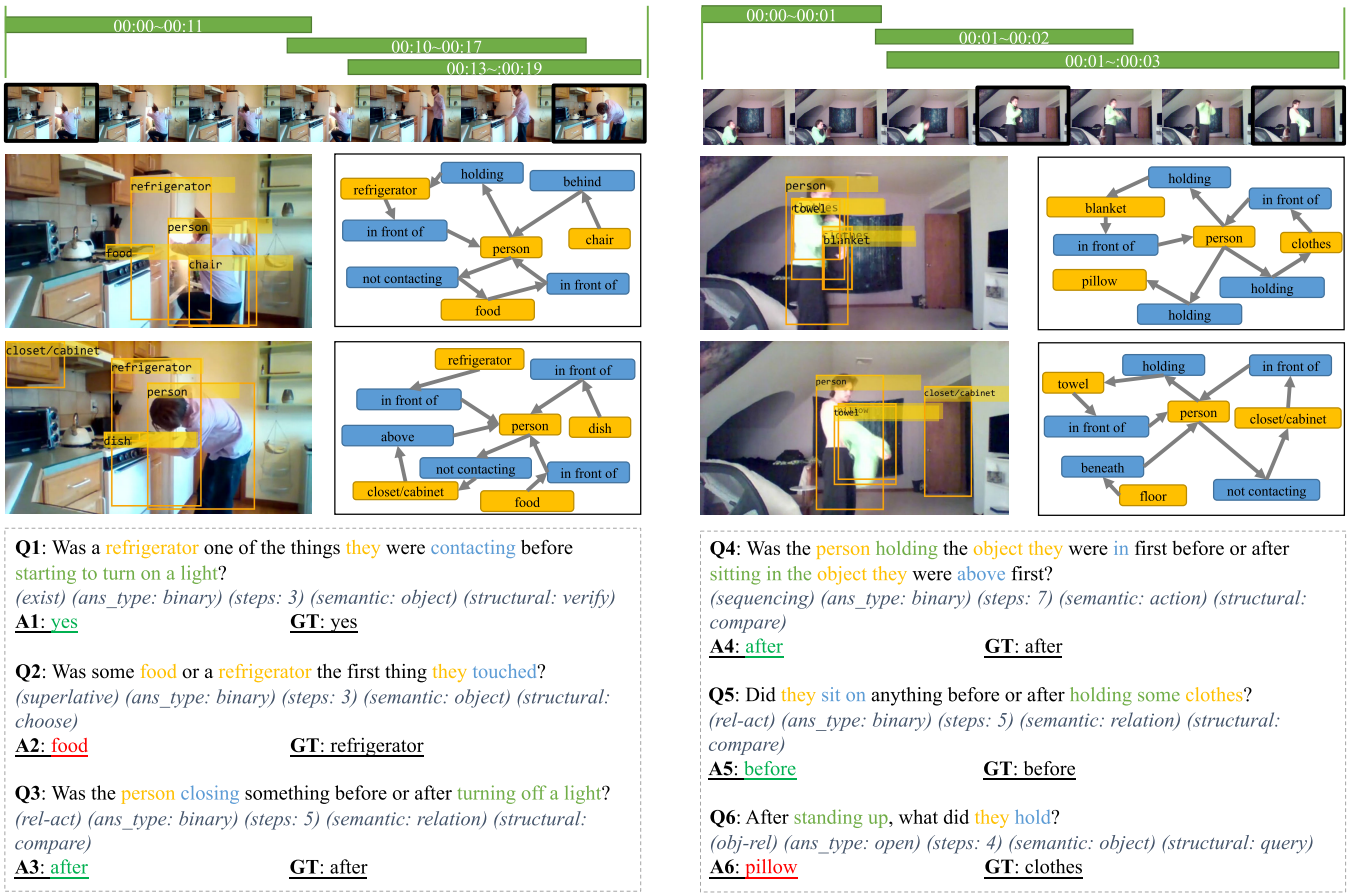


Fig. 11. Qualitative results for event graph generation and VideoQA. The green bars show the action proposals and the selected keyframes in videos are shown below the bars. Two static scene graphs related to the QA with yellow blocks for objects and blue blocks for relations (predicates) are shown. We only visualize the nodes with a top confidence score for clear display.

respectively. The darker the color of the grid is, the greater the correlation between cross-modal concepts is. In Figure 9, HSTT learns the alignment about key objects, *i.e.*, the "person", "bag" and the "clothes". Although the "clothes" is indirectly referred by "first thing they wore" in the question, it can still accurately correlate to the visual cues in the video. We randomly choose two attention maps from different heads (head 1 and head 4). Both of them can realize such alignment. In Figure 10, "<person,wear,clothes>" and "<person,hold,bags>" are aligned with the words "wore" and "person holding" in the first frame. <person,hold,clothes> are aligned with the words "holding the first thing" in the second frame. In the Language-Action part, "Action 0" and "Action 7" are aligned with the words "dressing themselves" and "tidying some clothes" in the question. Such phenomena sufficiently indicate our HSTT has effectively learned a multi-level correspondence between visual concepts and linguistic concepts. It is the key to solving the complex cross-level reasoning in the Video Question Answering task. We also make a case study on some of the videos in the AGQA dataset to give a detailed qualitative analysis. Figure 11 shows the constructed event graphs from 2 raw videos and the QA results of the proposed model. We choose different types of questions with various reasoning steps for comprehensive analysis. In each case, we display the key visual elements extracted from our event graph on the top. It can be seen that the constructed event graph can accurately represent abundant visual concepts at all levels. Besides, 3 different QA results are shown at the bottom. We highlight the cross-level semantic concepts in the questions with corresponding colors. By comparison, you can clearly find strong correspondence between the visual and semantic concepts. Therefore, our HSTT can exploit such correspondence to gather visual clues to predicate the right answers. In all 6 questions, HSTT answers 4 correctly (green answers). In **Q2**, HSTT fails to extract the relation "<person,touch,food>". In **Q6**, HSTT fails to distinguish the "pillow" and "clothes".

V. Conclusion

In this paper, we target at VideoQA task by first introducing a hierarchical graph structure representation for complex videos. Then we propose HSTT for compositional spatial-temporal reasoning guided by the event graph. The extensive experiments show some insights for the VideoQA task. 1) The structured parsing of the complex event is indispensable for the model to ground the local to global visual concepts concerned in the questions. 2) By precisely combining the spatial-temporal context of relevant visual

concepts, the model can be generalized to new concept compositional reasoning and multi-step reasoning tasks. 3) Explicitly constructing multi-granularity cross-modal correspondence is more effective than implicitly obtaining it through large-scale pretraining for addressing complex compositional reasoning tasks. Nevertheless, the current definition and formulation of the event graph is an early attempt. Future work will conduct further theoretical research to help build a more reasonable and effective event representation. Besides, dynamic reasoning with adaptive adjustment mechanism will be explored to reduce the impact of noise nodes/edges in the graphs. Overall, we hope our work can inspire more future works to address the video reasoning task from the perspective of explicit concepts grounding. It is believed in this way, we can build a more interpretative and reliable model to vigorously promote the VQA research.

REFERENCES

[1] J. P. Bigham et al., "VIZWIZ: Nearly real-time answers to visual questions," in *Proc. 23rd Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2010, pp. 333–342.

[2] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.

[4] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4971–4980.

[5] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6700–6709.

[6] X. Li et al., "Beyond RNNs: Positional self-attention with co-attention for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8658–8665.

[7] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1999–2007.

[8] J. Park, J. Lee, and K. Sohn, "Bridge to answer: Structure-aware graph interaction network for video question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2021, pp. 15526–15535.

[9] J. Wang, B.-K. Bao, and C. Xu, "DualVGR: A dual-visual graph reasoning unit for video question answering," *IEEE Trans. Multimedia*, vol. 24, pp. 3369–3380, 2022.

[10] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "AGQA: A benchmark for compositional spatio-temporal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11287–11297.

[11] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5410–5419.

[12] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 670–685.

[13] M. Rashid, H. Kjellstrom, and Y. J. Lee, "Action graphs: Weakly-supervised action localization with graph convolution networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 615–624.

[14] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10236–10247.

[15] M. Nawhal and G. Mori, "Activity graph transformer for temporal action localization," 2021, *arXiv:2101.08540*.

[16] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. R850–R852, Oct. 2004.

[17] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, "STAR: A benchmark for situated reasoning in real-world videos," in *Proc. 35th Conf. Neural Inf. Process. Syst. Track Datasets Benchmarks (Round 2)*, Dec. 2021, pp. 1–13.

[18] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6576–6585.

[19] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9972–9981.

[20] 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.

[21] P. Zeng, H. Zhang, L. Gao, J. Song, and H. T. Shen, "Video question answering with prior knowledge and object-sensitive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 5936–5948, 2022.

[22] D. Li et al., "A simple baseline for video restoration with grouped spatial–temporal shift," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9822–9832.

[23] R. Feng, Y. Gao, X. Ma, T. H. E. Tse, and H. J. Chang, "Mutual information-based temporal difference learning for human pose estimation in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17131–17141.

[24] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, and Y. Zhuang, "Multi-interaction network with object relation for video question answering," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 1193–1201.

[25] 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*, vol. 34, no. 7, 2020, pp. 11021–11028.

[26] F. Liu, J. Liu, W. Wang, and H. Lu, "HAIR: Hierarchical visual-semantic relational reasoning for video question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1698–1707.

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

[28] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE Trans. Image Process.*, vol. 30, pp. 2758–2770, 2021.

[29] L. Hoang Dang, T. Minh Le, V. Le, and T. Tran, "Hierarchical object-oriented spatio-temporal reasoning for video question answering," 2021, *arXiv:2106.13432*.

[30] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T.-S. Chua, "Video as conditional graph hierarchy for multi-granular question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2804–2812.

[31] A. Cherian, C. Hori, T. K. Marks, and J. Le Roux, "(2.5+ 1)D spatio-temporal scene graphs for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 444–453.

[32] Y. Yan et al., "Learning multi-attention context graph for group-based re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7001–7018, Jun. 2023.

[33] C. Rodriguez-Opazo, E. Marrese-Taylor, B. Fernando, H. Li, and S. Gould, "DORi: Discovering object relationships for moment localization of a natural language query in a video," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1079–1088.

[34] J. Li et al., "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3022–3031.

[35] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7464–7473.

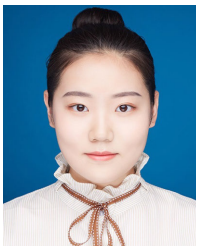
[36] L. Zhu and Y. Yang, "ActBERT: Learning global-local video-text representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8746–8755.

[37] R. Zellers et al., "MERLOT: Multimodal neural script knowledge models," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, Dec. 2021, pp. 23634–23651.

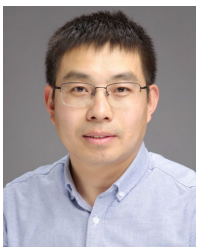
[38] J. Lei et al., "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7331–7341.

[39] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2021, pp. 1686–1697.

[40] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," 2022, *arXiv:2204.14198*.
[41] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," 2022, *arXiv:2206.03428*.
[42] C. Tan, Z. Lin, J.-F. Hu, W.-S. Zheng, and J. Lai, "Hierarchical semantic correspondence networks for video paragraph grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18973–18982.
[43] A. Urooj Khan et al., "Learning situation hyper-graphs for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14879–14889.
[44] M. Li et al., "WINNER: Weakly-supervised hIerarchical decompositioN and aligNment for spatio-tEmporal video gRounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23090–23099.
[45] V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," 2020, *arXiv:2003.00982*.
[46] S. Yao, T. Wang, and X. Wan, "Heterogeneous graph transformer for Graph-to-Sequence learning," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7464–7471.
[47] Z. Zhang, Q. Liu, Q. Hu, and C.-K. Lee, "Hierarchical graph transformer with adaptive node sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 21171–21183.
[48] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3713–3722.
[49] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[51] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial–temporal transformer for dynamic scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 16372–16382.
[52] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 510–526.
[53] K. Tang. (2020). *A Scene Graph Generation Codebase in Pytorch*. [Online]. Available: https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.
[55] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2017, pp. 1492–1500.
[57] 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, May 2017.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[59] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.
[60] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3889–3898.
[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[62] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32. Vancouver, BC, Canada, May 2019, pp. 8026–8037.
[63] L. Xu, H. Qu, J. Kuen, J. Gu, and J. Liu, "Meta spatio-temporal debiasing for video scene graph generation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 374–390.
[64] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[65] D. Gao, L. Zhou, L. Ji, L. Zhu, Y. Yang, and M. Zheng Shou, "MIST: Multi-modal iterative spatial–temporal transformer for long-form video question answering," 2022, *arXiv:2212.09522*.
[66] R. Zellers et al., "MERLOT RESERVE: Neural script knowledge through vision and language and sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16375–16387.
[67] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the 'video' in video-language understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2917–2927.



Ziyi Bai received the B.E. degree in computer science and technology from Xidian University, Xi'an, China, in 2019. She is currently pursuing the Ph.D. degree with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. Her research interests include computer vision and vision-language multimodal intelligence.



Ruiping Wang (Senior Member, IEEE) received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. From 2010 to 2012, he was a Postdoctoral Researcher with Tsinghua University. From 2010 to 2011, he was a Research Associate with the University of Maryland, College Park, MD, USA. In 2012, he joined as a Faculty Member with ICT, CAS, where he has been a Professor since 2017. His research interests include computer vision, pattern recognition, and machine learning. He is currently an Associate Editor of *Pattern Recognition* and *Neurocomputing*. He served as the Area Chair for IEEE WACV 2022–2024, ICCV 2021, CVPR 2021/2022, and ECCV 2022/2024.

Difei Gao (Member, IEEE) received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2015, and the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2022. He is currently a Postdoctoral Researcher with the National University of Singapore (NUS). His research interests include computer vision and vision-language multimodal intelligence.

Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 400 papers in refereed journals and conference proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of the ACM, IAPR, and CCF. He is also an Information Sciences Editorial Board Member of *Fundamental Research*, an Editorial Board Member of *Research*, a Senior Editor of the *Journal of Visual Communication and Image Representation*, and the Associate Editor-in-Chief of the *Chinese Journal of Computers* and *Chinese Journal of Pattern Recognition and Artificial Intelligence*. He served as an Organizing Committee Member for multiple conferences, including the General Co-Chair for FG 2013/FG 2018 and VCIP 2022, the Program Co-Chair for ICMI 2010/FG 2024, and the Area Chair/Senior PC for ICCV/CVPR/ECCV/NeurIPS/ICMI for more than ten times.