Graph Prompts: Adapting Video Graph for Video Question Answering

Yiming Li[1,5], **Xiaoshan Yang**[2,3,4], **Bing-Kun Bao**[1,4] and **Changsheng Xu**[2,3,4*]
[1]Nanjing University of Posts and Telecommunications
[2]Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Pengcheng Laboratory
[5]State Key Laboratory of Tibetan Intelligence
{yiming.li, bingkunbao}@njupt.edu.cn, {xiaoshan.yang, csxu}@nlpr.ia.ac.cn

Abstract

Due to the dynamic nature in videos, it is evident that perceiving and reasoning about temporal information are the key focus of Video Question Answering (VideoQA). In recent years, several methods have explored relationship-level temporal modeling with graph-structured video representation. Unfortunately, these methods heavily rely on the question text, thus making it challenging to perceive and reason about video content that is not explicitly mentioned in the question. To address the above challenge, we propose Graph Prompts-based VideoQA (GP-VQA), which adopts a video-based graph structure for enhanced video understanding. The proposed GP-VQA contains two stages, i.e., pre-training and prompt tuning. In pre-training, we define the pretext task that requires GP-VQA to reason about the randomly masked nodes or edges in the video graph, thus prompting GP-VQA to learn the reasoning ability with video-guided information. In prompt-tuning, we organize the textual question into question graph and implement message passing from video graph to question graph, therefore inheriting the video-based reasoning ability from video graph completion to VideoQA. Extensive experiments on various datasets have demonstrated the promising performance of GP-VQA.

1 Introduction

Video Question Answering (VideoQA) aims to assist humans in addressing everyday challenges [Wong *et al.*, 2022; Grauman *et al.*, 2022], such as helping users locate items, reminding them of past activities, and facilitating the completion of complex tasks, which is an intriguing cross-modal task bridging the domains of computer vision and natural language processing.

Each video is involved with dynamic actions, activities, and events, thus VideoQA models should be able to perceive and reason about the temporal information for answer generation. To capture detailed temporal information, recent methods have attempted to explore relation-level temporal modeling to avoid insufficient understanding of temporal context when modeling at the object level [Lei *et al.*, 2021; Xiao *et al.*, 2022; Liu *et al.*, 2021]. For example, some researchers [Li *et al.*, 2023d; Urooj *et al.*, 2023; Zong *et al.*, 2024] propose adopting multi-scale feature encoding to consider the temporal relationships of objects at different scales, such as objects, regions, and clips within the video. However, these methods do not explicitly define and model the temporal relationships between objects, thus they may encounter difficulties in capturing the fine-grained temporal dynamics and complex interactions among objects. Furthermore, to address explicitly define and model the temporal relationships between objects, recent approaches [Shi *et al.*, 2019; Park *et al.*, 2021] attempt to explore graph-structured video representation for VideoQA, as shown in Figure 1 (a). These methods adopt scene graph-liked structure which abstracts the visual content or textual question to relationship triplets, *e.g.*, $\langle subject, predicate, object \rangle$, thereby explicitly modeling the relationship. However, these question-guided temporal modeling methods rely heavily on the question text, which merely focuses on the question-mentioned objects and ignores other objects in the video that may be crucial for answer reasoning. For example, to deal with a reasoning-related question *"Why does the toddler cry in the last of video?"*, question-guided methods merely concern the mentioned "toddler" in the given question, while ignoring the unmentioned "dog" that interacts with the toddler, thus generating incorrect answers.

To address these challenge, we introduce the video-guided temporal modeling method, as shown in Figure 1 (b), which has the following advantages: (1) a more comprehensive understanding of the dynamic relationships between objects in the video, and (2) a stronger reasoning ability for video content not mentioned in the given question. Firstly, we construct a video graph and a question graph simultaneously, where the video graph contains objects and corresponding relationships of each frame, and the question graph is generated through semantic analysis of textual questions. Since the video graphs are generated based on video instead of textual questions, the perception of VideoQA model is no longer limited to the directly mentioned information in questions and achieving a comprehensive understanding of video. Then we can

*indicates corresponding author: Changsheng Xu

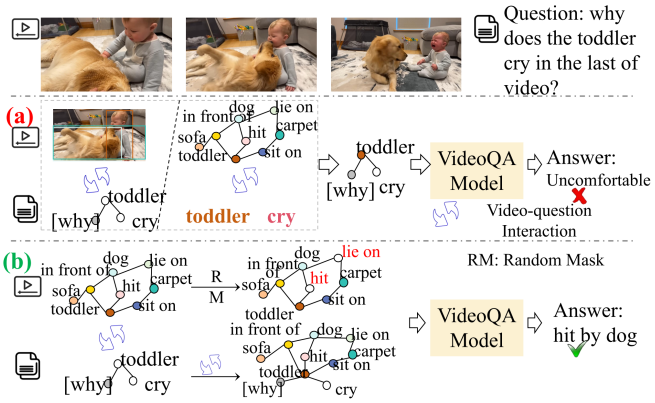

Figure 1: (a) Question-guided relationship-level temporal modeling by graph structure. (b) The proposed video-guided relationship-level temporal modeling.

adopt message passing between video and question graphs, thus integrating the video and question graph into a union graph structure. In this way, the VideoQA model can obtain stronger reasoning ability for all the video-contained objects by incorporating relationship-level information on the question semantics and all the video-contained objects. However, during the message-passing process, the interaction between the question graph and the video graph is based on nodes, inevitably leading to the model still being biased towards textual questions. Merely expanding the model's perception of the video is insufficient to address this issue.

Inspired by the success of recent prompt-tuning researches , we propose to adopt a two-stage training paradigm and introduce a video graph completion pretext task to address the above issue. In detail, we first randomly mask nodes and edges in the video graph, and the video graph completion is defined as completing the masked graph through temporal modeling, thus forcing the model to automatically learn each relationship in the video graph, rather than merely learning the question-mentioned object relationships. Moreover, since the video graph completion task is like a cloze-style question-answering task, it also facilitates the adaptation to VideoQA. Based on the proposed pretext task, we design a **G**raph **P**rompts based **V**ideo **Q**uestion **A**nswering model named GP-VQA. In pre-training, we introduce video SGG methods [Cong *et al.*, 2021; Li *et al.*, 2022b; Nag *et al.*, 2023] to provide a raw video graph, which makes the proposed GP-VQA able to adapt different models flexibly. Then we employ a vision-language model to learn temporal reasoning ability based on video graph completion pretext task. In the prompt-tuning stage, we first organize the questions as question graphs and adopt message-passing between the question graph and video graph, therefore perceiving the global information of the video and capturing the relevant information for answer reasoning. At the last, we take the question graph as a prompt and utilize the language module to predict the answer.

The main contributions are summarized as follows:

1. By incorporating video-based scene graphs instead of relying solely on question text for temporal modeling, we ensure that the proposed model obtains a more comprehensive perception of the video.
2. We propose a graph prompt-based VideoQA approach GP-VQA, which learns the reasoning ability from masked video graph completion pre-training and inherits this ability to VideoQA task in the prompt-tuning stage.
3. The evaluation on the three public VideoQA datasets verifies the effectiveness of the proposed GP-VQA, which also demonstrates the flexibility of GP-VQA.

2 Related Work

2.1 Video Question Answering

Some early works [Zeng *et al.*, 2017] apply the global video and question representations for answer prediction, which cannot construct robust fine-grained video semantics. For example, work [Zeng *et al.*, 2017] directly utilizes element-wise multiplication to capture the global multi-modal feature for answer classification. To capture the critical frames and regions, ST-VQA [Jang *et al.*, 2017] propose a dual-LSTM-based approach with spatial and temporal attention mechanisms. Work [Xu *et al.*, 2017] enhances the attention unit by incorporating the interaction between question words and both frame-level and clip-level visual features. Despite the aforementioned methods having the ability to attend to video frames and clips, they have later been shown to be weak in capturing long-term dependency.

To address the above problem, some work attempts to leverage the memory network [Bärmann and Waibel, 2022; Datta *et al.*, 2022] and Transformer [Li *et al.*, 2023c; Yuan *et al.*, 2023] to capture robust long-term representations. Specifically, memory networks can cache sequential inputs in memory slots and explicitly utilize long-term information, and Transformer utilizes the self-attention mechanism to model long-distance dependencies in context. Recently, graph-structured techniques have been favored for improving the reasoning ability of VideoQA models. Some work [Park *et al.*, 2021; Wang *et al.*, 2021; Gao *et al.*, 2023; Peng *et al.*, 2021; Huang *et al.*, 2020] build the graphs based on coarse-grained video segments. To capture the object-level information, L-GCN [Huang *et al.*, 2020] utilizes the object representation to construct graph, *e.g.,* appearance and location features. Furthermore, work [Peng *et al.*, 2021] concatenate the object-level, frame-level, and clip-level graphs to learn the visual relations

The most similar method to ours is [Park *et al.*, 2021], which also adopts question graph to connect textual and visual domains. The main difference is that we introduce video-based pre-training and design graph-based multimodal interaction to expand the model's perception and avoid limitations imposed by the question.

2.2 Scene Graph in VideoQA

Since the previous work succeeded in adopting scene graph in visual question answering, this idea has been extended to videoQA and captioning tasks [Tsai *et al.*, 2019; Cherian *et al.*, 2022; Park *et al.*, 2021]. For video understanding, scene graph generation is defined to decompose long-term actions into a series of frame-level relationships triplets [Ji

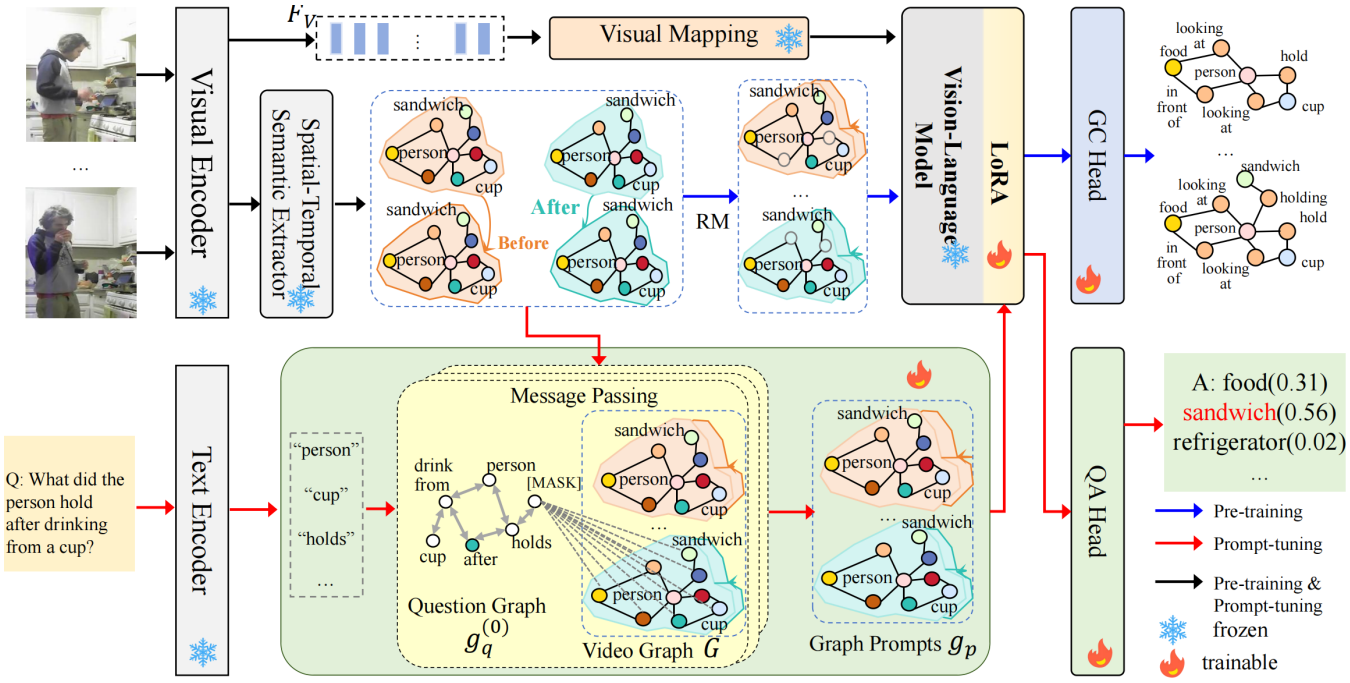

Figure 2: The framework of the proposed method. We employ the standard video scene graph generation model as spatial-temporal semantic extractor and adopt random mask operation to generate masked video graph, a language model with graph completion(GC) head is utilized in pre-training stage, and a question answering (QA) head is employed in prompt-tuning.

et al., 2020; Li *et al.*, 2022a; Li *et al.*, 2022b; Li *et al.*, 2023e]. STG [Pan *et al.*, 2020] explore combining video scene graphs with knowledge distillation for video captioning, while STSGR [Geng *et al.*, 2021] propose to adopt a multi-modal transformer guided by scene graph for VideoQA. Moreover, (2.5+1)D-Transformer [Cherian *et al.*, 2022] propose to model the semantic and motion information within video by scene graph, therefore conducting temporal reasoning. However, the above methods construct scene graphs based on the given questions, which limits the model's perception of global information.

To sum up, the major difference between GP-VQA with existing methods is we propose a video-guided temporal modeling approach, which induces the proposed model to utilize broader perception for answer reasoning.

3 Method

The proposed method is illustrated in Figure 2, which consists of two stages: pre-training with masked video graph completion, and prompt-tuning with VideoQA. In the pre-training stage, we adopt standard video SGG model, *e.g.,* STTran [Cong *et al.*, 2021], APT [Nag *et al.*, 2023], and Tempura [Li *et al.*, 2022b], to obtain the raw video scene graphs which abstract spatial-temporal semantic information within the given video. Then, we randomly mask nodes or edges in the raw video scene graph, and force the model to complete the masked parts. In the prompt-tuning stage, we organize the given question into a question graph, and update the feature of each node and edge in question graph by adopting message passing between video graph and question graph. After that, we take the question graph as prompt and feed it into the pre-trained language model with a new QA head to infer the answer of the given question.

3.1 Problem Formulation

GP-VQA consists of pre-training and prompt-tuning stages, where pre-training stage aims to complete the masked video graph, while the prompt-tuning stage aims to generate more accurate question answers. In other words, these two stages have different output, where pre-training stage generates a complete graph, and prompt-tuning stage utilizes classification to predict correct answer. Therefore, GP-VQA needs to address the following issues: *how to design a model to connect these two stages so that the prompt-tuning stage can inherit the temporal modeling knowledge from the pre-training stage?*

To address the above issue, we propose to bridge the pre-training and prompt-tuning stages with graph structure. Given the video $\mathcal{V} = \{I_1, I_2, ..., I_N\}$ and question X_q, VideoQA aims to generate the answer as follows:

$$\tilde{a} = \arg\max_{a \in \mathcal{A}} \mathcal{F}_\theta(a|X_q, \mathcal{V}), \tag{1}$$

where $\tilde{a}$ is the infferred answer in answer space $\mathcal{A}$ and is generated by $\mathcal{F}_\theta$. In contrast, regarding masked video graph completion as a variant of scene graph generation, the completed video scene graph sequence can be denoted as $G = \{g_1, g_2, ..., g_N\}$, where $g_t = \{O_t, E_t\}$ is composed of the objects O_t and there relationships E_t in I_t. To connect VideoQA and masked video graph completion, we formulate the infer-

ence of g_t similar to Eq. 1 as follows:

$$O_t = \arg\max_{o \in \mathcal{O}} \mathcal{F}_\epsilon(o|g_t^{'}, \mathcal{V}),$$
$$E_t = \arg\max_{e \in \mathcal{E}} \mathcal{F}_\epsilon(e|g_t^{'}, \mathcal{V}), \tag{2}$$

where $\mathcal{O}$ and $\mathcal{E}$ are the label space of object and relationship, respectively. We assume a dense graph $g_t^{'}$ with edges between all the nodes and predict the nodes and edges by $\mathcal{F}_\epsilon$. Moreover, by analyzing the semantic dependency of each word in the question X_q with [Manning *et al.*, 2014], we can organize X_q into question graph $g_q = \{O_q, E_q\}$. Thus, the aim of VideoQA can be regarded as predicting a specific node or edge in g_q and Eq. 1 can be reformulated as follows:

$$\tilde{a} = \arg\max_{a \in \mathcal{A}} \mathcal{F}_\theta^*(a|g_q, \mathcal{V}). \tag{3}$$

Since a could be regarded as a node or edge in g_q, it can be observed that if $g_t^{'}$ and g_q are embedded the same, it is possible to handle VideoQA and masked video graph completion with the same function. Therefore, connecting VideoQA and masked video graph completion is transformed into constructing a module or structure that can bridge $g_t^{'}$ and g_q.

3.2 Pre-training with Masked Graph Completion

Similar to recent VideoQA models, we adopt pre-trained feature extractors to capture the visual feature of $\mathcal{V} = \{I_1, I_2, ..., I_N\}$. The extracted feature is denoted as $F_V = \{f_1, f_2, ..., f_N\}$, where $f_t \in \mathbb{R}^{4096}$ is the visual feature of the t-th frame. After that, we adopt Mask-RCNN [He *et al.*, 2017] to detect the object in each frame, the visual features and categories of objects are also provided by the detector, For the i-th object in I_t, we denote the visual feature and predicted category as $v_{t,i} \in \mathbb{R}^{2048}$ and $o_{t,i}$.

With the extracted visual features and detected objects, we adopt the existing approach to obtain the raw video scene graphs. The adopted model is denoted as $\mathcal{F}_d$, and raw video scene graphs is denoted as $\hat{G} = \{\hat{g}_1, \hat{g}_2, ..., \hat{g}_N\}$, where $\hat{g}_t = \{\hat{O}_t, \hat{E}_t\}$ is the obtained raw video scene graph, $\hat{O}_t$ and $\hat{E}_t$ is the objects and the relation edge in I_t, respectively. We take the prediction result of $\mathcal{F}_d$ as a pseudo-label, and $\hat{o}_{t,i} \in \hat{O}_t$ can be represented as follows:

$$\hat{o}_{t,i} = Concate(s_{t,i}, f_{t,i}), \tag{4}$$

where $s_{t,i}$ is pseudo-label embedding provided by Glove [Pennington *et al.*, 2014] and $f_{t,i}$ is the original node representation extracted by $\mathcal{F}_d$. Moreover, the representation construction process of edges is the same as that of nodes. Then we establish hyperedges between scene graphs of different frames based on the temporal order, therefore locating the objects and relationships in the temporal dimension and facilitating subsequent question graph construction. For example, a node will be connected to all the nodes in the scene graphs of the previous frame by a directed hyperedge labeled as "after". Similarly, the directed hyperedge between this node and the nodes in the scene graphs of the subsequent frame is labeled as "before". Thus, we can obtain a video graph $\overline{G}$ to build a comprehensive perception of the video.

Then we construct masked video graph completion based on the $\overline{G}$. We randomly mask some nodes and edges with rate λ thus generating masked scene graph, *e.g.,* $\langle person - write\_on - [mask]\rangle$ and $\langle person - [mask] - paper\rangle$. After that, we employ a trainable language model $\mathcal{F}_l$ with graph completion head $\varphi_{gc}(\cdot)$ to predict a complete video graph, which takes visual feature and $\overline{G}$ as input. Denoting the complete video graph as G, it can be generated as follows:

$$G = \varphi_{gc}(\mathcal{F}_l(\phi(F_V), RM(\overline{G}, \lambda))), \tag{5}$$

where $\phi(\cdot)$ is a visual mapping function for dimension align, $RM(\cdot)$ is random mask operation, and $\mathcal{F}_l$ predicts the masked element in $\overline{G}$ as a cloze task. Specifically, we also follow [Dai *et al.*, 2019] to generate relative temporal position encoding and utilize it to determine temporal order, which is concatenated with each $f_t \in F_V$. Since there are multiple masked nodes or edges in $\overline{G}$, the video graph completion task can naturally guide GP-VQA to expand its perception and acquire stronger video understanding and reasoning abilities.

3.3 Prompt-tuning with VideoQA

In prompt-tuning, we aim to allow the transfer of learned comprehensive video understanding and reasoning capability from video graph completion to VideoQA. Therefore, it is necessary to transform the textual questions in VideoQA into a similar structure to $\overline{G}$, thus enabling the pre-trained $\mathcal{F}_l$ can process VideoQA.

Question Graph construction. To understand the semantic information within both the question and the video, we embed all words in the question and the labels of objects in video graph into 300-dimensional vectors with pre-trained word embeddings (*e.g.,* GloVe [Pennington *et al.*, 2014]). With the same semantic embedding, we can establish a connection between the video graph and the question, which allows the information integration and facilitates the association between video graph and the question. Moreover, we adopt Stanford CoreNLP [Manning *et al.*, 2014] to identify the semantic dependency within the text, thus abstracting the semantic structure of the question into a graph. As shown in Figure. 2, given the question "What did the person hold after drinking from a cup?", we first decompose this sentence into "what", "person", "cup", "hold", "drinking from" and "after". After that, we construct a question graph g_q based on the semantic dependency of these words. Following the construction of the video graph, "person", "cup", and "what" are presented as object nodes in g_q, while "what" with no definite semantic information, and we denote it as a masked node. "hold" and "drinking from" are denoted as spatial relationship edges, while "after" is denoted as temporal hyperedge.

Prompts Generation. Although we have organized the textual question into a question graph, it cannot be directly adopted to calculate the answer since g_q does not involve any visual information for reasoning. Therefore, we capture related visual information and update g_q based on video graph $\hat{G}$ that is generated by a frozen spatial-temporal semantic extractor in prompt-tuning. We first calculate the relevance between nodes in the question graph g_q and the video graph $\hat{G}$ with temporal order. With denote the node in g_q

Methods	obj.-rela.	rela.-act.	obj.-act.	superlative	sequencing	exists	dur. com.	acti. recog.	ALL
most likely	9.39	50.00	50.00	21.01	49.78	50.00	24.27	5.52	10.99
PSAC	37.84	49.95	50.00	33.20	49.78	49.94	45.21	4.14	40.18
HME	37.42	49.90	49.97	33.21	49.77	49.96	47.03	5.43	39.89
HCRN	40.33	49.86	49.85	33.55	49.70	50.01	43.84	5.52	42.11
AIO	48.34	48.99	49.66	37.53	49.61	50.81	45.36	18.97	48.59
SHG-VQA	46.42	60.67	64.63	38.83	62.17	56.06	48.12	10.12	49.20
Temp	50.15	49.76	46.25	39.78	48.25	51.79	49.59	18.96	49.79
MIST-AIO	51.43	54.67	55.37	41.34	53.14	53.49	47.48	**20.18**	50.96
MIST-CLIP	51.68	**67.18**	**68.99**	42.05	**67.24**	**60.33**	**54.62**	19.69	54.39
GF	**54.96**	-	-	**44.62**	53.24	59.13	52.80	14.17	**55.08**
GP-VQA'	**53.03**	**68.51**	**70.63**	**45.21**	**58.05**	**60.26**	**56.20**	**22.20**	**57.90**
GP-VQA*	52.76	69.25	72.04	46.28	68.66	62.39	55.82	25.50	60.43
GP-VQA$^+$	56.74	71.69	73.67	47.51	70.82	65.80	58.25	27.38	61.24

Table 1: QA accuracies of state-of-the-art methods w/o large model on AGQA v2.

as $V_q = \{v_{q,1}, v_{q,2}, ..., v_{q,N_q}\}$, the edge between $v_{q,a}$ and $v_{q,b}$ is defined as $e_{q,ab}$, the relevance weight between a node $v_{q,a} \in g_q$ and $v_{t,i} \in \hat{g}_t$ defined as follows:

$$w^o_{q,a\to t,i} = softmax(s_{q,a} \odot s_{t,i}), \tag{6}$$

where $s_{q,a}$ and $s_{t,i}$ are semantic embeddings generated with Glove [Pennington *et al.*, 2014]. And we also calculate the relevance weight $w^e_{q,ab\to t,ij}$ between each edge $e_{q,ab} \in g_q$ and $e_{t,ij} \in \hat{g}_t$. Finally, we form the similarity between relation triplelet as $w^r_{q,ab\to t,ij} = (w^o_{q,a\to t,i} + w^e_{q,ab\to t,ij} + w^o_{q,b\to t,j})/3$. Then we can update g_q as follows:

1) Triplelet Retrieve: We first retrieve triplets in the video graph that have a similarity higher than threshold α to the masked nodes or edges in the question graph. Then, based on the direction of hyperedges in the question graph, we sequentially search for the corresponding parts of other triplets in the question graph within the video graph, until all the triplets in the question graph have been traversed.

2) Graph Construction: We extract the video subgraphs of frames that correspond to the triplets in the question graph that we retrieved, along with the video subgraphs connected to them by no more than γ hyperedges. Then we establish dense connections between the obtained video subgraphs and the question graph, constructing a cross-modal graph.

3)Feature Update: We apply message passing on the cross-modal graph to update the nodes in the question graph.

In detail, the nodes and edges in g_q is updated as follows:

$$\begin{aligned} h^{(t)}_{v_{q,a}} &= \mathcal{F}_o(h^{(t-1)}_{o_{q,a}} \oplus \sum_{o_{t,i}\in\hat{g}_t} \sigma_o(w^r_{q,ab\to t,c}) h_{o_{t,i}}), \\ h^{(t)}_{e_{q,ab}} &= \mathcal{F}_e(h^{(t)}_{v_{q,a}} \oplus h^{(t)}_{v_{q,b}} \oplus \sum_{e_{t,ij}\in\hat{g}_t} \sigma_e(w^e_{q,ab\to t,c}) h_{e_{t,ij}}), \end{aligned} \tag{7}$$

where we process the updating operation with temporal order and $w^r_{q,ab\to t,c}$ is the similarity between triplelet $\langle a, ab, b\rangle \in g_q$ and the most similar triplet, h^t_* is the representation of question graph node or edge * in t-th frame, h_* is the representation of video graph node or edge *, $\oplus$ is an element-wise add operation, $\mathcal{F}_o$ and $\mathcal{F}_e$ are the learnable liner layer with relu activation. Moreover, we adopt σ_o and σ_e as score functions that adjust the relevance scores of nodes and edges in the video graph with respect to the question graph based on the hop count to the most similar triplet. For each additional hop, the relevance score decays by a factor of β. By passing messages between video graph $\hat{G}$ and question graph g_q, we denote the generated graph as prompt graph g_p which can be processed by the language modle pre-trained in the video graph completion.

Prompt-tuning with VideoQA. In prompt-tuning, we freeze the spatial-temporal semantic extractor and GC head while the language model and graph prompts generator are trainable. We take g_p to replace $\hat{G}$ as the language model input, the answer prediction can be formulated as follows:

$$Ans = \varphi_{qa}(\mathcal{F}_l(\phi(F_V), g_p)), \tag{8}$$

where φ_{qa} is trainable QA head. Furthermore, since there are different types of questions, *e.g.,* causal, temporal, and descriptive in NExT-QA [Xiao *et al.*, 2021], we adopt different classifiers as QA head for different question types. We utilize the standard cross-entropy loss function to optimize the parameters in prompt-tuning, which is denoted as L_{QA}.

4 Experiments

Dataset: We evaluate our model on three recently proposed challenging datasets for the long-form VideoQA, namely AGQA v2 [Grunde-McLaughlin *et al.*, 2022], NExT-QA [Xiao *et al.*, 2021], STAR[Wu *et al.*, 2021]. AGQA v2 is an open-ended VideoQA benchmark for compositional spatial-temporal reasoning, NExT-QA is a benchmark for causal and temporal reasoning, STAR is a benchmark for situated reasoning.

Baselines: We compare our model with several state-of-the-art VideoQA algorithms: SHG-VQA[Urooj *et al.*, 2023], VQA-T[Yang *et al.*, 2021], AIO[Wang *et al.*, 2023], VGT[Xiao *et al.*, 2022], MIST-AIO[Gao *et al.*, 2023], MIST-CLIP[Gao *et al.*, 2023], GF[Bai *et al.*, 2024], ATM[Chen *et al.*, 2023], CoVGT[Xiao *et al.*, 2023], TranSTR[Li *et al.*, 2023d], VideoChat2[Li *et al.*, 2023b], SeVila[Yu *et al.*, 2024], All-in-one[Wang *et al.*, 2023], BLIP-2[Li *et al.*, 2023a], ST-LLM [Liu *et al.*, 2024], TG-Vid [Hu *et al.*, 2024],

	Method	Casual	Temporal	Descriptive	All
w/o LM	HGA	44.22	52.49	44.07	49.74
	CLIP	46.30	39.00	53.10	43.70
	VQA-T	49.60	51.49	63.19	52.32
	AIO	48.04	48.63	63.24	50.60
	Temp	53.10	50.20	66.80	54.30
	VGT	52.28	55.09	64.09	55.02
	MIST-AIO	51.54	51.63	64.16	53.54
	MIST-CLIP	54.62	56.64	66.92	57.18
	GF	**56.93**	**57.07**	**70.53**	**58.83**
	GP-VQA'	**58.94**	**58.65**	**71.20**	**61.85**
w/ LM	TIGV	55.00	56.30	62.90	56.70
	ATM	56.04	58.44	65.38	58.27
	CoVGT	59.69	58.00	69.88	60.73
	TranSTR	59.70	60.20	70.00	61.50
	VideoChat2	64.70	68.70	76.10	68.60
	SeVila	68.10	72.90	81.20	72.60
	SeVila*	75.62	71.83	82.65	74.29
	ST-LLM	74.30	70.00	81.30	74.00
	TG-Vid	**77.40**	73.80	84.30	77.30
	GP-VQA*	77.22	**73.94**	**84.84**	**77.47**
	GP-VQA$^+$	**78.69**	**75.58**	**86.26**	**78.12**

Table 2: QA accuracies of SOTA methods on NExT-QA, where SeVila* is pre-trained with the proposed masked video graph completion and similar data.

ATT-4L[Jaiswal *et al.*, 2024]. We also adopt the proposed method on three video SGG methods: Sttran [Cong *et al.*, 2021], Tempura [Nag *et al.*, 2023], and APT [Li *et al.*, 2022b], to evaluate the flexibility of GP-VQA.

Implementation Details: We employ Mask RCNN [He *et al.*, 2017] with a ResNet-101 backbone as the object detector for adopted video scene graph models. For the vision-language model, we adopt the frozen LLAVA-7B and Qianwen2-VL-7B with LoRA. To ensure fairness, we also employed a standard Transformer [Vaswani *et al.*, 2017] to test the performance of the proposed model. During the pre-training stage, we use SGD optimizer with an initial learning rate of 0.001 and decay the learning rate by multiplying it with 0.9 after every epoch. The momentum is set to 0.9 and the size of mini-batch is set to 8. For hyper-parameters, we set the random mask rate λ to 0.08, while the α, β, and γ in prompts generation are set to 0.7, 0.8, and 2 respectively. Moreover, we sample 1 frame in every 3 frames for pre-training. For prompt-tuning, we use the same setting as pre-training, except the initial learning rate is $1e-5$. Moreover, the adopted video scene graph generation models are pre-trained on Action Genome [Ji *et al.*, 2020]. Since there is no scene graph annotation in NExT-QA, we do not pre-train video graph completion on NExT-QA. Instead, we adopt the module pre-trained on Action Genome and AGQA v2 for NExT-QA.

4.1 Comparison with State-of-the-arts

We compare GP-VQA with the state-of-the-art (SOTA) methods on three VideoQA datasets, as shown in Table 1, Table 2, and Table 3. GP-VQA', GP-VQA* and GP-VQA$^+$ are implemented with Transformer, LLAVA-7B and Qianwen2-VL-7B, respectively. GP-VQA achieves SOTA performances and outperforms the existing methods on all datasets.

Methods	Inte.	Seq.	Pred.	Feas.	All
All-in-One	47.5	50.8	47.7	44.0	47.5
MIST	55.5	54.2	54.2	44.4	51.1
InternVideo	62.7	65.6	54.9	51.9	58.8
BLIP-2^{voting}	52.3	54.8	49.0	51.2	51.8
BLIP-2^{concat}	65.4	69.0	59.7	54.2	62.0
SeViLa	63.7	70.4	63.1	62.4	64.9
Concat-Att-4L	68.1	71.4	66.6	55.2	65.3
Cross-Att-4L	67.5	72.1	64.4	58.5	65.6
GP-VQA'	66.5	70.7	64.4	53.7	65.6
GP-VQA*	**71.8**	**76.2**	**69.7**	**63.9**	**69.9**
GP-VQA+	**73.4**	**77.9**	**71.2**	**64.8**	**72.3**

Table 3: QA accuracies of SOTA methods on STAR.

The proposed method with VLM achieves a 6.16% improvement on the AGQA v2 dataset, especially obtaining 4.51% and 7.69% improvements on the relation-action and action recognition types. Since the SOTA method GF [Bai *et al.*, 2024] does not utilize LLM, to ensure fairness, we also evaluate GP-VQA with a standard Transformer, which still achieved a 2.82% improvement over the GF. As shown in Table 2, the performance improvement of the proposed method is lower on NExT-QA because we cannot perform masked video graph completion pre-training on NExT-QA, but it is worth noting that our model still has significant superiority and achieves the best performance. Moreover, it can be observed that GP-VQA exhibits a significant advantage in handling causal-type questions, with a 10.59% improvement compared to SeVila. Since GP-VQA utilizes additional scene graph data from Action Genome [Ji *et al.*, 2020], we implement a variant SiVila* of SiVila to ensure fairness. SiVila* also adopts masked video graph completion pre-training with the additional data. As shown in Table 2, pre-training effectively improved the answer reasoning ability of SiVila*, while the proposed model GP-VQA$^+$ still outperformed SiVila*. Furthermore, GP-VQA obtains obvious improvement on each metric of the STAR dataset, as shown in Table 3. This demonstrates the effectiveness of the proposed masked video graph completion pre-training in enhancing reasoning performance.

4.2 Ablation Study

In this part, we analyze the impacts of the designed pre-training & prompt-tuning strategy, graph prompt generation, and hyper-parameters in our model.

Effect of each component in GP-VQA: We ablate key components in GP-VQA, *e.g.,* Pre-training and Random Mask, denoted as, w/o PT and w/o RM.

- w/o PT: It removes the masked video graph completion pre-training in VideoQA, which we directly adopt video graph generated by frozen video SGG model.
- w/o RM: It removes the Random Mask in masked video graph completion pre-training, which vision-language model server as a scene graph generator in pre-training.

		(a)		(b)		(c)		
	Question Types	w/o RM	w/o PT	mean pooling	max pooling	Sttran*	Tempura*	GP-VQA
AGQA v2	object-relationship	53.69	53.49	50.96	53.03	56.11	57.14	56.74
	relationship-action	69.49	66.51	69.67	69.23	71.11	72.14	71.69
	object-action	67.93	71.15	69.72	70.63	73.73	74.53	73.67
	superlative	41.73	44.36	44.34	45.22	47.05	47.55	47.51
	sequencing	65.85	66.72	68.63	68.06	71.30	70.76	70.82
	exists	61.88	62.64	61.81	62.27	66.79	65.97	65.80
	duration comparison	53.41	55.18	53.21	56.21	58.96	58.95	58.25
	activity recognition	23.34	24.83	24.57	22.21	25.45	25.66	27.38
	ALL	58.76	56.58	57.11	57.91	59.87 +1.04	60.33+1.50	61.24+2.41
NExT-QA	Causal	75.74	74.35	74.88	74.35	77.17	77.57	78.69
	Temporal	72.98	73.31	70.95	68.53	74.62	73.69	75.58
	Descriptive	79.89	80.34	77.87	80.23	81.84	82.68	86.26
	All	73.40	72.87	70.96	71.60	75.57 +2.94	76.60+4.00	78.12+5.79

Table 4: Effect of (a) each component and (b) different strategies for prompts generation in GP-VQA. (c) Ablation Study of adopting different video SGG methods. The red numbers indicate the improvement compared with the previous SOTA methods.

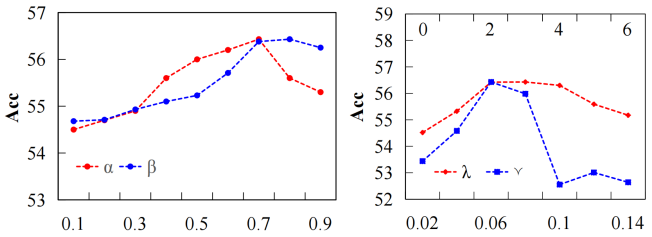

Figure 3: Effect of different parameters on AGQA v2.

The results of these variants on AGQA v2 and NExT-QA are shown in Tableb 4 (a). We can see that removing masked video graph completion pre-training causes a 4.66% and 4.25% accuracy drop on AGQA v2 and NExT-QA datasets, respectively. It proves that the proposed pre-training strategy can effectively enhance the comprehensive understanding ability of videos, thereby obtaining strong reasoning ability for question answering. In addition, the performance drop on both datasets is significant when removing pre-training, which also demonstrates the learned reasoning ability is universal in VideoQA. Tableb 4 also highlights the importance of random mask strategy, the full model GP-VQA obtains the best results on both datasets, which can be attributed to the random mask operation indeed helps the training of model's reasoning ability. Moreover, as it simulates the form of question answering in the masked video graph completion, the proposed model can more easily adapt to prompt-tuning.

Effect of different strategies for prompts generation: We compare three types of prompts generation strategies: mean pooling, max pooling, and learned layer (GP-VQA), which can be realized by modifying $\mathcal{F}_v$ and $\mathcal{F}_e$ in Eq. 7. As shown in Table 4 (b), graph prompts achieve the best performance, *e.g.,* 4.13% and 3.33% improvements compared with mean and max pooling on the AGQA v2 dataset. The reason is that a learned layer can effectively capture relevant information in video graph.

Effect of different video SGG models: To evaluate the flexibility of GP-VQA in incorporating different methods for obtaining raw video graphs, we conduct several experiments on different video scene graph generation methods, *e.g.,* Sttran*, Tempura*, and APT(GP-VQA). We conduct different video scene graph generation methods to replace the spatial-temporal semantic extractor in GP-VQA, as shown in Table 4 (c). We can see that all these methods outperform the previous SOTA method SiVila [Yu *et al.*, 2024] and GF [Bai *et al.*, 2024] on both AGQA v2 and NeXT-QA datasets. The improvement of performance further demonstrates the flexibility of the proposed GP-VQA, and verifies the advantage of introducing video graph for extending the proposed model's perception.

Effect of different hyperparameters: We analyze the performance with different random mask rates λ, and hyperparameters α, β and γ in prompts generation. As shown in Figure 3, it can be observed that accuracy first increases and then slowly decreases or level off. Thus, we set $\lambda = 0.08, \alpha = 0.7, \beta = 0.8$ and $\gamma = 2$.

5 Conclusion

In this work, we present a video-guided temporal modeling approach for VideoQA, comprising pre-training on masked video graph completion and prompt-tuning for VideoQA. The method uses random masking in pre-training for graph construction, leveraging video scene graphs and a graph-completion-equipped language model for multimodal reasoning. During prompt-tuning, question graphs are built, messages are passed between video and question graphs to create prompts, and answers are computed using the pre-trained model. This strategy enhances visual scene perception and overcomes text-question limitations. We conduct extensive experiments to show that the proposed method significantly outperforms the SOTA methods, with adaptable components such as the video SGG. Future exploration will focus on semi-supervised or unsupervised graph-based prompt-tuning, aiming for practical real-world adaptation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants (62325206, U23A20387, 62322212), the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, the CAS Project for Young Scientists in Basic Research (YSBR-116), the Opening Foundation of Artificial Intelligence Key Laboratory of Sichuan Province(2024RYY01), and the Natural Science Research Startup Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications under Grant NY224066.

References

[Bai *et al.*, 2024] Ziyi Bai, Ruiping Wang, and Xilin Chen. Glance and focus: Memory prompting for multi-event video question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

[Bärmann and Waibel, 2022] Leonard Bärmann and Alex Waibel. Where did i leave my keys? - episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1560–1568, June 2022.

[Chen *et al.*, 2023] Junwen Chen, Jie Zhu, and Yu Kong. Atm: Action temporality modeling for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4886–4895, 2023.

[Cherian *et al.*, 2022] Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. (2.5+ 1) d spatio-temporal scene graphs for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 444–453, 2022.

[Cong *et al.*, 2021] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, pages 16372–16382, 2021.

[Dai *et al.*, 2019] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[Datta *et al.*, 2022] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19119–19128, June 2022.

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

[Geng *et al.*, 2021] Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1415–1423, 2021.

[Grauman *et al.*, 2022] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[Grunde-McLaughlin *et al.*, 2022] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa 2.0: An updated benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2204.06105*, 2022.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[Hu *et al.*, 2024] Zi-Yuan Hu, Yiwu Zhong, Shijia Huang, Michael R. Lyu, and Liwei Wang. Enhancing temporal modeling of video llms via time gating, 2024.

[Huang *et al.*, 2020] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028, 2020.

[Jaiswal *et al.*, 2024] Shantanu Jaiswal, Debaditya Roy, Basura Fernando, and Cheston Tan. Learning to reason iteratively and parallelly for complex visual reasoning scenarios, 2024.

[Jang *et al.*, 2017] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[Ji *et al.*, 2020] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020.

[Lei *et al.*, 2021] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 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*, pages 7331–7341, 2021.

[Li *et al.*, 2022a] Yiming Li, Xiaoshan Yang, Xuhui Huang, Zhe Ma, and Changsheng Xu. Zero-shot predicate prediction for scene graph parsing. *IEEE Transactions on Multimedia*, 25:3140–3153, 2022.

[Li *et al.*, 2022b] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2023b] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023.

[Li *et al.*, 2023c] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Transformer-empowered invariant grounding for video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Li *et al.*, 2023d] Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878, 2023.

[Li *et al.*, 2023e] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Iterative learning with extra and inner knowledge for long-tail dynamic scene graph generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4707–4715, 2023.

[Liu *et al.*, 2021] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. Hair: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1698–1707, 2021.

[Liu *et al.*, 2024] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners, 2024.

[Manning *et al.*, 2014] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[Nag *et al.*, 2023] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy Chowdhury. Unbiased scene graph generation in videos. *arXiv preprint arXiv:2304.00733*, 2023.

[Pan *et al.*, 2020] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020.

[Park *et al.*, 2021] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 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*, pages 15526–15535, 2021.

[Peng *et al.*, 2021] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2871–2879, 2021.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Shi *et al.*, 2019] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8376–8384, 2019.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10433, 2019.

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2021] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 24:3369–3380, 2021.

[Wang *et al.*, 2023] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023.

[Wong *et al.*, 2022] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, pages 485–501. Springer, 2022.

[Wu *et al.*, 2021] Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[Xiao *et al.*, 2021] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[Xiao *et al.*, 2022] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022.

[Xiao *et al.*, 2023] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[Yang *et al.*, 2021] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[Yu *et al.*, 2024] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yuan *et al.*, 2023] Bowen Yuan, Sisi You, and Bing-Kun Bao. Self-pt: Adaptive self-prompt tuning for low-resource visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5089–5098, 2023.

[Zeng *et al.*, 2017] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[Zong *et al.*, 2024] Linlin Zong, Jiahui Wan, Xianchao Zhang, Xinyue Liu, Wenxin Liang, and Bo Xu. Video-context aligned transformer for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19795–19803, 2024.