Video Graph Transformer for Video Question Answering

Junbin Xiao[1,2,3], Pan Zhou[1], Tat-Seng Chua[2,3], and Shuicheng Yan[1]

[1] Sea AI Lab
[2] Sea-NExT Joint Lab, Singapore
[3] Department of Computer Science, National University of Singapore
junbin@comp.nus.edu.sg, zhoupan@sea.com, dcscts@nus.edu.sg, yansc@sea.com

Abstract. This paper proposes a Video Graph Transformer (VGT) model for Video Quetion Answering (VideoQA). VGT's uniqueness are two-fold: 1) it designs a dynamic graph transformer module which encodes video by explicitly capturing the visual objects, their relations, and dynamics for complex spatio-temporal reasoning; and 2) it exploits disentangled video and text Transformers for relevance comparison between the video and text to perform QA, instead of entangled cross-modal Transformer for answer classification. Vision-text communication is done by additional cross-modal interaction modules. With more reasonable video encoding and QA solution, we show that VGT can achieve much better performances on VideoQA tasks that challenge dynamic relation reasoning than prior arts in the pretraining-free scenario. Its performances even surpass those models that are pretrained with millions of external data. We further show that VGT can also benefit a lot from self-supervised cross-modal pretraining, yet with orders of magnitude smaller data. These results clearly demonstrate the effectiveness and superiority of VGT, and reveal its potential for more data-efficient pretraining. With comprehensive analyses and some heuristic observations, we hope that VGT can promote VQA research beyond coarse recognition/description towards fine-grained relation reasoning in realistic videos. Our code is available at https://github.com/sail-sg/VGT.

Keywords: Dynamic Visual Graph, Transformer, VideoQA

1 Introduction

Since the 1960s, the very beginning of Artificial Intelligence (AI), long efforts and steady progresses have been made towards machine systems that can demonstrate their understanding of the dynamic visual world by responding to humans' natural language queries in the context of videos which directly reflect our physical surroundings. In particular, since 2019 [11], we have been witnessing a drastic advancement in such multi-disciplinary AI where computer vision, natural language processing as well as knowledge reasoning are coordinated for accurate decision making. This advancement stems, in part from the success of *multi-modal pretraining* on web-scale vision-text data [8,21,31,34,38,44,52,53,54,63], and in

part from the unified deep neural network that can well model both vision and natural language data, *i.e.*, *transformer* [55]. As a typical multi-disciplinary AI task, Video Question Answering (VideoQA) has benefited a lot from these developments which helps to propel the field steadily forward over the use of purely conventional techniques [14,16,20,23,28,60,71].

Despite the excitement, we find that the advances made by such *transformer*-style models mostly lie in answering questions that demand the holistic recognition or description of video contents [30,48,62,63,64,68,72]. The problem of answering questions that challenge real-world visual relation reasoning, especially the causal and temporal relations that feature video dynamics [20,59], is largely under-explored. Cross-modal pretraining seems promising [29,67,70]. Yet, it requires the handling of prohibitively large-scale *video*-text data [15,70], or otherwise the performances are still inferior to the state-of-the-art (SoTA) conventional techniques [29,47,67]. In this work, we reveal two major reasons accounting for the failure: 1) **Video encoders are overly simplistic.** Current video encoders are either 2D neural networks (CNNs [18,45] or Transformers [13]) operated over sparse frames or 3D neural networks [5,37,61] operated over short video segments. Such networks encode the videos holistically, but fail to explicitly model the fine-grained details, *i.e.*, spatio-temporal interactions between visual objects. Consequently, the resulting VideoQA models are weak in reasoning and require large-scale video data for learning to compensate for such weak forms of input. 2) **Formulation of VideoQA problem is sub-optimal**. Often, in multi-choice QA, the video, question, and each candidate answer are appended (or fused) into one holistic token sequence and fed to a cross-modal Transformer to gain a global representation for answer classification [72,29]. Such a global representation is weak in disambiguating the candidate answers, because the video and question portions are the same and large, which may overwhelm the short answer and dominate the overall representation. In open-ended QA (popularly formulated as a multi-class classification problem [62]), answers are treated as class indexes and their word semantics (which are helpful for QA.) are ignored. The insufficient information modelling exacerbates the data-hungry issue and leads to sub-optimal performance as well.

To improve visual relation reasoning and also reduce the data demands for video question answering, we propose the Video Graph Transformer (VGT) model. VGT addresses the aforementioned problems and advances over previous *transformer*-style VideoQA models mainly in two aspects: 1) For video encoder, it designs a dynamic graph transformer module which explicitly captures the objects and relations as well as their dynamics to improve visual reasoning in dynamic scenario. 2) For problem formulation, it exploit *separate* vision and text transformers to encode video and text respectively for similarity (or relevance) comparison instead of using a single cross-modal transformer to fuse the vision and text information for answer classification. Vision-text communication is done by additional cross-modal interaction modules. Through more sufficient video information modelling and more reasonable QA problem solution, we show that VGT can achieve much better performances on benchmarks featuring dynamic relation reasoning than previous arts including those pretrained on million-scale

vision-text data. Such strong performance comes even without using external data to pretrain. When pretraining VGT with a small amount of data, we can observe further and non-trivial performance improvements. The results clearly demonstrate VGT's effectiveness and superiority in visual reasoning, as well as its potential for more data-efficient[4] video-language pretraining.

To summarize our contributions: 1) We propose Video Graph Transformer (VGT) that advances VideoQA from shallow description to in-depth reason. 2) We design a dynamic graph transformer module which shows strength for visual reasoning. The module is task-agnostic and can be easily applied to other video-language tasks. 3) We achieve SoTA results on NExT-QA [59] and TGIF-QA [20] that task visual reasoning of dynamic visual contents. Also, our structured video representation gives a promise for data-efficient video-language pretraining.

2   Related Work

Conventional Techniques for VideoQA. Prior to the success of Transformer for vision-language tasks, various techniques, e.g., cross-modal attention [20,33,22], motion-appearance memory [16,14,36], and graph neural networks [23,35,41], have been proposed to model informative videos contents for answering questions. Yet, most of them leverage frame- or clip-level video representations as information source. Recently, graphs constructed over object-level representations [19,36,47,60] have demonstrated superior performance, especially on benchmarks that emphasize visual relation reasoning [20,49,50,59]. However, these graph methods either construct monolithic graphs that do not disambiguate between relations in 1) space and time, 2) local and global scopes [19,57], or build static graphs at frame-level without explicitly capturing the temporal dynamics [36,42,60]. The monolithic graph is cumbersome to long videos where multiple objects interact in space-time. Besides, the static graphs may lead to incorrect relations (e.g., `hug` *vs.* `fight`) or fail to capture dynamic relations (e.g., `take away`). In this work, we model video as a local-to-global dynamic visual graph, and design graph transformer module to explicitly model the objects, their relations, and dynamics, for exploiting object and relations in adjacent frames to calibrate the spurious relations obtained at static frame-level. Importantly, we also integrate strong language models and explore cross-modal pretraining techniques to learn the structured video representations in a self-supervised manner.

Transformer for VideoQA. Pioneer works [32,48,63,64,72] learn generalizable representations from HowTo100M [40] by either applying various proxy tasks [72], or curating more tailored-made supervisions (e.g., future utterance [48] and QA pairs [64]) for VideoQA. However, they focus on answering questions that demand the holistic recognition [62] or shallow description [68], and their performances on visual relation reasoning [20,59] remains unknown. Furthermore, recent works [3,70] reveal that these models may suffer from performance lose on open-domain questions due to the heavy noise [1,39] and limited data scope of HowTo100M. Recent efforts tend to use open-domain vision-text

---

[4] The model demands on less training data to achieve good performance.

data for end-to-end learning. ClipBERT [29] takes advantage of image-caption data [7,27] for pretraining, but it only has limited performance improvement on temporal reasoning tasks [20], as the temporal relations are hard to learn from static images. In addition, ClipBERT relies on human annotated descriptions which are expensive to annotate and hard to scale up. More recent works [15,70] collect million-scale user-generated (vastly abundant on the Web) vision-text data [3,51,70] for pretraining, but suffers from huge computational cost to train on such large-scale datasets. Two latest works [6,12] reveal the potential of Transformers for learning on the target datasets (relatively small scale). While promising, they either target at revealing the single-frame bias of benchmark datasets by using image-text pretrained features (e.g. from CLIP [44]), or only demonstrate the model's effectiveness on synthesized data [65]. Overall, the *poor-dynamic-reasoning* and *data-hungry* problems in existing *transformer*-style video-language models largely motivate this work. To alleviate these problems, we explicitly model the objects and relations for dynamic visual reasoning and incorporate structure priors (or relational inductive bias [4]) into transformer architectures to reduce the demand on data.

Graph Transformer. The connection between graph neural networks and Transformer has earned increasing attention [56,66,69]. Nonetheless, the major advancements are made in modelling natural graph data (e.g. social connections) by either incorporating graph expertise (e.g., node degrees) into self-attention block of Transformer [66], or designing *transformer*-style convolution blocks to fuse information from heterogeneous graphs [69]. A recent work [17] combines graphs and Transformers for video dialogues. Yet, it simply applies global transformer over pooled graph representations built from static frames and does not explicitly encode object and relation dynamics. Our work differs from it by designing and learning dynamic visual graph over video objects and using transformers to capture the temporal dynamics at both local and global scopes.

3   Method

3.1   Overview

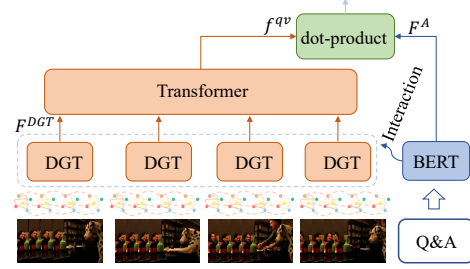Given a video v and a question q, VideoQA aims to combine the two stream information v and q to predict the answer a. Depending on the task settings, a can be given in multiple choices along with each question for multi-choice QA, or it is given in a global answer set for open-ended QA. In this work, we handle both types of VideoQA by optimizing the following objective:

$$a^* = \arg\max_{a \in \mathcal{A}} \mathcal{F}_W(a|q, v, \mathcal{A}), \qquad (1)$$

in which $\mathcal{A}$ can be $\mathcal{A}_{mc}$ corresponding to the candidate answers of each question in multi-choice QA, or $\mathcal{A}_{oe}$ corresponding to the global answer set in open-ended QA. $\mathcal{F}_W$ denotes the mapping function with learnable parameters W.

To solve the problem, we design a video graph transformer (VGT) model to perform the mapping $\mathcal{F}_W$ in Eqn. (1). As illustrated in Fig. 1, at the visual part (Orange), VGT takes as input visual object graphs, and drives a global feature f^{qv} with the integration of textual information, to represent the query-relevant video content. At the textual part (Blue), VGT extracts the feature

representations $F^{\mathcal{A}}$ for all the candidate answers via a language model (e.g., BERT [11]). The final answer a^* is determined by returning the candidate answers with maximal similarity (relevance score) between f^{qv} and $f^a \in F^{\mathcal{A}}$ via dot-product. At the heart of the model is the dynamic graph transformer module (DGT). The module clip-wisely reasons over the input graphs, and ag-

Fig. 1. Overview of video graph transformer (VGT) for VideoQA.

gregates them into a sequence of feature representations F^{DGT} which are then fed to a global transformer to achieve f^{qv}. During training, the whole framework is end-to-end optimized with Softmax cross-entropy loss. For pretraining with weakly-paired video-text data, we adopt cross-modal matching as the major proxy task and optimize the model in a contrastive manner [44] along with masked language modelling [11].
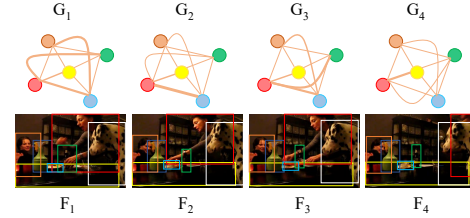
3.2 Video Graph Representation

Given a video, we sparsely sample l_v frames in a way analogous to [60]. The l_v frames are evenly distributed into k clips of length $l_c = \frac{l_v}{k}$. For each sampled frame (see Fig. 2), we extract n RoI-aligned features as object appearance representations $F_r = \{f_{r_i}\}_{i=1}^n$ along with their spatial locations $B = \{b_{r_i}\}_{i=1}^n$ with a pretrained object detector [2,45], where r_i represents the i-th object region in a frame.

Fig. 2. Illustration of graph construction in a short video clip of $l_c = 4$ frames. The nodes of same color denote same object.

Additionally, we obtain an image-level feature $F_I = \{f_{I_t}\}_{t=1}^{l_v}$ for all the sampled frames with a pretrained image classification model [18]. F_I serve as global contexts to augment the graph representations aggregated from the local objects.

To find the same object across different frames within a clip, we define a linking score s by considering their appearance and spatial location:

$$s_{i,j} = \psi(f_{r_i}^t, f_{r_j}^{t+1}) + \lambda * \mathrm{IoU}(b_i^t, b_j^{t+1}), \quad t \in \{1, 2, \ldots, l_c - 1\}, \quad (2)$$

where ψ denotes the cosine similarity between two detected objects i and j in adjacent frames. Intersection-over-union (IoU) computes the location overlap of objects i and j. Our experiments always set λ as one. The n detected objects in the first frame of each clip are designated as anchor objects. Detected objects in consecutive frames are then linked to the anchor objects by greedily maximizing s frame by frame[5]. By aligning objects within a clip, we ensure the consistency of the node and edge representations for the graphs constructed at different frames.

---

[5] We assume that the group of objects do not change in a short video clip.

Next, we concatenate the object appearance f_r and location f_{loc} representations and project the combined feature into the d-dimensional space via

$$f_o = \text{ELU}(\phi_{W_o}([f_r; f_{loc}])), \tag{3}$$

where $[;]$ denotes feature concatenation and f_{loc} is obtained by applying a 1×1 convolution over the relative coordinates as in [60]. The function ϕ_{W_o} denotes a linear transformation with parameters W_o. With $F_o = \{f_{o_i}\}_{i=1}^n$, the relations in the t-th frame can be initialized as pairwise similarities:

$$R_t = \sigma(\phi_{W_{ak}}(F_{o_t})\phi_{W_{av}}(F_{o_t})^\top), \quad t \in \{1, 2, \ldots, l_v\}, \tag{4}$$

where $\phi_{W_{ak}}$ and $\phi_{W_{av}}$ denote linear transformations with parameters W_{ak} and $W_{av} \in \mathbb{R}^{d \times \frac{d}{2}}$ respectively. We use different transformations to reflect the asymmetric nature of real-world subject-object interactions [58,26]. For symmetric relations, we expect that the learned parameters W_{ak} and W_{av} are quite similar. σ is the Softmax operation that normalizes each row. For brevity, we use $G_t = (F_{o_t}, R_t)$ to denote the graph representation of the t-th frame where F_o are node representations and R are edge representations of the graph.

3.3   Dynamic Graph Transformer

Our dynamic graph transformer (DGT) takes as input a set of visual graphs $\{G_t\}_{t=1}^{L_v}$ clip-wisely, and outputs a sequence of representations $F^{DGT} \in \mathbb{R}^{d \times k}$ by mining the temporal dynamics of objects and their spatial interactions. To this end, we sequentially operate a temporal graph transformer unit, a spatial graph convolution unit and a hierarchical aggregation unit as detailed below.

Temporal Graph Transformer As illustrated in Fig. 3, the temporal graph transformer unit takes as input a set of graphs G_{in} and outputs a new set of graphs G_{out} by mining the temporal dynamics among them via a node transformer (NTrans) and an edge transformer (ETrans). For completeness, we briefly recap the self-attention in Transformer [55]. It uses a multi-head self-attention (MHSA) to fuse a sequence of input features $X_{in} = \{x_{in}^t\}_{t=1}^l$:

$$X_{out} = \text{MHSA}(X_{in}) = \phi_{W_c}([h_1; h_2; \ldots, h_e]), \tag{5}$$

where ϕ_{W_c} is a linear transformation with parameters W_c, and

$$h_i = \text{SA}(\phi_{W_{i_q}}(X_{\text{in}}), \phi_{W_{i_k}}(X_{\text{in}}), \phi_{W_{i_v}}(X_{\text{in}})), \tag{6}$$

where $\phi_{W_{i_q}}$, $\phi_{W_{i_k}}$ and $\phi_{W_{i_v}}$ denote the linear transformations of the query, key, and value vectors of the i-th self-attention (SA) head respectively. e denotes the number of self-attention heads, and SA is defined as:

$$\text{SA}(X_q, X_k, X_v) = \sigma\left(X_k X_q^\top / \sqrt{d_k}\right) X_v, \tag{7}$$

in which d_k is the dimension of the key vector. Finally, a skip-connection with layer normalization (LN) is applied to the output sequence $X = LN(X_{out} + X_{in})$. X can undergo more MHSAs depending on the number of transformer layers.

In temporal graph transformer, we apply H self-attention blocks to enhance the node (or object) representations by aggregating information from other nodes of the same object from all adjacent frames within a clip:

$$F'_{o_i} = \text{NTrans}(F_{o_i}) = \text{MHSA}^{(H)}(F_{o_i}), \tag{8}$$

in which $F_{o_i} \in \mathbb{R}^{l_c \times d}$ denotes a sequence of feature representations corresponding to object i in a video clip of length l_c. Our motivation behind the node

transformer is that it models the change of single object behaviours and thus infer the dynamic actions (e.g. `bend down`). Also, it is helpful in improving the objects' appearance feature in the cases where the object at certain frames suffer from motion blur or partial occlusion.

Based on the new nodes $F'_o = \{F'_{o_i}\}_{i=1}^n$, we update the relation matrix R via Eqn. (4). Then, to explicitly model the temporal relation dynamics, we apply an edge transformer on the updated relation matrices:

Fig. 3. Illustration of temporal graph transformer in a short video clip.

$$\mathcal{R}' = \text{ETrans}(\mathcal{R}) = \text{MHSA}^{(H)}(\mathcal{R}), \tag{9}$$

where $\mathcal{R} = \{R_t\}_{t=1}^l \in \mathbb{R}^{l_c \times d_n}$ $(d_n = n^2)$ is the l_c adjacency matrices that are row-wisely expanded. Our motivation is that the relations captured at static frames may be spurious, trivial or incomplete. The edge transformer can help to calibrate the wrong relations and recall the missing ones. For brevity, we refer to the temporally contextualized graph at the t-th frame as $G_{out_t} = (F'_{o_t}, R'_t)$.

Spatial Graph Convolution The temporal graph transformer focuses on temporal relation reasoning. To reason over the object spatial interactions, we apply a U-layer graph attention convolution [25] on all the l_v graphs:

$$F'^{(u)}_o = \text{ReLU}((R' + I)F'^{(u-1)}_o W^{(u)}), \tag{10}$$

where $W^{(u)}$ is the graph parameters at the u-th layer. I is the identity matrix for skip connections. $F'^{(u)}_o$ are initialized by the output node representations F'_o as aforementioned. The index t is omitted for brevity. A last skip-connection: $F_{o_{out}} = F'_o + F'^{(U)}_o$ is used to obtain the final node representations.

Hierarchical Aggregation The node representations so far have explicitly token into account the objects' spatial and temporal interactions. But such interactions are mostly atomic. To aggregate these atomic interactions into higher-level video elements, we adopt a hierarchical aggregation strategy in Fig. 4. First, we aggregate the graph nodes at each frame by a simple attention:
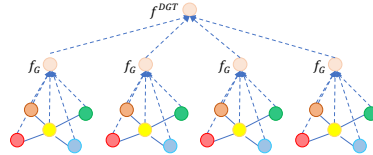

Fig. 4. Hierarchical Aggregation.

$$f_G = \sum_{i=1}^N \alpha_i F_{o_{out_i}}, \quad \alpha = \sigma(\phi_{W_G}(F_{o_{out}})), \tag{11}$$

where ϕ_{W_G} is linear transformation with parameters $W_G \in \mathbb{R}^{d \times 1}$. The graph representation f_G captures a local object interactions. It may lose sight of a global picture of a frame, especially since we only retain n objects and cannot guarantee that they include all the objects of interest in that frame. As such, we complement f_G with the frame-level feature f_I by concatenation:

$$f_G = \text{ELU}(\phi_{W_m}([\phi_{W_f}(f_I); f_G])) \tag{12}$$
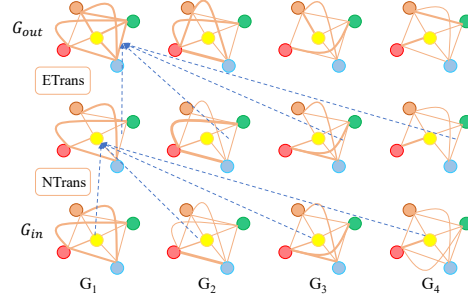
in which ϕ_{W_m} and ϕ_{W_f} are linear transformations with parameters $W_m \in \mathbb{R}^{2d \times d}$ and $W_f \in \mathbb{R}^{2048 \times d}$ respectively. We next pool the local interactions to obtain a sequence of clip-level feature representations via:

$$f^{\mathrm{DGT}} = \mathrm{MPool}(F_G) = \frac{1}{l_c} \sum_{t=1}^{l_v} f_{G_t} \tag{13}$$

The set of k clips are finally represented by $F^{\mathrm{DGT}} = \{f_c^{\mathrm{DGT}}\}_{c=1}^k$.

3.4   Cross-modal Interaction

To find the informative visual contents with respect to a particular text query, a cross-model interaction between the visual and textual nodes is essential. Given a set of visual nodes denoted by X^v, we integrate textual information $X^q = \{x_m^q\}_{m=1}^M$ into the visual nodes via a simple cross-modal attention:

$$x^{qv} = x^v + \sum_{m=1}^M \beta_m x_m^q, \quad \text{where} \quad \beta = \sigma(x^v (X^q)^\top), \tag{14}$$

where M is the number of tokens in the text query. In principle, the X^v can be visual representations from different levels of the DGT module. In our experiment, we explore performing the cross-modal interaction with visual representations at the object-level (F_O in Eqn. (3)), frame-level (F_G in Eqn. (12)), and clip-level (F^{DGT} in Eqn. (13)). We find that the results vary among different datasets. As a default, we perform cross-modal interaction at the clip-level outputs (*i.e.*, the outputs of the DGT module $X^v := F^{\mathrm{DGT}}$), since the number of nodes at this stage is much smaller, and the node representations have already absorbed the information from the preceding layers. For the text node X^q, we obtain them by a simple linear projection on the token outputs of a language model [11]:

$$X^q = \phi_{W_Q}(\mathrm{BERT}(Q)), \tag{15}$$

where $W_Q \in \mathbb{R}^{768 \times d}$. The text query Q can be questions in open-end QA or QA pairs in multi-choice QA. Note that in multi-choice QA, we max-pool the obtained query-aware visual representations with respect to different QA pairs to find the one that is mostly relevant to the video.

3.5   Global Transformer

The aforementioned DGT module pays attention to extract informative visual clues from video clips. To capture the temporal dynamics between these clips, we employ another H-layer transformer over the cross-modal interacted clip feature (*i.e.* F^{DGT}), and add learnable sinusoidal temporal position embeddings [11]. Finally, the transformer's outputs are mean-pooled to obtain the global representation $f^{qv} \in \mathbb{R}^d$ for the entire video, which is defined as follows:

$$f^{qv} = \mathrm{MPool}(\mathrm{MHSA}^{(H)}(F^{\mathrm{DGT}})). \tag{16}$$

The global transformer has two major advantages: 1) It retains the overall hierarchical structure which progressively drives the video elements at different granularity as in [60]. 2) It improves the feature compatibility of vision and text, which may benefit cross-modal comparison.

3.6   Answer Prediction

To obtain a global representation for a particular answer candidate, we mean-pool its token representations from BERT by $f^A = \text{MPool}(X^A)$, where X^A denotes a candidate answer's token representations, and is obtained in a way analogous to Eqn. (15). Its similarity with the query-aware video representation f^{qv} is then obtained via a dot-product. Consequently, the candidate answer of maximal similarity is returned as the final prediction:

$$s = f^{qv}(F^A)^\top, \quad a^* = \arg\max(s), \tag{17}$$

in which $F^A = \{f_a^A\}_{a=1}^{|\mathcal{A}|} \in \mathbb{R}^{|\mathcal{A}|\times d}$, and $|\mathcal{A}|$ denotes the number of candidate answers. Additionally, for open-ended QA, we follow previous works [60] and enable a video-absent QA by directly computing the similarities between the question representation f^q (obtained in a way similar to f^A) and the answer representations F^A. As a result, the final answer can be a joint decision:

$$s = f^{qv}(F^A)^\top \odot f^q(F^A)^\top \tag{18}$$

in which $\odot$ is element-wise product. During training, we maximize the $\langle \text{VQ}, \text{A}\rangle$ similarity corresponding to the correct answer of a given sample by optimizing the Softmax cross entropy loss function. $\mathcal{L} = -\sum_{i=1}^{|\mathcal{A}|} y_i \log s_i$, where s_i is the matching score for the i-th sample. $y_i = 1$ if the answer index corresponds to the i-th sample's ground-truth answer and 0 otherwise.

3.7   Pretraining with Weakly-Paired Data

For cross-model matching, we encourage the representation of each video-text interacted representation f^{qv} to be closer to that of its paired description f^q and be far away from that of negative descriptions which are randomly collected from other video-text pairs in each training iteration. This is formally achieved by maximizing the following contrastive objective:

$$\sum_i \log\left(\frac{\exp\left(f_i^{qv}(f_i^q)^\top\right)}{\exp\left(f_i^{qv}(f_i^q)^\top\right) + \sum_{(f^{qv},f^q)\in\mathcal{N}_i} \exp\left(f^{qv}(f^q)^\top\right)}\right), \tag{19}$$

where $\mathcal{N}_i$ denotes the representations of all the negative video-description pairs of the i-th sample. The parameters to be optimized are hidden in the process of calculating f^{qv} and f^q as introduced above. For negative sampling, we sample them from the whole training set at each iteration. For masked language modelling, we only corrupt the positive description of each video for efficiency.

4   Experiment

4.1   Dataset and Configuration

We conduct experiments on benchmarks whose QAs feature temporal dynamics: 1) NExT-QA [59] is a manually annotated dataset that features causal and temporal object interaction in space-time. 2) TGIF-QA [20] features short GIFs; it asks questions about repeated action recognition, temporal state transition

Table 1. Results on NExT-QA [59]. (Acc@C, T, D: Accuracy for Causal, Temporal and Descriptive questions respectively. *: Results reproduced with the official code.)

Method	CM-Pretrain	NExT-QA Val				NExT-QA Test			
		Acc@C	Acc@T	Acc@D	Acc@All	Acc@C	Acc@T	Acc@D	Acc@All
HGA [23]	-	46.26	50.74	59.33	49.74	48.13	49.08	57.79	50.01
IGV [35]	-	-	-	-	-	48.56	51.67	59.64	51.34
HQGA [60]	-	48.48	51.24	61.65	51.42	49.04	**52.28**	59.43	51.75
P3D-G [9]	-	51.33	52.30	62.58	53.40	-	-	-	-
VQA-T* [64]	-	41.66	44.11	59.97	45.30	42.05	42.75	55.87	44.54
VQA-T* [64]	How2VQA69M	49.60	51.49	63.19	52.32	47.89	50.02	61.87	50.83
VGT (Ours)	-	**52.28**	**55.09**	**64.09**	**55.02**	**51.62**	51.94	**63.65**	**53.68**

and frame QA which invokes a certain frame for answer. For better comparison, we also experiment on MSRVTT-QA [62] which challenges a holistic visual recognition or description. Other data statistics are presented in Appendix A.

We decode the video into frames following [60], and then sparsely sample $l_v = 32$ frames from each video. The frames are distributed into $k = 8$ clips whose length $l_c = 4$. For each frame, we detect and keep $N = 20$ regions of high confidence for NExT-QA (Top-5 are used in the pretraining-free experiments, refer to our analysis in Appendix C.2), and $N = 10$ for the other datasets, using the object detection model provided by [2]. The dimension of the models' hidden states is $d = 512$. The default number of layers and self-attention heads in transformer are $H = 1$ and $e = 8$ ($e = 5$ for edge transformer in DGT) respectively. Besides, the number of graph layers is $U = 2$. For training, we use Adam optimizer with initial learning rate 1×10^{-5} of a cosine annealing schedule. The batch size is set to 64, and the maximum epoch varies from 10 to 30 among different datasets. Our pretraining data (~ 0.18M) are collected from WebVid [3]. More details are presented in Appendix B.

4.2 Sate-of-the-Art Comparison

In Table 1, we compare VGT with the prior arts on NExT-QA [60]. The results show that VGT surpasses the previous SoTAs by clear margins on both the val and test sets, improving the overall accuracy by 1.6% and 1.9% respectively. VGT even outperforms a latest work ATP [6] which is based on CLIP features [44] (VGT *vs.* ATP: 55.02% *vs.* 54.3%), and thus sets the new SoTA results. In particular, we note that such strong results come without considering large-scale cross-modal pretraining. When pretraining VGT with (relatively) small amount of data, we can further increase the results to 56.9% and 55.7% on NExT-QA val and test sets respectively (refer to our analysis of Table 5 in Sec. 4.4).

Compared with VQA-T [64] which also formulates VideoQA as problem of similarity comparison instead of classification, VGT outperforms it almost in all metrics. The strong results could be due to that VGT explicitly models the object interactions and dynamics for visual reasoning, instead of holistically encoding video clips with S3D [39,61]. For a better analysis, we further replace the S3D encoder in VQA-T with our DGT module. As shown in Table 2 (S3D → DGT), our DGT encoder significantly improves VQA-T's result by 4.7%, in which most of the improvements are from answering reasoning

type of questions. Aside from the DGT module, we encode the candidate answers in the context of the corresponding question with a single language model, whereas VQA-T encodes Q and A independently with two language models [46]. Our

Table 2. Detailed comparison with VQA-T [64]. CMTrans: Cross-Modal Transformer.

Models	Size (M)	NExT-QA Val			
		Acc@C	Acc@T	Acc@D	Acc@All
VQA-T [64]	600	41.66	44.11	59.97	45.30
S3D→DGT	641	47.53	48.08	62.42	50.02
CMTrans→CM	573	42.27	44.29	58.17	45.40
VGT (DistilBERT)	346	50.71	51.67	66.41	53.46
VGT (BERT)	511	52.28	55.09	64.09	55.02

method improves answer encoding with contexts and reduces the model size (or parameters), as shown in Table 2 (VGT (DistilBERT)). Finally, VQA-T adopts cross-modal transformer to fuse the video-question pair, whereas we design lightweight cross-modal interaction module. The module is more parameter efficient but has little impact on the performances (CMTrans→CM in Table 2).

Compared with other graph based methods [9,23,60], VGT enjoys several advantages: 1) It explicitly model the temporal dynamics of both objects and their interactions. 2) It solves VideoQA by explicit similarity comparison between the video and text instead of classification. 3) It represents both visual and textual data with Transformers which may improve the feature compatibility and benefit cross-modal interaction and comparison [11]. 4) VGT uses much few frames for training and inference (e.g., VGT *vs.* HQGA [60]: 32 *vs.* 256), which benefits efficiency for video encoding. The detailed analyses are given in Sec. 4.3.

In Table 3, we compare VGT with previous arts on the TGIF-QA and MSRVTT-QA datasets. The results show that VGT performs pretty well on the tasks of repeating action recognition and state transition that feature temporal dynamics, surpassing the previous pretraining-free SoTA results significantly by 10.6% (VGT *vs.* MASN [47]: 95.0% *vs.* 84.4%) and 6.8% (VGT *vs.* MHN [43]: 97.6% *vs.* 90.8%) respectively. It even beats the pretraining SoTA (*i.e.* MERLOT [70]) by about 1.0%, yet without using external data for cross-modal pretraining. On TGIF-QA-R [42] which is curated by making the negative answers in TGIF-QA more challenging, we can also observe remarkable improvements. Besides, VGT also achieves competitive results on normal descriptive QA tasks as defined in FrameQA and MSRVTT-QA though they are not our focus.

4.3   Model Analysis

DGT. The middle block of Table 4 shows that removing the DGT module (w/o DGT) (i.e. directly summarizing the object representations in each clip) leads to clear performance drops (∼2.0%) on all tasks that challenge spatio-temporal reasoning. We then study the temporal graph transformer module (w/o TTrans) by removing both NTrans and ETrans. It shows better results than removing the whole DGT module. Yet, its performances on tasks featuring temporal dynamics are still weak. We further ablate the temporal graph transformer module to investigate the independent contribution of the node transformer (NTrans) and edge transformer (ETrans). The results (w/o NTrans and w/o ETrans) demonstrate that both transformers benefit temporal dynamic modelling. Finally, the ablation study on the global frame feature F_I reveals its vital role to DGT.

Table 3. Results on TGIF-QA and MSVTT-QA. † denotes TGIF-QA-R [42] whose multiple choices for repeated action and state transition are more challenging. We grey out the results reported in [42] regarding these two sub-tasks, because the candidate answers are slightly different as we have further rectified the redundant choices.
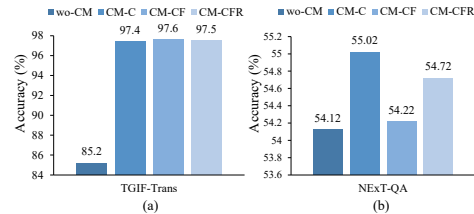
Models	CM-Pretrain	TGIF-QA					MSRVTT-QA
		Action	Transition	FrameQA	Action†	Transition†	-QA
LGCN [19]	-	74.3	81.1	56.3	-	-	-
HGA [23]	-	75.4	81.0	55.1	-	-	35.5
HCRN [28]	-	75.0	81.4	55.9	55.7	63.9	35.6
B2A [41]	-	75.9	82.6	57.5	-	-	36.9
HOSTR [10]	-	75.0	83.0	58.0	-	-	35.9
HAIR [36]	-	77.8	82.3	60.2	-	-	36.9
MASN [47]	-	84.4	87.4	59.5	-	-	35.2
PGAT [42]	-	80.6	85.7	61.1	58.7	65.9	38.1
HQGA [60]	-	76.9	85.6	61.3	-	-	38.6
MHN [43]	-	83.5	90.8	58.1	-	-	38.6
ClipBERT [29]	VG+COCO Caption	82.8	87.8	60.3	-	-	37.4
SiaSRea [67]	VG+COCO Caption	79.7	85.3	60.2	-	-	41.6
MERLOT [70]	Youtube180M, CC3M	94.0	96.2	**69.5**	-	-	**43.1**
VGT (Ours)	-	**95.0**	**97.6**	61.6	**59.9**	**70.5**	39.7

Similarity Comparison *vs.* Classification. We study a model variant by concatenating the outputs of the DGT module with the token representations from BERT in a way analogous to ClipBERT [29]. The formed text-video representation sequence is fed to a cross-modal transformer for

Table 4. Study of model components

Models	TGIF-QA		NExT-QA Val			
	Action	Trans	Acc@C	Acc@T	Acc@D	Acc@All
VGT	**95.0**	**97.6**	**52.28**	**55.09**	64.09	**55.02**
w/o DGT	89.6	95.4	50.10	52.85	64.48	53.22
w/o TTrans	94.0	97.6	50.86	53.04	64.86	53.74
w/o NTrans	94.5	97.4	50.79	54.22	63.32	53.84
w/o ETrans	94.8	97.4	51.25	54.34	64.48	54.30
w/o F_I	93.5	97.0	50.44	53.97	63.32	53.58
Comp→CLS	70.1	79.9	42.96	46.96	53.02	45.82

information fusion. Then, the output of the '[CLS]' token is fed to a $|\mathcal{A}|$-way classifier in open-ended QA or a 1-way classifier for binary relevance in multi-choice QA following [20,28,60]. As can be seen from the bottom part of Table 4, this classification model variant (Comp → CLS) leads to drastic performance drops. To be complete, we also conduct additional experiments on the FrameQA task which is set as open-ended QA. Again, we find that the accuracy drops from 61.6% to 56.9%. A detailed analysis of the performances on the training and validation sets (see Appendix C.1.) reveals that the CLS-model suffers from serious over-fitting on the target datasets. The experiment demonstrates the superiority of solving QA by relevance comparison instead of answer classification.

Cross-modal Interaction. Fig. 5 investigates several implementation variants of the cross-modal interaction module as depicted in Sec. 3.4. The results suggest that it is better to integrate textual information at both the frame- and clip-level outputs (CM-CF) for TGIF-QA, while our default interaction at the clip-level outputs (CM-C) brings the optimal results on NExT-QA. Compared with the baselines that do not use cross-modal interaction, all three kinds of interactions improve the performances. We notice

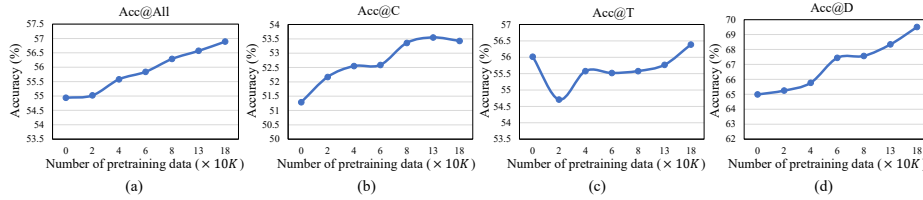**Fig. 5.** Study of Cross-modal Interaction.

Fig. 6. Results of pretraining with different amounts of data.

Table 5. Study of cross-model pretraining. Results on NExT-QA are with 20 regions.

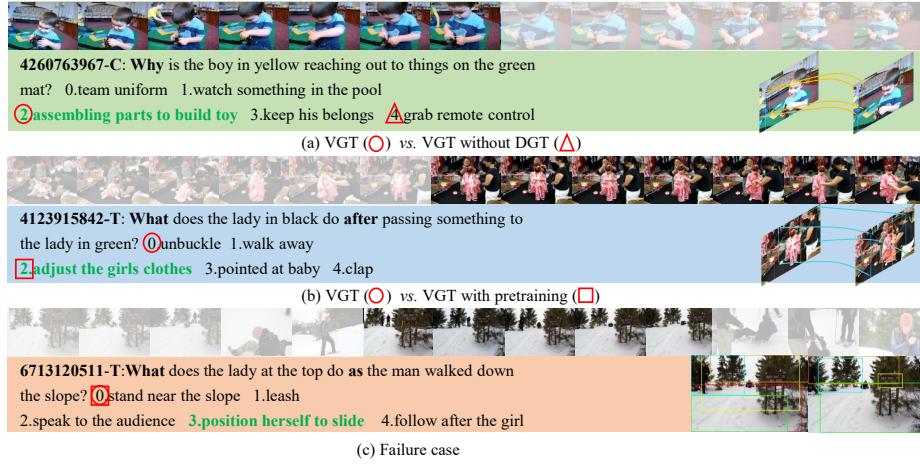| Methods	TGIF-QA		NExT-QA Val				NExT-QA Test			
	Action†	Trans†	Acc@C	Acc@T	Acc@D	Acc@All	Acc@C	Acc@T	Acc@D	Acc@All
VGT	59.9	70.5	51.29	56.02	64.99	54.94	50.82	52.29	63.27	53.51
VGT (FT w/ QA)	60.2	71.0	**53.93**	56.20	**70.14**	**57.19**	51.73	53.78	67.05	54.88
VGT (FT w/ QA & MLM)	**60.5**	**71.5**	53.43	**56.39**	69.50	56.89	**52.78**	**54.54**	**67.26**	**55.70**

that the cross-modal interaction improves the accuracy on TGIF-QA by more than 10%. A possible reason is that the GIFs are trimmed short videos that only contain the QA-related visual contents. This greatly eases the challenge in spatial-temporal grounding of the positive answers, especially when most of the negative answers are not presence in the short GIFs. Thus, the cross-modal interaction performs more effectively on this dataset. The videos in NExT-QA are not trimmed, thereby the improvements are relatively smaller. Base on these observations, we perform cross-modal interaction at both the frame- and clip-level outputs for the temporal reasoning tasks in TGIF-QA, and keep the default implementation for other datasets.

4.4   Pretraining and Finetuning

Table 5 presents a comparison between VGT with and without pretraining. We can see that pretraining can steadily boost the QA performance, especially on NExT-QA. The relatively smaller improvements on TGIF-QA could be due to that TGIF-QA dataset is large, and has enough annotated data for fine-tuning. As such, pretraining helps little [73]. Besides, we find that finetuning with masked language modelling (MLM) can improve the generalization from val to test set, and thus achieves the best overall accuracy (*i.e.* 55.7%) on NExT-QA test set. Fig. 6 studies the QA performances on NExT-QA val set with respect to different amounts of pretraining data. Generally, there is a clear tendency of performance improvements for the overall accuracy (Acc@All) when more data is available. A more detailed analysis shows that these improvements mostly come from a stronger performance in answering causal (Acc@C) and descriptive (Acc@D) questions. For temporal questions, it seems that pretraining with more data does not help much. Therefore, to boost performance, it is promising to add more data or explore a better way to handle temporal languages.

4.5   Qualitative Analysis

In Fig. 7, we qualitatively analyze the benefits of both dynamic graph transformer and pretraining. The example in (a) shows that the model without the

4260763967-C: Why is the boy in yellow reaching out to things on the green mat?   0.team uniform   1.watch something in the pool   ②assembling parts to build toy   3.keep his belongs   △grab remote control

(a) VGT (○) *vs.* VGT without DGT (△)

4123915842-T: What does the lady in black do **after** passing something to the lady in green? ○unbuckle   1.walk away   2.adjust the girls clothes   3.pointed at baby   4.clap

(b) VGT (○) *vs.* VGT with pretraining (□)

6713120511-T:What does the lady at the top do **as** the man walked down the slope? ○stand near the slope   1.leash   2.speak to the audience   3.position herself to slide   4.follow after the girl

(c) Failure case

Fig. 7. Result visualization on NExT-QA [59]. The ground-truth answers are in green.

DGT module is prone to predicting atomic or contact actions (e.g. 'grab') that can be captured at static frame-level. (b) shows that the model without pretraining fails to predict the answer that is highly abstract (e.g. 'adjust'). Finally, we show a failure case in (c). It indicates that our model tends to predict distractor answers that are semantically close to the questions when the object of interests in the video are small and the detector fails to detect it. Keeping more detected regions could be helpful, but one needs to carefully balance the graph complexity as well as the inference efficiency. Another alternative is to perform modulated detection as in [24], we leave it for future exploration.

5   Conclusions

We presented video graph transformer which explicitly exploits the objects, their relations, and dynamics, to improve visual reasoning and alleviate the data-hungry issue for VideoQA. Our extensive experiments show that VGT can achieve superior performances as compared with previous SoTA methods on tasks that challenge temporal dynamic reasoning. The performance even surpasses those methods that are pretrained on large-scale vision-text data. To study the learning capacity of VGT, we further explored pretraining on weakly-paired video-text data and obtained promising results. With careful and comprehensive analyses of the model, we hope this work can encourage more efforts in designing effectiveness models to alleviate the burden of handling large-scale data, and also promote VQA research that goes beyond a holistic recognition/description to reason about the fine-grained video details.

References

1. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 35, pp. 6644–6652 (2021)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086 (2018)
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1728–1738 (2021)
4. Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261 (2018)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. pp. 813–824. PMLR (2021)
6. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the" video" in video-language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2917–2927 (2022)
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision (ECCV). pp. 104–120. Springer (2020)
9. Cherian, A., Hori, C., Marks, T.K., Le Roux, J.: (2.5+ 1) d spatio-temporal scene graphs for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 444–453 (2022)
10. Dang, L.H., Le, T.M., Le, V., Tran, T.: Hierarchical object-oriented spatio-temporal reasoning for video question answering. In: IJCAI (August 2021)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
12. Ding, D., Hill, F., Santoro, A., Reynolds, M., Botvinick, M.: Attention over learned object embeddings enables complex visual reasoning. Advances in neural information processing systems (NeurIPS) **34** (2021)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Representation Learning (ICLR) (2020)
14. Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1999–2007 (2019)
15. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. In: arXiv preprint arXiv:2111.12681 (November 2021)

16. Gao, J., Ge, R., Chen, K., Nevatia, R.: Motion-appearance co-memory networks for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6576–6585 (2018)
17. Geng, S., Gao, P., Chatterjee, M., Hori, C., Le Roux, J., Zhang, Y., Li, H., Cherian, A.: Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In: AAAI Conference on Artificial Intelligence (AAAI) (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
19. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 11021–11028 (2020)
20. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2758–2766 (2017)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021)
22. Jiang, J., Chen, Z., Lin, H., Zhao, X., Gao, Y.: Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 34, pp. 11101–11108 (2020)
23. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: AAAI Conference on Artificial Intelligence (AAAI) (2020)
24. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1780–1790 (2021)
25. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Representation Learning (ICLR) (2017)
26. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6867–6876 (2018)
27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017)
28. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9972–9981 (2020)
29. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7331–7341 (2021)
30. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: Empirical Methods in Natural Language Processing (EMNLP) (2018)
31. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in neural information processing systems (NeurIPS). vol. 34 (2021)
32. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. In: Proceedings of the 2020

Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2046–2065 (2020)

33. Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: AAAI Conference on Artificial Intelligence (AAAI). pp. 8658–8665 (2019)

34. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision (ECCV). pp. 121–137. Springer (2020)

35. Li, Y., Wang, X., Xiao, J., Ji, W., Chua, T.S.: Invariant grounding for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2928–2937 (2022)

36. Liu, F., Liu, J., Wang, W., Lu, H.: Hair: Hierarchical visual-semantic relational reasoning for video question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1698–1707 (October 2021)

37. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3202–3211 (2022)

38. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems (NeurIPS). pp. 13–23 (2019)

39. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9879–9889 (2020)

40. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2630–2640 (2019)

41. Park, J., Lee, J., Sohn, K.: Bridge to answer: Structure-aware graph interaction network for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15526–15535 (2021)

42. Peng, L., Yang, S., Bin, Y., Wang, G.: Progressive graph attention network for video question answering. In: ACM MM. pp. 2871–2879 (2021)

43. Peng, M., Wang, C., Gao, Y., Shi, Y., Zhou, X.D.: Multilevel hierarchical network with multiscale sampling for video question answering. IJCAI (2022)

44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)

45. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (NeurIPS) **28** (2015)

46. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Advances in neural information processing systems (NeurIPS) (2019)

47. Seo, A., Kang, G.C., Park, J., Zhang, B.T.: Attend what you need: Motion-appearance synergistic networks for video question answering. In: ACL. pp. 6167–6177 (2021)

48. Seo, P.H., Nagrani, A., Schmid, C.: Look before you speak: Visually contextualized utterances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16877–16887 (2021)
49. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR). pp. 279–287 (2019)
50. Shang, X., Xiao, J., Di, D., Chua, T.S.: Relation understanding in videos: A grand challenge overview. In: Proceedings of the 27th ACM International Conference on Multimedia (MM). pp. 2652–2656 (2019)
51. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL. pp. 2556–2565 (2018)
52. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Representation Learning (ICLR) (2020)
53. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7464–7473 (2019)
54. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP. pp. 5100–5111 (2019)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems (NeurIPS). vol. 30 (2017)
56. Wang, L., Chang, X., Li, S., Chu, Y., Li, H., Zhang, W., He, X., Song, L., Zhou, J., Yang, H.: Tcl: Transformer-based dynamic graph modelling via contrastive learning. arXiv preprint arXiv:2105.07944 (2021)
57. Wang, X., Gupta, A.: Videos as space-time region graphs. In: European conference on computer vision (ECCV). pp. 399–417 (2018)
58. Xiao, J., Shang, X., Yang, X., Tang, S., Chua, T.S.: Visual relation grounding in videos. In: European Conference on Computer Vision (ECCV). pp. 447–464. Springer (2020)
59. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9777–9786 (2021)
60. Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., Chua, T.S.: Video as conditional graph hierarchy for multi-granular question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 2804–2812 (2022)
61. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: European Conference on Computer Vision (ECCV). pp. 305–321 (2018)
62. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM MM. pp. 1645–1653 (2017)
63. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP. pp. 6787–6800 (2021)
64. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1686–1697 (2021)

65. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. In: International Conference on Learning Representations (ICLR) (2019)
66. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., Liu, T.Y.: Do transformers really perform badly for graph representation? Advances in neural information processing systems (NeurIPS) **34** (2021)
67. Yu, W., Zheng, H., Li, M., Ji, L., Wu, L., Xiao, N., Duan, N.: Learning from inside: Self-driven siamese sampling and reasoning for video question answering. Advances in neural information processing systems (NeurIPS) **34** (2021)
68. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: European Conference on Computer Vision (ECCV). pp. 471–487 (2018)
69. Yun, S., Jeong, M., Kim, R., Kang, J., Kim, H.J.: Graph transformer networks. Advances in neural information processing systems (NeurIPS) **32** (2019)
70. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. In: Advances in neural information processing systems (NeurIPS). vol. 34 (2021)
71. Zhong, Y., Ji, W., Xiao, J., Li, Y., Deng, W., Chua, T.S.: Video question answering: Datasets, algorithms and challenges. arXiv preprint arXiv:2203.01225 (2022)
72. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8746–8755 (2020)
73. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. In: Advances in neural information processing systems (NeurIPS). vol. 33, pp. 3833–3845 (2020)