LiVLR: A Lightweight Visual-Linguistic Reasoning Framework for Video Question Answering

Jingjing Jiang, Ziyi Liu, Member, IEEE, and Nanning Zheng, Fellow, IEEE

Abstract—Video Question Answering (VideoQA), aiming to correctly answer a given question based on understanding multimodal video content, is challenging due to the richness of the video content. From the perspective of video understanding, a complete VideoQA framework needs to understand the video content at different semantic levels and flexibly integrate diverse video content to distill question-related content. To this end, we propose a Lightweight Visual-Linguistic Reasoning framework named LiVLR. Specifically, LiVLR first utilizes graph-based visual and linguistic encoders to obtain multi-grained visual and linguistic representations, respectively. Subsequently, the obtained representations are integrated with the devised Diversity-aware Visual-Linguistic Reasoning module (DaVL). DaVL distinguishes different types of representations with the learnable index embedding in graph embedding. Therefore, DaVL can flexibly adjust the importance of different representations when generating the question-related joint representation. The proposed LiVLR is lightweight and shows its performance advantage on three VideoQA benchmarks, MRSVTT-QA, KnowIT VQA, and TVQA. Extensive ablation studies demonstrate the effectiveness of the key components of LiVLR.

Index Terms—Video question answering, relational reasoning, graph convolutional network, representation integration.

I. INTRODUCTION

VIDEO Question Answering (VideoQA) is a typical multimodal understanding task that aims to correctly answer a given question based on understanding video content. Due to the richness of the video content, it is challenging to find evidence of the correct answer from the massive amount of video information. From a video understanding perspective, a complete VideoQA framework requires two crucial functions: (i) understanding of the video content at different semantic levels, (ii) flexible integration of the diverse content for distilling question-related content.

For the first function, pioneering works [1]–[3] capture the spatiotemporal information of the video and represent them with appearance and motion features. These image- and clip-level representations carry the information needed to answer the types of questions conditioned on holistic video understanding. For example, to answer question Q1 in Fig. 1, the VideoQA model requires capturing the holistic event (i.e., the two guys follow the girl into the building) described in the video stream. To answer the type of questions based on video details, such as Q2 in Fig. 1, the VideoQA model needs to identify the seat that Penny is sitting on and to capture the fine-grained relationship between the seat and Penny in one frame of the video. To this end, relational reasoning-based VideoQA methods [4]–[6] have been proposed to model relationships between visual objects. In addition to the aforementioned multi-grained visual content, some videos contain linguistic content, such as subtitles [7], [8], knowledge [9], and descriptions [1]. Analogously, the VideoQA framework also needs to properly understand the holistic and fine-grained linguistic content to answer the questions concerning the linguistic content (such as Q3 in Fig. 1) or even to support visual understanding. Therefore, to achieve a versatile VideoQA framework, it should consider all the cases listed above and flexibly react to each case.

For the second function, i.e., effective integration of the obtained diverse representations for answer prediction, the existing solutions can be roughly divided into two categories. The first, attention-based solutions, design different attention mechanisms, such as memory-enhanced attention [3], [10], spatial-temporal attention [2], [11], and cross-modality transformer [12], [13], for fusing diverse representations. The other is attention- and graph-based solutions, which adopt both attention mechanisms and graph reasoning for the representation fusion. For example, the works in references [14], [15] sequentially apply question-related attention and graph reasoning to perform diverse representations fusion. However, the attention mechanism includes many matrix multiplication operations with high-dimensional dense representations, which increases the number of model parameters and reduces computational efficiency. As shown in Fig. 2, VideoQA models using attention mechanisms to achieve the function (i) and (ii). For example, VQA-T [13] is usually more heavyweight than the model that utilizes graph neural networks like DualVGR [15]. Therefore, the graph reasoning network is one feasible solution to devise a lightweight VideoQA model.

In this paper, we propose a Lightweight Visual-Linguistic Reasoning framework, named LiVLR, which mainly consists of Visual Encoder, Linguistic Encoder, and the Diversity-aware Visual-Linguistic Reasoning module (DaVL). Firstly, the proposed framework respectively applies the graph-based Visual...
Fig. 1. Examples of the VideoQA task. Answering Q1 requires understanding the holistic event described in the video stream. Answering Q2 requires capturing the fine-grained relationship between the seat and Penny in one frame of the video. Answering Q3 requires understanding the linguistic content of the video.

Fig. 2. Comparison of the number of model parameters and accuracies on MSRVTT-QA.

Encoder and Linguistic Encoder to encode the visual and linguistic content of the video at different semantic levels and yield multi-grained visual and linguistic representations. Subsequently, the obtained multi-grained visual and linguistic representations and the question representation are passed into the Diversity-aware Visual-Linguistic Reasoning module (DaVL). In DaVL, we construct a diversity-aware graph with the multi-grained visual and linguistic representations as initial node representations. The initial node representations are first associated with the question representation using an attention block and then enhanced by the learnable index embeddings of the different representations. Facilitated by learnable embeddings, which prompt the differences between different types of representations and adjusts their importance, DaVL can flexibly react to different case of question in using the graph convolutional network to obtain a joint representation for answer prediction.

Our main contributions are summarized as follows:

- We propose a Lightweight Visual-Linguistic Reasoning framework for VideoQA, named LiVLR, which separately generates multi-grained visual and linguistic representations using graph-based Visual and Linguistic Encoders, and effectively integrates multi-grained visual and linguistic representations via a proposed representation integration method DaVL.
- We propose the Diversity-aware Visual-Linguistic Reasoning module (DaVL), a powerful and general representation integration method considering the diversity of multi-grained visual and linguistic representations.
- The proposed VideoQA framework LiVLR is lightweight and shows its performance advantage on three standard VideoQA benchmarks. Extensive ablation studies on key components of LiVLR demonstrate the effectiveness of the proposed framework.

II. RELATED WORK

A. Video Question Answering

Video Question Answering (VideoQA) aims to answer a given question concerning video content. Most current works [1]–[3], [10], [16]–[18] extract holistic visual appearance and motion features to represent video contents and design different attention mechanisms, such as question-guided attention [1], [11] and co-attention [3], [18], to integrate these features. These methods focus on the holistic understanding of the video contents, which may neglect meaningful and fine-grained video content that semantic-complicated questions concern.

To answer such semantic-complicated questions that based on a fine-grained comprehension of video content, relational reasoning-based methods [4]–[6], [14], [15], [19], [20] have been proposed. More specifically, Jin et al. [4] proposed a multimodal and multi-level interaction network to capture relations between objects. Jiang et al. [14] developed a heterogeneous graph alignment network to integrate both inter- and intra-modal relations for cross-modal reasoning. Le et al. [19] explored more robust multimodal interaction by constructing a general-purpose neural reasoning unit. Huang et al. [5] proposed a location-aware graph convolutional network to model the location and relation among objects explicitly. Wang et al. [15] adopted a stacked dual-visual graph reasoning unit, DualVGR, to iteratively model the rich relationship between video clips. Seo et al. [6] utilized graph convolutional networks to compute the relationships among objects in both appearance and motion modules. Park et al. [21] constructed graphs for both video and question and encoded question-to-visual relationships and visual-to-visual relationships.

In addition, to better understand the video content, which is usually multimodal data containing visual and linguistic information, extra linguistic information, such as subtitles [7], [8], captions [22], [23], and knowledge [9], [24], have been introduced to VideoQA tasks. Our work aims to handle generalized VideoQA tasks that consider both visual and linguistic information in a more practical manner than the abovementioned visual-specific VideoQA tasks. The Visual Encoder in LiVLR is closely related to relational reasoning-based methods.

B. Relational Reasoning

Relational reasoning is extensively exploited in vision-and-language tasks [25]–[30] to model intra-modal or cross-modal relations among visual/semantic elements. Recent approaches
to relational reasoning can broadly be classified as graph-based [31]–[33], neuro-symbolic-based [34]–[36], and others [19], [37], [38]. Graph-based methods have been shown to be powerful in performing visual and semantic reasoning and have become prevailing methods for performing relational reasoning in vision-and-language tasks, typically by considering the explicit relation, which can be directly denoted by a relation triplet, and the implicit relation, which is not predefined. Specifically, Li et al. [31] encoded explicit semantic and spatial relations as well as implicit fully-connected relations between objects by a graph-based attention network. Huang et al. [5] focused on the location and relations among object interaction and proposed a location-aware graph convolutional network (GCN) to model implicit relations between objects. A more recent study [39] considered explicit relations in visual, semantic, and knowledge modalities and proposed a modality-aware heterogeneous GCN to encode them. In addition, the graph learner module [40] conditioned on the context of a given question was developed to better uncover and exploit these implicit relations between objects. Similarly, we also employ attention-based GCNs to reveal explicit and implicit relations underlying visual and linguistic content.

C. Graph Neural Network

Graph Neural Network (GNN) [41] is a type of deep learning models that handles graph-structure data, utilizing the graph structure to aggregate node information from neighborhoods. The power of GNNs in modeling relationships between graph nodes has led to its widespread use in various tasks and applications, such as graph classification [42], [43], cross-modality retrieval [44], [45], and visual question answering [5], [46]. In recent years, many GNNs have been proposed. These existing GNNs can be broadly grouped into two categories: spectral-based GNNs [47]–[50] and spatial-based GNNs [51]–[53]. Specifically, spectral-based GNNs first transform graphs to the spectral domain by graph Fourier transform, then perform the convolution operator defined in the spectral domain, and finally transform the encoded graphs back to spatial domain with the inverse graph Fourier transform. For example, Defferrard et al. [50] utilized the Chebyshev expansion of the graph Laplacian matrix to define spectral filters, which alleviates the computational complexity of the eigen-decomposition.

Spatial-based GNNs directly define convolution operators on the graph based on the graph topology. For example, GAT [52] is a typical spatial-based GNN. It incorporates the attention mechanism into the propagation step, which assigns different weights to neighbors to alleviate node noise. In this paper, we adopt self-attention based GCNs in Visual and Linguistic Encoders and utilize multi-head attention based GCN in DaVL.

III. PRELIMINARY

In this section, we first state the problem definition and the inputs of LiVLR, i.e., the pre-extracted visual and linguistic features. After that, we introduce the attention-based GCN, which is the basic block that will be utilized in LiVLR.

Problem Definition: The goal of the VideoQA task is to infer the answer $\hat{a}$ to a given question $q$ conditioned on understanding the video content. The answer $\hat{a}$ can be found in an answer set $A$, a predefined set of possible answers for an open-ended (OE) question setting or a list of answer candidates for a multiple-choice (MC) question setting. Since the proposed VideoQA framework LiVLR independently encodes the visual $(V)$ and linguistic $(L)$ content of the given video, we formulate the VideoQA task as

$$\hat{a} = \arg \max_{a \in A} p_\theta(a \mid q, V, L),$$

where $\theta$ denotes trainable model parameters.

Visual Features: For each video clip, we sample $N_f$ frames of images and represent these sampled images in three forms: (i) image-level appearance features: $[o_{11}, \ldots, o_{N_f}] \in \mathbb{R}^{N_f \times 2048}$, (ii) object-level region features: $[O_1^0, \ldots, O_{N_f}^0] \in \mathbb{R}^{N_f \times N_v \times 2048}$, where $N_v$ is the number of objects in an image, and $O_f^0 = [o_{11}^f, \ldots, o_{N_v}^f] \in \mathbb{R}^{N_v \times 2048}$ $1 \leq f \leq N_f$, and (iii) phrase-level class-attribute features: $[C_1^0, \ldots, C_{N_f}^0] \in \mathbb{R}^{N_f \times N_c \times 768}$, where $C_f^0 = [c_{11}^f, \ldots, c_{N_c}^f] \in \mathbb{R}^{N_c \times 768}$ is the class-attribute phrase embedding of the $N_v$ objects in the $f$-th image.

Linguistic Features: For a given question, we first extract the token-level features $Q^0 = [q_{11}^T, \ldots, q_{N_o}^T] \in \mathbb{R}^{N_o \times 768}$ of the question, where $N_o$ is the number of tokens in the sentence. In addition to the given question, there are $N_s$ linguistic description sentences corresponding to the video-question pair. Therefore, we extract linguistic features $[L_1^0, \ldots, L_{N_f}^0] \in \mathbb{R}^{N_f \times N_v \times 768}$ for all $N_s$ sentences, where $L_f^0 = [l_{11}^f, \ldots, l_{N_v}^f] \in \mathbb{R}^{N_v \times 768}$, $1 \leq s \leq N_s$. That is, the linguistic inputs of LiVLR are the token-level question features and the linguistic description features.

Attention-Based GCN: For the given graph $G = (V, E)$, $V = \{v_1, \ldots, v_{N_v}\}$ is a set of nodes, $N_v = |V|$ denotes the number of nodes, $E$ is a set of edges, and $V = \{v_1, \ldots, v_{N_v}\} \in \mathbb{R}^{N_v \times d}$ is the initial set of node representations. Then, the formula for updating node $i$ in the $l$-th GCN layer can be expressed as:

$$v_i^{(l+1)} = \text{ReLU}(\alpha_{i,i,j} \cdot W(l)v_j^{(l)}),$$

where $\text{ReLU}$ denotes the ReLU activation function, and $N_i$ denotes the neighborhoods of node $i$, which is determined by $E$. $W(l) \in \mathbb{R}^{d \times d}$ is a transformation matrix for node $i$ in the $l$-th GCN layer. The attention coefficient $\alpha_{i,j}$ is defined as

$$\alpha_{i,j} = \frac{\exp((W_qv_i)^T \cdot W_kv_j)}{\sum_{v_j \in N_i} \exp((W_qv_i)^T \cdot W_kv_j)},$$

where $W_q \in \mathbb{R}^{d \times d}$ and $W_k \in \mathbb{R}^{d \times d}$ are learnable transformation matrices.

IV. METHOD

Fig. 3 shows the details of the proposed LiVLR, consisting of Visual Encoder, Linguistic Encoder, Question Encoder, the proposed DaVL, and the Answer Prediction module. In a nutshell, LiVLR first exploits graph-based encoders to encode fine-grained visual inputs and linguistic inputs and yields multi-grained visual and linguistic representations. Subsequently,
LiVLIR integrates the obtained multi-grained visual and linguistic representations via the Diversity-aware Visual-Linguistic Reasoning module (DaVL). LiVLIR separately encodes the holistic and fine-grained visual content and yields multi-grained visual representations. Linguistic Encoder uniformly encodes holistic and fine-grained linguistic content and generates multi-grained linguistic representations. DaVL aims to integrate multi-grained visual and linguistic representations in a diversity-aware manner and outputs the joint question-related representation for answer prediction.

### A. Visual Encoder

The Visual Encoder separately encodes the holistic and fine-grained visual contents of the video. For each video clip, the image-level appearance features \([a_1, \ldots, a_N]\) are mapped into a \(d\)-dimensional \((d-D)\) space by a fully-connected (FC) layer to obtain the holistic visual representation \(X_{V, p} = [a_1, \ldots, a_N] \in \mathbb{R}^{N \times d}\). To obtain the fine-grained visual representation, the visual relationships between the objects are encoded into graphs. The specific process is as follows.

1. **Object Relation Graph Construction**: Intuitively, visual relations imply spatial relationships reflecting the relative locations of objects and semantic relationships depicting semantic coherence of visual concepts. As shown in Fig. 3, for the \(t\)-th sampled image, we construct a spatial graph \(G_{sp} = (V, E_{sp}, R_{sp})\) and a semantic graph \(G_{se} = (V, E_{se})\) using objects as graph nodes. \(R_{sp}\) is a set of edges types.

2. **Object Relation Graph Embedding**: To better represent object-spatial and object-semantic relations, we improve the node embeddings of \(G_{sp}\) by concatenating the position features and improve node embeddings of \(G_{se}\) by concatenating the class-attribute features. Specifically, we denote the position feature of object \(i\) in the \(t\)-th image as \(p_i = [p_x, p_y, p_w, p_h, p_w, p_h]^T\), where \(p_x, p_y\) is the top-left coordinate of the bounding box, and \(p_w\) and \(p_h\) are the weight and height of the box, respectively. Given the feature \(o_i^0\) of object \(i\) in the \(t\)-th image, the \(i\)-th node embedding in \(G_{sp}\) can be initialized by

\[
v_{sp,i}^{(0)} = W_{sp}^0([W_o o_i^0 + b_o, W_p p_i + b_p]),
\]

where \(W_o \in \mathbb{R}^{2048 \times d}\) and \(b_o \in \mathbb{R}^d\) map the extracted object-level feature \(o_i^0\) into a \(d\)-D representation, \(W_p \in \mathbb{R}^{6 \times d}\) and \(b_p \in \mathbb{R}^d\) map the position feature \(p_i\) to a \(d\)-D representation, and \(W_{sp}^0 \in \mathbb{R}^{2d \times d}\) transforms the concatenated feature into a \(d\)-D representation space.

Given the class-attribute feature \(c_i^0\) of object \(i\) in the \(t\)-th image, we initialize the \(i\)-th node embedding in \(G_{se}\) as:

\[
v_{se,i}^{(0)} = W_{se}^0([W_o o_i^0 + b_o, W_c c_i^0 + b_c]),
\]

where \(W_o\) and \(b_o\) are the same as (4). \(W_c \in \mathbb{R}^{768 \times d}\) and \(b_c \in \mathbb{R}^d\) transform \(c_i^0\) into a \(d\)-D representation space, and \(W_{se}^0 \in \mathbb{R}^{2d \times d}\) transforms the concatenated feature into a \(d\)-D representation space.

3. **Object Relation Encoding**: To encode the information of known edge types \(R_{sp}\) into \(G_{sp}\), we modify the information aggregation between node \(i\) and a neighboring node \(j\) in (2) to

\[
v_{sp,i}^{(l)} = W_{sp}^{(l)} v_{sp,j}^{(l)} \oplus b_{sp}^{(l)} (r_{i,j}),
\]

where, \(W_{sp}^{(l)} \in \mathbb{R}^{d \times d}\) and \(b_{sp}^{(l)} \in \mathbb{R}^{d}\) are the node transformation matrix and the learnable vector of edge types in the \(l\)-th GCN layer, respectively. \(r_{i,j} \in \mathbb{R}_{sp}\) indicates the edge type between node \(i\) and \(j\), which are classified into 11 categories according to recent studies [25, 31]. \(b_{sp}^{(l)} (r_{i,j})\) denotes the \(r_{i,j}\)-th element of \(b_{sp}^{(l)}\).

For \(G_{se}\), considering the complexity of detecting relation triplets between objects in a video, we do not explicitly define the semantic relation but have the framework implicitly learn the relation with a graph learner [40]. More concretely, the adjacency matrix \(A_{se}\) can be obtained using the initial node embeddings

\[
A_{se} = (W_1 v_{se}^{(0)T})(W_2 v_{se}^{(0)})^T,
\]

where \(W_1, W_2 \in \mathbb{R}^{d \times d}\) are the transformation matrices of the node embeddings. Additionally, we adopt a ranking strategy to

---

**Fig. 3.** Overall architecture of LiVLIR. The framework mainly consists of Visual Encoder, Linguistic Encoder, and the Diversity-aware Visual-Linguistic Reasoning module (DaVL). Visual Encoder separately encodes the holistic and fine-grained visual content and yields multi-grained visual representations. Linguistic Encoder uniformly encodes holistic and fine-grained linguistic content and generates multi-grained linguistic representations. DaVL aims to integrate multi-grained visual and linguistic representations in a diversity-aware manner and outputs the joint question-related representation for answer prediction.
constrain the graph sparsity; that is, we only retain the top $N_n$ maximum values for each row of $A_{sc}$. After determining the adjacency matrix, the node in $G_{sc}$ can be updated by (2).

So far, for the $f$-th image in one video clip, we can obtain two node-level representations: $V_{sp} = [v_{sp,1}, \ldots, v_{sp,N_r}] \in \mathbb{R}^{N_r \times d}$ from $G_{sp}$ and $V_{se} = [v_{se,1}, \ldots, v_{se,N_r}] \in \mathbb{R}^{N_r \times d}$ from $G_{sc}$. Then, we apply graph pooling on the two node-level representations to generate the graph-level embeddings $\bar{v}_{sp,f}, \bar{v}_{se,f} \in \mathbb{R}^d$, and stack $N_f$ graph-level representations to obtain the fine-grained visual representation $X_{V,l} = [x_1, \ldots, x_{N_f}] \in \mathbb{R}^{N_f \times d}$, where $x_f = \bar{v}_{sp,f} + \bar{v}_{se,f}$.

### B. Linguistic Encoder

For a given video-question pair, there are $N_s$ linguistic sentences. We construct a semantic role graph for each sentence. The nodes of the semantic role graph include the sentence itself, describing an holistic event, and the linguistic components in the sentence reflecting fine-grained semantic coherence. Therefore, the Linguistic Encoder, which has a similar network architecture to Visual Encoder, uniformly encodes the holistic and fine-grained linguistic contents. The specific process is as follows:

1) **Semantic Role Graph Construction:** To construct the semantic role graph for the $s$-th ($1 \leq s \leq N_s$) linguistic sentence, as shown in Fig. 3, we first adopt an off-the-shelf SRL toolkit [54] to obtain predicates, arguments, and roles of arguments corresponding to the predicates in the sentence. With the sentence itself and $N_r$ semantic roles, inspired by previous works [55], [56], we construct the semantic role graph $G_{sr} = (V, E, T_{sr})$, where $|V| = N_r + 1$, $T_{sr}$ is a set of node types, and $G_{sr}$ is a directed hierarchical graph. More specifically, the $s$-th sentence itself serves as a global event node. Predicates and arguments are deemed local action nodes and entity nodes, respectively. Each action node is directly connected to an event node, while an entity node is connected with different action nodes according to the semantic role type related to the action node.

2) **Semantic Role Graph Embedding:** For the $G_{sr}$ of the $s$-th sentence, we initialize the global event node with a sentence-level embedding $e \in \mathbb{R}^d$. To obtain the sentence-level embedding, we first use an FC layer to transform the token-level feature $L_0 \in \mathbb{R}^{N_s \times d}$ of the $s$-th sentence into a $d$-D representation space ($L_s \in \mathbb{R}^{N_s \times d}$). Then, we apply a one-layer bidirectional long-short term memory (BiLSTM) [57] on $L_s$:

$$I = [\text{Bi-LSTM}(\bar{L}_s; \bar{\theta}_1); \text{Bi-LSTM}(\bar{L}_s; \bar{\theta}_2)],$$

where $\bar{\theta}_1$ ($\bar{\theta}_2$) are the forward (reverse) learned parameters and $[; :]$ means the concatenation operation. Each action/entity node is initialized with a token-level feature generated by a nonlinear projection:

$$v_{sr,i}^{(0)} = W_0^{sr} l_{i+1}^{0}, \quad 2 \leq i \leq N_r + 1,$$

where $l_{i+1}^{0}$ is the token-level feature corresponding to the predicate/argument feature of the $i$-th node in $L_0$, $W_0^{sr} \in \mathbb{R}^{768 \times d}$ maps the feature into a $d$-D representation space.

The semantic role itself implies underlying relationships between the local action node and entity node. To introduce the semantic role types into $G_{sr}$, we enhance the $i$-th local node in the $l$-th layer with a role embedding, which can be expressed as

$$\bar{v}_{sr,i}^{(l)} = v_{sr,i}^{(l)} \odot W_{sr}^{(l)} [t_{sr,i}; ;] , \quad 2 \leq i \leq N_r + 1,$$

where $\odot$ is element-wise multiplication, $W_{sr}^{(l)} \in \mathbb{R}^{N_r \times d}$ is a learnable role embedding matrix, $t_{sr,i} \in T_{sr} = \{1, \ldots, N_r\}$ is the semantic role type of node $i$, and $W_{sr}^{(l)} [t_{sr,i}; ;]$ denotes the $t_{sr,i}$-th row of $W_{sr}^{(l)}$.

3) **Semantic Relation Encoding:** We employ the attention-based GCN to encode the contextual semantic correlations in $G_{sr}$. Specifically, we first adopt the attention mechanism described in (3) to characterize semantic relations of different hierarchical nodes. Subsequently, the $i$-th node is updated by the update formula in (2).

After encoding $G_{sr}$, we can obtain the event node representation $v_{sr,1}$, which is the generated holistic linguistic representation for the $s$-th sentence. We stack $N_s$ event node representations to get the holistic linguistic representations $X_{L,g} \in \mathbb{R}^{N_s \times d}$. To obtain the fine-grained linguistic representations of $N_s$ sentences, we first apply an average graph pooling on the action and entity node embeddings to get the fine-grained linguistic representation $\tilde{v}_{sr,s} \in \mathbb{R}^d$ for the $s$-th sentence and then stack $N_s$ such pooled representations as the fine-grained linguistic representations $X_{V,l} \in \mathbb{R}^{N_s \times d}$.

### C. Diversity-Aware Visual-Linguistic Reasoning Module

To better fuse the multi-grained visual and linguistic representations for answer prediction, we consider the diversity of the representations and aligning the visual and linguistic representations at different semantic levels (e.g., sentence $\leftrightarrow$ image, semantic roles $\leftrightarrow$ object instances). Therefore, we propose constructing a heterogeneous graph with diversity-aware nodes and utilize a GCN module to further encode and capture relationships between them.

1) **Diversity-Aware Graph Construction:** To integrate the obtained multi-grained visual and linguistic representations in Sec. IV-A and Sec. IV-B in a diversity-aware manner, we construct an undirected heterogeneous graph $G_{da}$. Graph $G_{da}$ consists of four types of nodes: $N_f$ image-level nodes, $N_j$ object-level nodes, $N_s$ sentence-level nodes, and $N_r$ semantic role-level nodes.

2) **Diversity-Aware Graph Embedding:** Since the obtained representations $\{X_{V,g}, X_{V,l}, X_{L,g}, X_{L,l}\}$ are high-level semantic but question agnostic, we first use an attention block to associate the video content with the given question and distill the question-related representations. Specifically, we respectively apply a multi-head attention block [58] on the four representations, which can be expressed as:

$$Q_{att}(X, Q) = \sum_{n=1}^{N_h} W_h \sigma \left( \frac{W_h^0 X (W_h^0 Q)^T}{\sqrt{d_k/N_h}} \right) W_h^1 Q,$$

where $X \in \{X_{V,g}, X_{V,l}, X_{L,g}, X_{L,l}\}$, $Q \in \mathbb{R}^{N_s \times d}$ is the token-level question embedding generated by a nonlinear projection that maps $Q_0$ into a $d$-D representation space, $\cup$ denotes the concatenation operation in (11), $N_h$ is the number of heads in the attention module.
of heads, $\sigma$ indicates the softmax operation, and $d_k$ is a scaling factor. $W_q^h$, $W_k^h$, $W_v^h \in \mathbb{R}^{d/N_h \times d}$ and $W_x \in \mathbb{R}^{d \times d/N_h}$ are learned parameters.

The resulting question-related representations are the initial node representations of $\mathcal{G}_{da}$, i.e., $V^0 \in \mathbb{R}^{(2N_t + 2N_s) \times d}$. In addition, to inject the diversity-aware information of multi-grained visual and linguistic representations in representation integration, we use the index embeddings of different types of representations to enhance the initial node representations. The index embeddings are learnable and can be used to dynamically adjust the importance of different types of nodes. For the $i$-th node in the 0-th layer, this improvement process can be expressed as:

$$\tilde{\mathbf{v}}^{(0)}_{da,i} = \mathbf{v}^{(0)}_{da,i} \odot W^{(0)}_{da}(g_i, \cdot),$$  \hspace{1cm} (12)

where $W^{(0)}_{da} \in \mathbb{R}^{4 \times d}$ is a learnable transformation matrix of the index embedding, $g_i \in \{1, \ldots, 4\}$ denotes the index of $\{X_{v,g}, X_{v,t}, X_{L,g}, X_{L,t}\}$, and $W^{(0)}_{da}[g_i, \cdot]$ denotes the $g_i$-th row of $W^{(0)}_{da}$.

3) Diversity-Aware Graph Encoding: There are shallow correlations among the multi-grained and multi-source nodes in $\mathcal{G}_{da}$, such as the temporal correlations among visual nodes and the semantic consistency between visual and linguistic nodes. To delineate these correlations, we apply a vanilla GCN to update the representation of node $i$ in $\mathcal{G}_{da}$:

$$\mathbf{v}^{(l+1)}_{da,i} = \text{ReLU} \left( \mathbf{v}^{(l)}_{da,i} + \sum_{e_{da,j} \in \mathcal{N}_i} W^{(l)}_{da} \tilde{\mathbf{v}}^{(l)}_{da,j} \right),$$  \hspace{1cm} (13)

where $\mathcal{N}_i$ is the neighborhood of node $i$, which is defined by a sparse adjacency matrix learned by (7), and $W^{(l)}_{da} \in \mathbb{R}^{d \times d}$ are transformation matrix of node embeddings.

After effectively encoding these multi-grained and multi-source representations, we use a granularity- and source-sensitive graph reasoning network (i.e., DaVL), we perform average graph pooling on the node embeddings in $\mathcal{G}_{da}$ to obtain the joint representation $\hat{x} = \tilde{\mathbf{v}}_{da} \in \mathbb{R}^d$ for answer prediction.

D. Question Encoder

To further encode the contextual content of the pre-extracted token-level question embedding, we apply a one-layer BiLSTM [57] on the token-level question embedding $\mathbf{Q}$ to obtain the final sentence-level question representation $\hat{\mathbf{q}} \in \mathbb{R}^d$:

$$\hat{\mathbf{q}} = [\text{Bi-LSTM}(\tilde{\mathbf{Q}}; \theta_q^h); \text{Bi-LSTM}(\tilde{\mathbf{Q}}; \theta_q^r)],$$  \hspace{1cm} (14)

where $\tilde{\mathbf{Q}}$ and $\hat{\mathbf{Q}}$ are the forward and reverse hidden states respectively, $\theta_q^h$ and $\theta_q^r$ are learned parameters, and $[\cdot; \cdot]$ represents the concatenation operation.

E. Answer Prediction

1) Open-Ended: In the open-ended question setting, one correct answer is chosen from a predefined answer set $A$, which can be regarded as a multi-label classification problem trained with a cross-entropy loss function. Therefore, we feed the final joint representation $\hat{x}$ and the final question representation $\hat{\mathbf{q}}$ into a classifier with two FC layers ($M_{cl}$) to compute label probabilities:

$$y_o = M_{cl}([\hat{x}; \hat{\mathbf{q}}]), y_o \in \mathbb{R}^{|A|}.$$  \hspace{1cm} (15)

2) Multiple-Choice: In the multiple-choice question setting, one correct answer is chosen from $N_c$ candidates. In this case, we first generate the answer embedding $\hat{e}_k$ of the $k$-th candidate using a one-layer BiLSTM such as that in (14). Then, $\hat{x}$, $\hat{\mathbf{q}}$ and $\hat{e}_k$ are fed into a classifier with a linear regression ($M_{reg}$) to output the $k$-th answer score:

$$s_k = M_{reg}([\hat{x}; \hat{\mathbf{q}}; \hat{e}_k]), 1 \leq k \leq N_c,$$  \hspace{1cm} (16)

where the score of the correct candidate is the positive score $s_p$, and the remaining scores are negative ($s_1, \ldots, s_{N_c-1}$). During training, we utilize the summed pairwise hinge loss $\sum_{k=1}^{N_c-1} \max(0, 1 - (s_p - s_k))$ between the positive score and each negative score to train our model.

V. EXPERIMENTS

A. Experimental Settings

1) Evaluation Datasets: We evaluate the proposed LiVLR framework on three VideoQA benchmarks, MSRVTT-QA [1], KnowIT VQA [24], and TVQA [7]. MSRVTT-QA provides captions related to video content, KnowIT VQA provides subtitles (sub) and highly structured knowledge (know), and TVQA provides subtitles. These annotated captions, subtitles, and knowledge serve as the extra linguistic inputs and generate multi-grained linguistic representations. Table I summarizes the statistics of the experimental datasets. Specifically, MSRVTT-QA consists of 10K videos and 243,680 question-answer pairs. The question setting is open-ended, and the size of the pre-defined answer set is 1000. There are five question types: What, Who, How, When, and Where. KnowIT VQA is a small-scale multiple-choice VideoQA dataset comprised of 12,087 video clips and 24,282 question-answer pairs. It provides four candidate answers for each question. There are four question types: Visual (Vis.), Textual (Text.), Temporal (Temp.), and Knowledge (Know.). TVQA is a large-scale multiple-choice VideoQA dataset, consisting of 21,793 video clips from six TV shows, and 152,545 question-answer pairs. It provides five candidate answers for each question.

2) Feature Extraction Details: To obtain the inputs of LiVLR (i.e., the visual and linguistic features), we utilize ResNet-101 [64] pre-trained on ImageNet [70] to extract the holistic image appearance features for all experimental datasets, and the bottom-up attention Faster R-CNN [71] pre-trained on Visual Genome [72] to detect objects and corresponding class-attributes in each sampled image. More specifically, for MSRVTT-QA, we sample 64 frames at an equal interval from each video clip and 12 sentences as the linguistic sentences for each video clip from the caption annotations provided in [73]. The number of detected objects in each sampled image is 10. For KnowIT VQA, the

\[1^\mathrm{Online} \text{. Available: https://knowit-vqa.github.io/}\]

\[2^\mathrm{Online} \text{. Available: https://tvqa.cs.uncc.edu/}\]

\[3^\mathrm{Online} \text{. Available: https://github.com/peteanderson80/bottom-up-attention}\]
For LiVLR, we set the standard in the group (to 512, and the number of layer d utilize extra-linguistic in-
by the proposed DaVL. In-
For MSRVTT-QA, we
in Sec. IV-B is 16,
, 2
) use an attention mechanism
in (7) is set to 5. The number of heads 3
vs. (8) adopt large-scale video-language pre-
). Moreover, compared with
in (11) is re-
et al. (21], DualVGR [15], HGA [14], MASN [6],
, MiNOR [4], HME [10], GRA [1], Co-mem [3],
ST-VQA [2], VQA-T [13], CoMVT [69], ClipBERT [12], and
SSML [68]. It is worth noting that VQA-T, CoMVT, Clip-
BERT, and SSML (1) adopt large-scale video-language pre-
training to enhance the downstream VideoQA task. Generally,
the performance of pretraining-based methods is better than that of
methods that are not pretrained. ST-VQA, Co-mem, GRA,
HME, MiNOR, and HCR (2) use an attention mechanism
to achieve cross-modal representation interaction and fusion.
MASN, HGA, DualVGR, and Park et al. (ding174) adopt graph
neural networks (GNNs). Specifically, MASN adopts GNNs
to encode visual representations. HGA sequentially applies an
attention mechanism and GNNs for representation fusion.
DualVGR and Park et al. are most similar to the proposed LiVLR
utilizing GNNs for both representation encoding and fusion.

Table II summarizes the comparisons with the aforementioned
methods on MSRVTT-QA. Since none of the compared
methods listed in Table II (2, 3) utilize extra-linguistic
information in addition to the given question, for a fairer
comparison, we mainly compare them to LiVLR-V (i.e., #2 in
Table VII), which only integrates the obtained multi-grained vi-
sual representations (X,v,g X,v,l) by the proposed DaVL.
Integrating only multi-grained visual representations, our method
(LiVLR-V) outperforms the best in the same group (40.6% vs.
36.9%). After using the extra multi-grained linguistic represen-
tations, LiVLR-V (3) is further improved by 18.8% in the group (2).
Moreover, compared with the best pretraining-based method, VQA-T, the performance
of our LiVLR-V is also comparable.

B. Comparisons With State-of-the-Art Methods

We compare LiVLR with state-of-the-art methods on the open-ended (MSRVTT-QA) and two multiple-choice (KnowIT
VQA and TVQA) datasets.

1) Comparisons on MSRVTT-QA: For MSRVTT-QA, we
compare the proposed LiVLR with recent methods including
Park et al. [21], DualVGR [15], HGA [14], MASN [6],
HCR [19], MiNOR [4], HME [10], GRA [1], Co-mem [3],
ST-VQA [2], VQA-T [13], CoMVT [69], ClipBERT [12], and

Table I

| Dataset       | #Question | #Video Clip | #Sentence | LType | QType |
|---------------|-----------|-------------|-----------|-------|-------|
|               | Train | Val | Test | Train | Val | Test | Train | Val | Test | Ns | No | Nl | OE | MC | Year |
| MSRVTT-QA     | 138,581 | 12,278 | 72,824 | 6,513 | 497 | 2,990 | 78,156 | 3,964 | 35,880 | 64 | 10 | 12 | caption | OE | 2016 |
| KnowIT-VQA    | 19,569 | 2,352 | 2,361 | 9,731 | 1,178 | 1,178 | 19,569 | 2,352 | 2,361 | 32 | 12 | 12/1 | sub/know | MC | 2020 |
| TVQA          | 122,039 | 15,253 | 7,623 | 17,435 | 2,179 | 1,089 | 17,435 | 2,179 | 1,089 | 32 | 12 | 16 | Subtile | MC | 2018 |

For TVQA, the number of sampled images and detected
objects are 32 and 12, respectively. We utilize 16 subtitles as the

number of sampled images and detected objects are 32 and 12,
respectively. We use original subtitles and the provided knowl-
edge condensed from subtitles of the video as the linguistic sen-
tences. For TVQA, the number of sampled images and detected
objects are 32 and 12, respectively. We utilize 16 subtitles as the
linguistic sentences.

3) Implementation Details: For LiVLR, we set the standard
feature dimensionality d to 512, and the number of layer l of
GCN to 1. The number of semantic roles Ns in Sec. IV-B is 16,
and the number of remaining maximum values Nh in matrix
Ain in (7) is set to 5. The number of heads Nh in (11) is re-
spectively set to 16, 8, and 16 for MSRVTT-QA, KnowIT-VQA,
and TVQA. We implement LiVLR on two NVIDIA GeForce
GTX 2080Ti GPUs and utilize the AdamW optimizer with
an initial learning rate of 8e-5 and a batch size of 256
and TVQA. We implement LiVLR on two NVIDIA GeForce

GTX 2080Ti GPUs and utilize the AdamW optimizer [74]
with an initial learning rate of 8e-5 and a batch size of 256
for 80 epochs. The code is available at https://github.com/
jingjing12110/LiVLR-VideoQA.

Table II

| # | Method | Video Representation | #Param | PT | Accuracy (%) |
|---|--------|----------------------|--------|----|--------------|
| 1 | SSML [68] | ResNeXt-101 | 113.5M | ✓ | 35.1 |
| 2 | ClipBERT [12] | ResNet-50 | 25.4M | ✓ | 37.4 |
| 3 | CoMVT [69] | Faster R-CNN | - | ✓ | 39.5 |
| 4 | VQA-T [13] | Faster R-CNN | 156.5M | - | 41.5 |
| 5 | ST-VQA [2] | ResNet-152 | - | ✓ | 39.0M |
| 6 | Co-mem [3] | ResNet-152 | 24.5 | ✓ | 39.9 |
| 7 | GRA [1] | VGG16 | - | ✓ | 39.9 |
| 8 | HME [10] | VGG16 | - | ✓ | 43.8 |
| 9 | MiNOR [4] | Mask R-CNN | 24.5 | ✓ | 43.8 |
| 10 | HCR [19] | Faster R-CNN | - | ✓ | 46.7 |
| 11 | MASN [6] | Faster R-CNN | 25.4M | - | 35.2 |
| 12 | HGA [14] | Faster R-CNN | 25.4M | - | 35.2 |
| 13 | DualVGR [15] | Faster R-CNN | 25.4M | - | 35.2 |
| 14 | Park et al. [21] | Faster R-CNN | 25.4M | - | 35.2 |
| 15 | LiVLR-V | Faster R-CNN | 25.4M | - | 35.2 |
| 16 | LiVLR-V | Faster R-CNN | 25.4M | - | 35.2 |
| 17 | LiVLR-V | Faster R-CNN | 25.4M | - | 35.2 |
| 18 | LiVLR-V | Faster R-CNN | 25.4M | - | 35.2 |
| 19 | LiVLR-V | Faster R-CNN | 25.4M | - | 35.2 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
2) Comparisons on KnowIT VQA: We next compare LiVLR with the latest reported results on KnowIT VQA (four different settings of ROCK [24] and TVQA [7]) and ROLL [9]. Specifically, ROCK adopts four different techniques to describe the visual contents of video frames: (a) image, image-level features extracted using ResNet50 [64]. (b) concept, bag-of-words representations of the objects and their attributes obtained using a detector [71]. (c) facial, bag-of-faces representations of main characters in the clip detected with a face detector [76]. (d) caption, representations of sentences describing the visual content of the frames and sentences obtained using the method in reference [77]. ROLL generates unsupervised video scene descriptions (des.) as the visual input. Our LiVLR utilizes image-level appearance features and object-level region features (I + O) as the visual input. To obtain multi-grained linguistic representation, we respectively exploit the provided subtitles (Ns=12) and knowledge (Ns=1) as the original inputs of our Linguistic Encoder.

The results are shown in Table III. Overall, the proposed LiVLR outperforms the previous methods by a large margin. Particularly, LiVLR outperforms ROCK, the second-best performing method, by approximately 4%. Comparing the two cases of LiVLR (using subtitles/knowledge as the linguistic input), they achieve similar overall accuracy. However, using knowledge as the linguistic input performs poorer than using subtitles in answering textual-based questions. The possible reason is that LiVLR can only obtain one pair of multi-grained linguistic representations from the provided knowledge sentence for one image-question pair, which is far less than the number of obtained multi-grained visual representations, causing the visual information to dominate the representation integration process and weaken the effect of linguistic information.

3) Comparisons on TVQA: For TVQA, we compare the proposed LiVLR with recent methods, including Multi-Stream [7], MSAN [20], PAMN [75], and STAGE [8]. All utilize both visual and linguistic representations for VideoQA. The results are shown in Table IV. Compared with the two types of methods (using or not using timestamp annotations), the proposed LiVLR achieves state-of-the-art performance without using timestamp annotations.

C. Ablation Studies

We conduct ablation studies on MSRVTT-QA and KnowIT-VQA to demonstrate the effectiveness of key components in the proposed LiVLR. For KnowIT-VQA, we use the knowledge provided by the KnowIT-VQA benchmark to obtain multi-grained linguistic inputs for LiVLR in all ablation studies.

1) Effectiveness of the Proposed RI Method (DaVL): In our LiVLR, DaVL is designed to better integrate multi-grained visual and linguistic representations. To evaluate its effectiveness, we compare DaVL with three alternative methods for representation integration (RI) on the above two benchmarks: ◼RI-GCN: using a vanilla GCN to integrate the obtained question-related multi-grained visual and linguistic representations \(\{X_{V,g}, X_{V,l}, X_{L,g}, X_{L,l}\}\). RI-GCN is the most similar method to our proposed DaVL. However, RI-GCN does not encode the diversity-aware information for graph \(G_{da}\). ◼RI-AT: integrating \(\{X_{V,g}, X_{V,l}, X_{L,g}, X_{L,l}\}\) using the co-attention operation similar to the recent work on HGA [14]. ◼RI-Concat: integrating the obtained representations \(\{X_{V,g}, X_{V,l}, X_{L,g}, X_{L,l}\}\) with a vector concatenation operation.

The results are shown in Table V. The large performance gap with respect to the three alternative RI methods suggests the effectiveness of our proposed DaVL. Furthermore, compared with the best alternative RI method (RI-GCN), our proposed DaVL shows the improvements of 3.28% (59.44 vs. 56.16) and 3.31% (77.10 vs. 73.79) on MSRVTT-QA and KnowIT-VQA, respectively. The performance gains on the two datasets demonstrate that encoding diversity-aware information is important for the integration of multi-grained visual and linguistic representations.

2) Effectiveness of DaVL for Multi-Grained Representations: To demonstrate that our proposed DaVL is also effective in integrating multi-grained representations derived from a single source, we conduct the comparisons in Table VII (I). ◼#1 vs. #2: using RI-GCN and our DaVL respectively to integrate multi-grained visual representations \(\{X_{V,g}, X_{V,l}\}\). ◼#3 vs. #4: using RI-GCN and our DaVL respectively to integrate multi-grained visual representations \(\{X_{V,g}, X_{V,l}\}\).
linguistic representations ($X_{V,g}$, $X_{L,l}$). The results in Table VII (I) suggest that DaVL is also effective in integrating single-source multi-grained representations.

3) Effectiveness of DaVL for Cross-Modal Representations: To evaluate the effectiveness of our DaVL in integrating single-granularity cross-modal representations, we conduct the comparisons in Table VII (II). $\triangleright$ #5 vs. #6: using RI-GCN and our DaVL respectively to integrate holistic cross-modal representations ($X_{V,g}$, $X_{L,l}$). $\triangleright$ #7 vs. #8: using RI-GCN and our DaVL respectively to integrate fine-grained visual and linguistic representations ($X_{V,l}$, $X_{L,l}$). The results in Table VII (II) illustrate that the proposed DaVL is also effective in integrating single-granularity cross-modal representations.

4) Impact of Multi-Grained Visual and Linguistic Representations: The proposed LI-LVR encodes visual and linguistic content by Visual and Linguistic Encoders with similar architectures. This guarantees to some extent that the obtained holistic (fine-grained) representations from different modalities are at the same semantic level (sentence $\leftrightarrow$ image, semantic roles $\leftrightarrow$ object instances). To analyze the impact of the multi-grained visual and linguistic representations, we first consider the two comparisons: $\triangleright$ Table VII #9 vs. #1 vs. #3 and $\triangleright$ Table VII #10 vs. #2 vs. #4. From the results in the table, we observe that the performance when considering multi-grained visual and linguistic representations (#9/#10) is markedly better than that when using single visual (#1/#2) or linguistic (#3/#4) representations. Secondly, we consider another pair of comparisons: $\triangleright$ Table VII #9 vs. #5 vs. #7 and $\triangleright$ Table VII #10 vs. #6 vs. #8. Analogously, we find that the performance when considering multi-grained visual and linguistic representations (#9/#10) is markedly better than that when using single holistic (#5/#6) or fine-grained (#7/#8) representations. Finally, to further illustrate the superiority of the multi-grained visual and linguistic representations, especially the fine-grained visual and linguistic representations, we conduct the following experiment: $\triangleright$ replacing the GCN in Visual and Linguistic Encoders with a two-layer FC network to obtain fine-grained visual representations. The results on MSRVTT-QA and KnowIT-VQA are shown in Table VIII, which suggests that obtaining fine-grained visual and linguistic representations that encode relationships between visual objects or linguistic components is crucial.

5) Hyperparameters: To conduct more detailed parameter analyses, we consider the key hyperparameter $N_h$ in (11), which may directly affect the effectiveness of the proposed RI method DaVL. Specifically, the question-related attention block ($Q_{att}$) is employed to associate the multi-grained visual and linguistic representations ($X_{V,g}$, $X_{V,l}$, $X_{L,g}$, $X_{L,l}$) to the question-related representation ($Q$). This is important for distilling the question-related information from multi-grained visual and linguistic information. Specifically, we consider the following settings: $N_h = 1, 4, 8, 16, 32$. From the experimental results in Fig. 4, we observe that compared with the overall performance improvement, the performance fluctuation of LiLVR using different $N_h$ is slight, demonstrating that our method is robust to the hyperparameter $N_h$.

D. Qualitative Results

1) Qualitative Examples: To qualitatively evaluate the effectiveness of the proposed representation integration method (DaVL), we visualize some prediction examples on MSRVTT-QA [1] and KnowIT-VQA [24].

![Fig. 4. Ablation studies on the number of heads $N_h$ in Eq.(11) on MSRVTT-QA [1] and KnowIT-VQA [24].](image-url)
Fig. 5. Visualization examples of the comparison between two representation integration (RI) method on MSRVTT-QA [1]. DaVL is our proposed RI method and RI-GCN is the previous graph-based RI method. The incorrect and correct answers are highlighted in red and green, respectively.

Fig. 6. Two failure cases on MSRVTT-QA [1]. (a) A failure case caused by multi-grained linguistic representations in DaVL. The incorrect and correct answers are respectively highlighted in red and green. Using RI-GCN answers Q2 incorrectly. In Fig. 5(b), although using RI-GCN and DaVL correctly answer the Q3 related to the visual content, using RI-GCN can not answer the Q4 related to the linguistic content. The two groups of comparisons between RI-GCN and DaVL demonstrate the effectiveness of the learnable index embeddings in graph-based representation integration, and the embeddings, to some extent, adaptively choose the needed representations for the specific question. 

2) Failure Cases Analysis: Fig. 6 shows two failure cases on MSRVTT-QA. In Fig. 6(a), considering only the multi-grained linguistic representations in DaVL, our proposed LiVLR instead answers the question incorrectly. The case suggests that the learnable index embeddings may not be sufficient for DaVL to select the needed visual representations and ignore the irrelevant linguistic representations when answering questions only related to visual content. In Fig. 6(b), using graph-based RI methods (ie, RI-GCN and DaVL) to integrate multi-grained visual and linguistic representations, the VideoQA model answers the question incorrectly, while the model answers the question correctly when using the other two simple RI methods. The case shows that graph-based RI methods sometimes may lose the discriminability between nodes (ie, different types of representations) when answering the semantic-complicated question, which require a jointly understanding of both visual and linguistic content and represents the inherent limitation in using graph-based representation integration methods.

VI. CONCLUSION

In this paper, we propose a Lightweight Visual-Linguistic Reasoning framework (LiVLR), which mainly consists of Visual Encoder, Linguistic Encoder, and the proposed Diversity-aware Visual-Linguistic Reasoning module (DaVL). Specifically, LiVLR first adopts the Visual and Linguistic Encoders to obtain
multi-grained visual and linguistic representations and then utilizes DaVL to integrate the obtained representations and yield a joint representation for answer prediction. Extensive ablation studies are conducted to explore the performance contributions of the crucial components of LiVLR. The proposed LiVLR is lightweight and shows its superiority on an open-ended and a multiple-choice VideoQA datasets. In the future, we aim to explore a new representation integration method that can more flexibly selects the needed representations according to the given question.

REFERENCES

[1] D. Xu et al., “Video question answering via gradually refined attention over appearance and motion,” in Proc. 25th ACM Int. Conf. Multimedia, 2017, pp. 1645–1653.
[2] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, “TGIIF-QA: Toward spatio-temporal reasoning in visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1359–1367.
[3] J. Gao, R. Ge, K. Chen, and R. Nevatia, “Motion-appearance co-memory networks for video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6576–6585.
[4] W. Jin et al., “Multi-interaction network with object relation for video question answering,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 1193–1201.
[5] D. Huang et al., “Location-aware graph convolutional networks for video question answering,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11021–11028.
[6] A. Seo, G.-C. Kang, J. Park, and B.-T. Zhang, “Attend what you need: Motion-appearance synergetic networks for video question answering,” in Proc. Assoc. Comput. Linguistics, 2021, pp. 6167–6177.
[7] J. Lei, L. Yu, M. Bansal, and T. Berg, “TVQA: Localized, compositional video question answering,” in Proc. Joint Conf. Empirical Methods Natu- ral Lang. Process., 2018, pp. 1369–1379.
[8] J. Lei, L. Yu, T. L. Berg, and M. Bansal, “TVQA: Spatio-temporal grounding for video question answering,” in Proc. Assoc. Comput. Linguistics, 2019, pp. 8211–8225.
[9] N. Garcia and Y. Nakashima, “Knowledge-based video question answering with unsupervised scene descriptions,” 2020, arXiv:2007.08751.
[10] C. Fan et al., “Heterogeneous memory enhanced multimodal attention model for video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1999–2008.
[11] J. Jiang, Z. Chen, H. Lin, X. Zhao, and Y. Gao, “Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11101–11108.
[12] 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., 2021, pp. 7327–7337.
[13] A. Yang, A. Miech, J. Siclic, I. Laptev, and C. Schmid, “Just ask: Learning to answer questions from millions of narrated videos,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 1666–1677.
[14] P. Jiang and Y. Han, “Reasoning with heterogeneous graph alignment for video question answering,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11109–11116.
[15] J. Wang, B. Bao, and C. Xu, “DualVGR: A dual-visual graph reasoning unit for video question answering,” IEEE Trans. Multimedia, early access, Jul. 16, 2021, doi: 10.1109/TMM.2021.3097171.
[16] T. Yang, Z.-J. Zha, H. Xie, M. Wang, and H. Zhang, “Question-aware tube-switch network for video question answering,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 1184–1192.
[17] X. Li et al., “Learnable aggregating net with diversity learning for video question answering,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 1166–1174.
[18] Z.-J. Zha, J. Liu, T. Yang, and Y. Zhang, “Spatiotemporal-textual-co- attention network for video question answering,” ACM Trans. Multim. Comput. Commun. Appl., vol. 15, no. 2s, pp. 1–18, 2019.
[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., 2020, pp. 9972–9981.
[20] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, “Modality shifting attention network for multi-modal video question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10103–10112.
Jiang et al.: LIVLR: A LIGHTWEIGHT VISUAL-LINGUISTIC REASONING FRAMEWORK FOR VIDEO QUESTION ANSWERING

[49] O. Rippel, J. Snoek, and R. P. Adams, “Spectral representations for convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 2449–2457.

[50] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 3848–3852.

[51] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, arXiv:1609.02907.

[52] P. Velickovic et al., “Graph attention networks,” 2017, arXiv:1710.10903.

[53] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1025–1035.

[54] P. Shi and J. Lin, “Simple BERT models for relation extraction and semantic role labeling,” 2019, arXiv:1904.05255.

[55] D. Marcheggiani and I. Titov, “Encoding sentences with graph convolutional networks for semantic role labeling,” 2017, arXiv:1703.04826.

[56] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, “Fine-grained video-text retrieval with hierarchical graph reasoning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10635–10644.

[57] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[58] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[59] K. Harra, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6546–6555.

[60] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 305–321.

[61] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4489–4497.

[62] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4724–4733.

[63] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[64] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[65] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[66] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Toward real-time object detection with region proposal networks,” Proc. Adv. Neural Inf. Process. Syst., vol. 28, pp. 91–99, 2015.

[67] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2900–2908.

[68] E. Amran, R. Ben-Ari, D. Rotman, and A. Bronstein, “Noise estimation using density estimation for self-supervised multimodal learning,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 6644–6652.

[69] P. H. Seo, A. Nagrani, and C. Schmid, “Look before you speak: Visually contextualized utterances,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16872–16882.

[70] J. Deng et al., “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 248–255.

[71] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6077–6086.

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

[73] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5288–5296.

[74] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, arXiv:1711.05101.

[75] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo, “Progressive attention memory network for movie story question answering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8329–8338.

[76] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in Proc. BMVC. Brit. Mach. Vis. Assoc., 2015, pp. 41.1–41.12.

[77] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.

Jingjing Jiang received the B.S. degree in mechanical engineering in 2017 from Xi’an Jiaotong University, Shaanxi, China, where she is currently working toward the Ph.D. degree with the Institute of Artificial Intelligence and Robotics. Her research interests include vision-and-language and deep learning.

Ziyi Liu (Member, IEEE) received the B.S. degree in control science and engineering in 2015 from Xi’an Jiaotong University, Shaanxi, China, where he is currently working toward the Ph.D. degree with the Institute of Artificial Intelligence and Robotics. From 2019 to 2020, he was a Visiting Ph.D. Student with the State University of New York at Buffalo, Albany, NY, USA. His research interests include computer vision and machine learning.

Nanning Zheng (Fellow, IEEE) received the graduation degree from the Department of Electrical Engineering, the M.S. degree in information and control engineering from Xi’an Jiaotong University, Shaanxi, China, in 1975 and 1981, respectively, and the Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. He is the Founder of the Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, where he is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics. His research interests include computer vision, pattern recognition, autonomous vehicle, and brain-inspired computing. He became a Member of the Chinese Academy of Engineering in 1999, and a Council Member of the International Association for Pattern Recognition. He is currently the President of Chinese Association of Automation and Vice President of Chinese Society for Cognitive Science.