Dynamic Multistep Reasoning based on Video Scene Graph for Video Question Answering

Jianguo Mao[1,2]*, **Wenbin Jiang**[3], **Xiangdong Wang**[1], **Zhifan Feng**[3], **Yajuan Lyu**[3], **Hong Liu**[1], **Yong Zhu**[3]

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Baidu Inc., Beijing, China

{maojianguo20s,xdwang,hliu}@ict.ac.cn

{jiangwenbin,fengzhifan,lvyajuan,zhuyong}@baidu.com

Abstract

Existing video question answering (video QA) models lack the capacity for deep video understanding and flexible multistep reasoning. We propose for video QA a novel model which performs dynamic multistep reasoning between questions and videos. It creates video semantic representation based on the video scene graph composed of semantic elements of the video and semantic relations among these elements. Then, it performs multistep reasoning for better answer decision between the representations of the question and the video, and dynamically integrate the reasoning results. Experiments show the significant advantage of the proposed model against previous methods in accuracy and interpretability. Against the existing state-of-the-art model, the proposed model dramatically improves more than $4\%/3.1\%/2\%$ on the three widely used video QA datasets, MSRVTT-QA, MSRVTT multi-choice, and TGIF-QA, and displays better interpretability by backtracing along with the attention mechanisms to the video scene graphs.

1 Introduction

Video question answering (video QA) aims to answer questions according to the given videos. It is usually defined as a classification task, where the most appropriate answer is chosen from a candidate list for the given question and video. Existing methods for video QA conduct direct answering selection based on the multimodal encoding of questions and videos (Jang et al., 2017; Lei et al., 2018, 2020). In recent years, researchers have proposed many optimization strategies for better performance in video question answering, e.g., designing delicate encoding mechanisms (Kim et al., 2020a; Nuamah, 2021; Gao et al., 2018; Li et al., 2019; Fan et al., 2019; Le et al., 2020; Jiang et al., 2020; Kim et al., 2020b; Seo et al., 2021), introducing video scene graphs (Garcia and Nakashima, 2020), adopting video pre-trained language models (Li et al., 2020; Zellers et al., 2021; Li and Wang, 2020; Lei et al., 2021; Sun et al., 2019), and leveraging external knowledge or resources (Chadha et al., 2020; Garcia et al., 2020; Liu et al., 2020b; Song et al., 2021; Garcia and Nakashima, 2020). Compared with conventional monomodal question answering tasks such as text QA (Oguz et al., 2021; Zhou et al., 2018; Lin et al., 2018) and table QA (Cao et al., 2021; Wang et al., 2019). Video QA is more difficult due to the need for crossmodal understanding and reasoning of the video and the question. Existing methods are mainly concerned with how to encode the crossmodal features better. When faced with a complex question, they usually lack the abilities of deep understanding and complex reasoning.

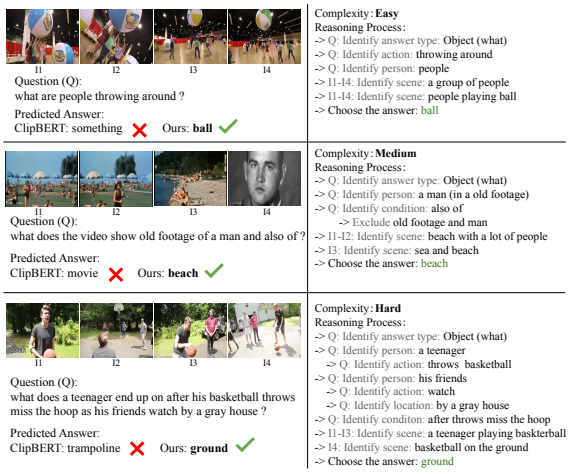

Figure 1: Error cases of the previous state-of-the-art method (Lei et al., 2021) that needed multistep reasoning. We demonstrate the multistep reasoning process based on question and video(left).

Similar to the situations in text QA and table QA, it is also necessary for video QA to deeply understand the semantics of the context, namely, the video, and the reasoning on the context and the question. Statistical analysis on several datasets

*This is joint work of CASICT and Baidu.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3894 - 3904

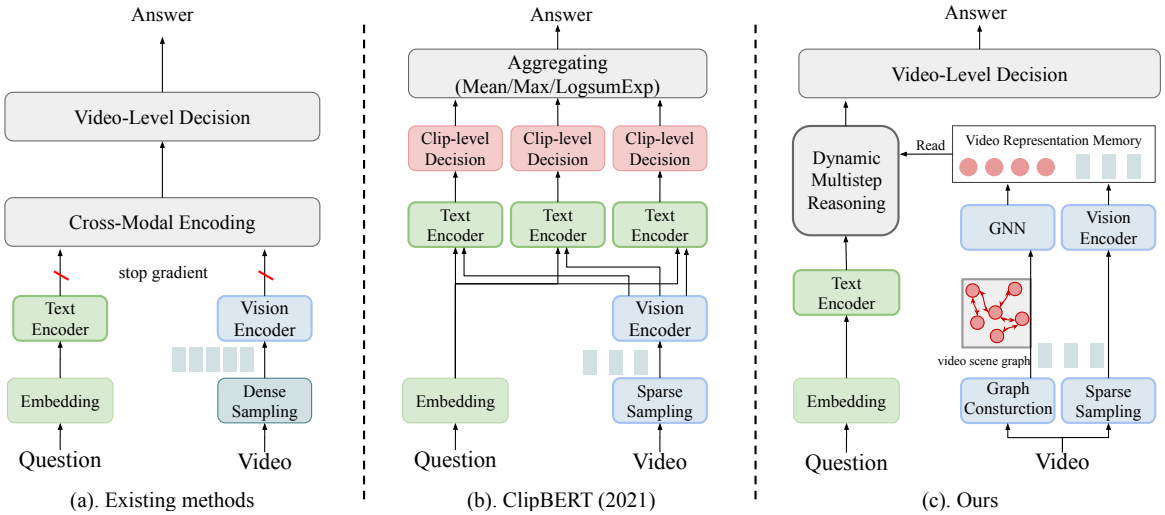

Figure 2: Comparison of our model architecture with (a). popular video-and-language learning paradigm and (b).ClipBERT. In contrast to most previous methods that adopt simple and implicit reasoning mechanisms, Our method uses video scene graphs to represent the video semantic information and adopt a dynamic multistep reasoning mechanism.

reveals that a significant percentage of failed cases of the state-of-the-art (SOTA) model ClipBERT (Lei et al., 2021) is caused by the lack of deep understanding and reasoning. Figure 1 shows some error cases of the SOTA model on MSRVTT-QA. We randomly select one hundered error cases and find that about 24% of the error cases need multiple steps of reasoning on the question and the video, and about 15% of the error cases need a deep understanding of the semantic of the video. These cases could be solved by a model emphasizing deep understanding and reasoning. The questions in these three datasets are relatively simple according to the building procedures (Xu et al., 2017a; Yu et al., 2018a; Jang et al., 2017). It will be more valuable for video QA to enable deep understanding and reasoning on the question and the video in realistic application scenarios.

In this work, we propose a novel dynamic reasoning model for video QA to overcome the weakness of previous models in deep understanding and reasoning. It first creates the video semantic representation from the video scene graph, which is composed of the semantic elements of the video and the semantic relations between these elements. Then it conducts multistep reasoning of the question based on the video semantic representation to generate a series of video-aware question representations. Finally, it generates the most appropriate question representation for the final answering decision by dynamically integrating these video-aware question representations according to the reasoning complicity prediction. Figure 2 shows the overall architecture of the proposed model and the comparison with previous methods. It simulates the reasoning procedure of human beings, while previous methods follow the pipeline of multimodal encoding and answering selection. In addition, the proposed model enables the decomposition of question understanding and video understanding, thus leading to more opportunities for future optimization. On the one hand, more external knowledge resources and better reasoning architectures can be introduced for better video QA performance. On the other hand, it can act as a unified framework for different QA tasks such as video QA, table QA, and text QA.

We verify the proposed model on three well-known datasets, MSRVTT-QA (Xu et al., 2017a), MSRVTT multi-choice (Yu et al., 2018a), and TGIF-QA (Jang et al., 2017), widely used in recent video QA works (Jang et al., 2017; Gao et al., 2018; Li et al., 2019; Fan et al., 2019; Le et al., 2020; Zhu and Yang, 2020; Lei et al., 2021; Seo et al., 2021). Experiments show that our model achieves dramatic improvement over the powerful state-of-the-art model ClipBERT (Lei et al., 2021), with an average accuracy increment of more than 3 percentage points. Ablation studies show that the dynamic reasoning strategy significantly outperforms previous implicit simple reasoning strategies. The video semantic representation based on the video scene graph makes the dynamic reasoning strategy work better. Backtracing along with the at-

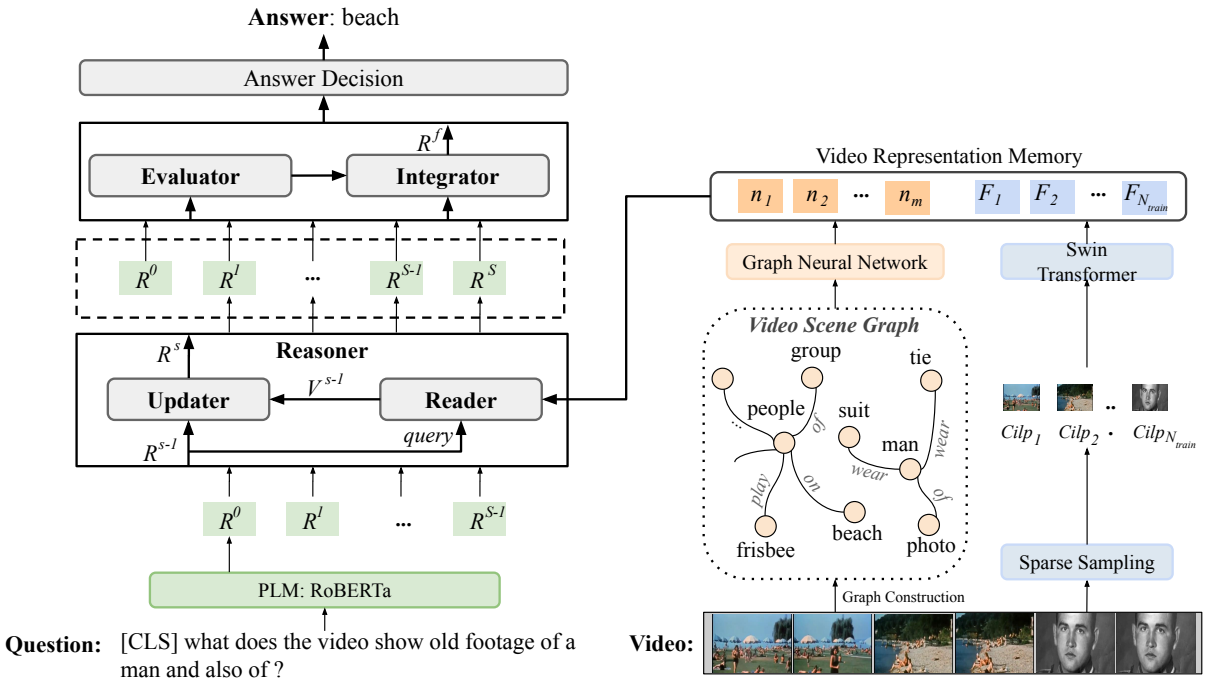

Figure 3: Overview of our model architecture. It contains two part: 3.2. Video Representation Learning (right) and 3.3. Dynamic Multistep Reasoning (left)

.

tention mechanism to the video scene graph clearly shows the semantic elements that answer decision relies on each reasoning step, thus giving better interpretability than most of the previous methods.

2 Background

Given the question q, video question answering requires choose the correct answer $\hat{a}$ from the candidates set Ω_a according to the video content V.

$$\hat{a} = \underset{a \in \Omega_a}{\text{argmax}} p(a|q, V; \theta) \quad (1)$$

As Figure 2 (a) shows, most existing works utilize offline (stop gradient) extracted dense video features and text features. As Figure 2 (b) shows, ClipBERT (Lei et al., 2021) achieves the state-of-the-art by using sparsely sampled clips and raw text for end-to-end training, yet suffer from two main drawbacks: (i) Lack of a deep understanding of the video content and reasoning on the question based on the video. (ii) Strong coupling between video and question modeling process, which needs additional image-text pair data for pre-training to enhance the ability of the text encoder to model multimodal features and leads to poor scalability and repeated computation. As shown in Figure 2 (c), we propose a simple but effective architecture to solve the weakness of previous works. It first decouples the question understanding module and the video understanding module. Then, we introduce the video scene graph to get a better representations of the video semantic information, which can enhance the understanding of the video content and its graph structure is also better for reasoning. At last, we design a dynamic multistep reasoning mechanism to iteratively deepen the understanding of the question according to the video content.

3 Method

3.1 Overall Architecture

Figure 3 gives an overview of the model architecture. For the visual representation, we use both video scene graphs and image features. On the one hand, we construct a video scene graph to represent semantic information in a structural form. On the other hand, we use Swin Transformer (Liu et al., 2021) to extract image features to make up for the missing information of the scene graph. The reasoning module (***Reasoner***) will iteratively updates the understanding of the question according to the video content. The ***Evaluator*** will decide the number of reasoning steps according to the complexity of the question. The ***Integrator*** will integrate all intermediate reasoning results to get the final reasoning results. The answer decision

module chooses an answer according to the final comprehensive understanding of the question.

3.2 Video Representation Learning

We chose a structured video scene graph to describe video semantics which is better for reasoning. We also extract image features by Swin Transformer (Liu et al., 2021) to make up for the missing information in the scene graph. The video scene graph and the image features constitute the *Video Representation Memory* shown in Figure 3, a memory for the ***Reader*** module to access.

3.2.1 Video Scene Graph

The video scene graph is the basis for conducting dynamic reasoning, it is a graph-based semantic representation of video content, representing the objects in the video, their attributes, and their relationships in a structured form. Unlike the image scene graph commonly used in visual question answering, our video scene graph is semantically richer and contains spatio-temporal information of the video. Specifically, We first use an image captioning model to generate captions for each clip. Then we use the scene parser (Schuster et al., 2015) to convert each caption sentence into a semantic sub-graph and integrate the same nodes of each sub-graph to obtain a video scene graph. Compared to caption sentences, the video scene graph represents the video-level semantic information in a structured form, better modeling the visual semantic information.

Graph Representation Learning We first obtain the embeddings of nodes $\mathbf{n} = \{n_1, n_2, ..., n_m\}$ and edges in the video scene graph via a parameter-sharing language encoder. Then, we use graph attention neural network (Veličković et al., 2017) iteratively to update the representation of the scene graph. Each node updates its representation based on the correlation with its neighbor nodes.

$$\mathbf{n}_i^{'} = a_{ii}\mathbf{W}\mathbf{n}_i + \sum_{j \in \mathcal{N}_i} a_{ij}\mathbf{W}\mathbf{n}_j \tag{2}$$

where $\mathbf{W}$ is a weight matrix, a_{ij} is the attention weight of node $\mathbf{n}_i$ and $\mathbf{n}_j$, and $\mathcal{N}_i$ is the neighbors of the node $\mathbf{n_i}$ in the graph. In our experiments, we use standard graph attention neural network setting, applying the LeakyReLU nonlinearity (with negative input slope $\alpha = 0.2$). At the same time, the edges have explicit meanings of relations between nodes, so we also consider edges features $\mathbf{e}_{ij}$ when calculating attention weight. The attention weight a_{ij} are computed as

$$a_{ij} = \frac{\exp(\text{LeakyReLU}(\text{A}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{e}_{ij})))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\text{A}(\mathbf{n}_i, \mathbf{n}_k, \mathbf{e}_{ik})))} \tag{3}$$

$$\text{A}(\mathbf{n}_i, \mathbf{n}_j, \mathbf{e}_{ij}) = \mathbf{W_a}^\top [\mathbf{W}\mathbf{n}_i || \mathbf{W}\mathbf{n}_j || \mathbf{W}_e\mathbf{e}_{ij}] \tag{4}$$

where $.^\top$ represents transposition, $||$ is the concatenation operation and the $\mathbf{W_a}$ and $\mathbf{W_e}$ are weight matrices.

3.2.2 Image Features

We extract image features to make up for the missing information in the scene graph. First, we use Lei et al.'s sparse sampling method to sparsely and randomly sample N_{train} clips $\{c_i\}_{i=1}^{N_{train}}$ from video. N_{train} is typically much smaller than the entire video length N. This sampling method can reduce the computation cost and obtain better performance than dense sampling. For inference, we uniformly sample N_{test} clips of the same duration. Swin Transformer (Liu et al., 2021) is one of the mainstream visual backbone networks. It alleviates the problem of large variations in the scale of visual entities and the high resolution of pixels in images. We use it as a vision encoder $\mathcal{E}_v$ to extract clip features $\{F_i\}_{i=1}^{N_{train}}, F_i = \mathcal{E}_v(c_i) \in \mathbb{R}^{w \times w \times d}$, where w is the window size and d is the feature dimension.

3.3 Dynamic Multistep Reasoning

Humans will deepen their understanding of a complex question through repeated reading the context information. The more complex the question, the more repetitions are required (Chang and Millett, 2013; Gorsuch and Taguchi, 2008; Carver and Hoffman, 1981). At each step of reading, people will focus on different parts of the context information. Inspired by it, we designed the dynamic multistep reasoning mechanism. It will iteratively update the understanding of the question based on the video representations. We first extract the question representation by language model RoBERTa (Liu et al., 2019). Specifically, we concatenate the question text with a special token [CLS] as the input and take the [CLS]'s hidden state R^0 as the representation of the question. At the first reasoning step, we select R^0 as the input. Then, we get question-related information from Video Representation Memory

through the ***Reader***, which is an attention mechanism. Then, we use this retrieved video information to update the understanding of the question and get the first reasoning step result R_1 through the ***Updater***. At the next reasoning step, we select R_1 as the input. After S steps of reasoning, we obtain all results $\mathbf{R} = \{R^0, R^1, R^2, ..., R^S\}$.

$$R^s = \text{Updater}(R^{s-1}, V^{s-1}) \quad (5)$$

Updater consists of two linear transformations with a ReLU activation in between.

$$\text{Updater}(Q, K) = Q + \text{ReLU}(\mathbf{W_1}K + b_1)\mathbf{W_2} + b_2 \quad (6)$$

where $\mathbf{W_1}$ and $\mathbf{W_2}$ are weight matrices and b_1 and b_2 are biases. V_{s-1} represents question-related video information.

$$V^{s-1} = \text{Reader}(R^{s-1}, \mathbf{Vid}) \quad (7)$$

where $\mathbf{Vid}$ consists of the node features $\{n_1, n_2, ..., n_m\}$ of the video scene graph and image features $\{F_1, F_2, ..., F_{N_{train}}\}$. We use Scaled Dot-Product Attention (Vaswani et al., 2017) as the ***Reader***. The input consists of query Q, and context K of dimension d_k.

$$\text{Reader}(Q, K) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})K \quad (8)$$

The word *dynamic* has two meanings: (i). dynamically decide the number of reasoning steps according to the question's complexity. ***(3.3 Evaluator)***. (ii). dynamically integrate the results of all reasoning steps as the final result. ***(3.3 Integrator)***.

Evaluator When humans begin faced with different complexity questions, they will dynamically adjust the number of times to read relevant information (Chang and Millett, 2013; Gorsuch and Taguchi, 2008; Carver and Hoffman, 1981). We propose the first dynamic reasoning strategy by imitating the human reading and understanding mechanisms. It will decides the number of reasoning steps according to the complexity of the question. Specifically, we perform a nonlinear transformation with GumbelSoftmax as activation function on the question representation R^0, and output an S-dimensional vector $\text{D}(R_0) \in \mathbb{R}^{1\times \text{S}}$ to represents the distribution probability of the number of reasoning steps from 1 to S.

$$\text{D}(R_0) = \text{GumbelSoftmax}(\mathbf{W_d}R^0 + b) \quad (9)$$

Where $\mathbf{W_d}$ is a weight matrix, and b is a bias. We choose the one with the greatest probability as the number of the reasoning steps S.

Integrator In the reasoning process, the ***Reader*** pays attention to the different parts of the video content at each step to gradually deepen the question's understanding. Therefore, we think the intermediate reasoning results are also helpful in choosing an answer. After S steps of reasoning, we select all intermediate reasoning results as input and perform a nonlinear transformation with softmax function to calculate the distribution of the weight of $\mathbf{R}$ and get the weighted sum as the final reasoning results R^f.

$$R^f = \text{softmax}(\mathbf{W_f}\mathbf{R} + \mathbf{b})\mathbf{R} \quad (10)$$

where $\mathbf{W_f}$ is a weight matrix and b is a bias.

3.4 Answer Decision

Given the final reasoning results R^f, we use two fully-connected layers as a classifier to obtain the logits l_f for the answer options. Then we use a softmax function to obtain the probability distribution of each answer option and apply cross-entropy loss as our model loss $\mathcal{L}$.

$$l_f = \text{classifier}(R^f) \quad (11)$$

$$\hat{y} = \text{softmax}(l_f), \ \mathcal{L} = -\sum_{i=1}^{M} y_i log\hat{y}_i \quad (12)$$

4 Related Work

Video QA requires fine-grained modeling of multimodal features. We have witnessed many efforts devoted to video understanding for video QA. Some methods use visual techniques such as object detection (Ren et al., 2016) and image captioning (Johnson et al., 2016; Rennie et al., 2017) to extract additional visual information (Kim et al., 2020a). In recent years, visual pre-training based on large-scale data has become a popular method to improve video applications including video QA (Li et al., 2020; Zellers et al., 2021; Li and Wang, 2020; Sun et al., 2019). In addition, several advanced techniques such as contrastive self-supervised learning (Kim et al., 2020b) and symbolized video scene graph (Garcia and Nakashima, 2020) are proposed to improve the performance of video understanding

for video QA. We extract video semantic representations based on both the visual pre-training model (Liu et al., 2021) and the video scene graphs. Experiments show the amazing complementarity of the two kinds of information.

Video QA also requires flexible reasoning for complicated questions and videos. Although the reasoning capability is rarely emphasized in previous work for video QA, it is broadly investigated in other QA tasks (Clark et al., 2018). For example, text QA resorts to iterative update of the question and the context (Das et al., 2019; Liu et al., 2020a), table QA generates a structural query such as SQL which is then executed on the tabular data (Guo et al., 2019), and visual QA builds a specific module network according to the question and runs it on the image (Andreas et al., 2016; Cao et al., 2018; Hu et al., 2017). The overall architecture of the proposed method is similar to the iterative reasoning strategies for text QA but with significant innovation. Our model can dynamically determine the best integration strategy for the intermediate reasoning results according to the given question and video, leading to better interpretability and much better performance.

5 Experiments

In this section, we validate our method on three mainstream video QA datasets. We conduct comparison experiments with previous works and perform ablation experiments to analyze the critical improvement in our proposed method. We use standard train/val/test splits for all datasets and use accuracy to measure the performance. All experimental results are the mean and standard deviation of ten replicate experiments.

5.1 Datasets

MSRVTT-QA MSR-VTT (Xu et al., 2016) is a large video description dataset. It provides 10k web video clips with 41.1 hours and 200k clip-sentence pairs in total. MSRVTT-QA (Xu et al., 2017a) is created based on clip-sentence pairs in MSR-VTT automatically through a program. It contains 243k open-ended questions with 1500 answers.

MSRVTT-MC MSRVTT-MC (multiple-choice) (Yu et al., 2018a) is a dataset for video-text matching tasks built on MSR-VTT with videos are used as queries, captions as answers. Each video contains five captions. Only one is correct.

TGIF-QA TGIF-QA (Jang et al., 2017) dataset contains 165K QA pairs for the animated GIFs from the TGIF dataset. We experiment on 3 TGIF-QA tasks: Repeating Action and State Transition for multiple-choice QA and Frame QA for open-ended QA. We follow most previous works and ClipBERT's (Lei et al., 2021) approach to leave the Count task as future work as it requires directly modeling full-length videos.

5.2 Results and Analysis

5.2.1 Comparison with existing approaches

As shown in Table 1, our method reaches the new state-of-the-art and achieves $\mathbf{41.6\%/91.4\%/84.6\%, 90.1\%, 62.5\%}$ accuracy on MSRVTT-QA/MSRVTT-MC (multi-choice) /TGIF-QA (Action, Transition, FrameQA), with $\mathbf{4.2\%/3.2\%/1.8\%, 2.3\%, 2.2\%}$ improvement over the previous state-of-the-art method ClipBERT (Lei et al., 2021). The *Pre-training data* column represents that the model was pre-trained with additional data. The results show that our method achieves the best performance on all three video question answering datasets without using additional data.

5.2.2 Ablations Analysis

Comparsion of Different Architectures We compare the three architectures shown in Figure 2, namely, the widely-adopted architecture based on Cross-Modal Encoding, ClipBERT, and the architecture proposed in this paper. For all architectures, we use RoBERTa (Liu et al., 2019) as a language encoder and Swin Transformer (Liu et al., 2021) as a vision encoder. We sparsely sample 8 clips from each video, then uniformly sample a single frame within each clip. In addition, we remove the video scene graph from our model and only conduct single-step reasoning to ensure fairness. Other hyper-parameters are the same for all models. The results are given in Table 2, which show that our architecture achieves the best performance, in spite of the removing of the video scene graph and only conducting single-step reasoning. This may because that the proposed architecture makes decisions based on the global video information, unlike ClipBERT (Lei et al., 2021) and most other methods, which integrates the decision made by each clip to get the final decision. Furthermore, unlike previous works, we make a more apparent distinction between the process of question understanding, video understanding, reasoning, and answer deci-

Method	Pre-training data	MSRVTT		TGIF-QA		
		QA	MC	Action	Transition	FrameQA
SNUVL (Yu et al., 2016) (by Yu et al.)	-	-	65.4	-	-	-
ST-VQA (Jang et al., 2017) (by Fan et al.)	-	30.9	66.1	60.8	67.1	49.3
CT-SAN (Yu et al., 2017) (by Yu et al.)	-	-	66.4	-	-	-
MLB (Kim et al., 2016) (by Yu et al.)	-	-	76.1	-	-	-
JSFusion (Yu et al., 2018b)	-	-	83.4	-	-	-
ActBERT (Zhu and Yang, 2020)	HowTo100M	-	85.7	-	-	-
Co-Memory (Gao et al., 2018) (by Fan et al.)	-	32.0	-	68.2	67.1	49.3
AMU (Xu et al., 2017b)	-	32.5	-	-	-	-
PSAC (Li et al., 2019)	-	-	-	70.4	76.9	55.7
Heterogenous Memory (Fan et al., 2019)	-	33.0	-	73.9	77.8	53.8
QuesST (Jiang et al., 2020)	-	-	-	75.9	81.0	59.7
HCRN (Le et al., 2020)	-	35.6	-	75.0	81.4	55.9
MASN (Seo et al., 2021)	-	35.2	-	84.4	87.4	59.5
VQA-T (Yang et al., 2021)	-	39.6	-	-	-	-
VQA-T (Yang et al., 2021)	HowTo100M	40.4	-	-	-	-
VQA-T (Yang et al., 2021)	HowToVQA69M	41.5	-	-	-	-
HGQA (Xiao et al., 2021)	-	38.6	-	76.9	85.6	61.3
CLIPBERT (Lei et al., 2021)	COCO & VG	37.4	88.2	82.8	87.8	60.3
Ours	-	**41.6**	**91.4**	**84.6**	**90.1**	**62.5**

Table 1: Comparison with state-of-the-art methods on video question answering. We verified performance on standard test sets of three datasets. The evaluation metric is accuracy. It is worth specifying that ActBERT, VQA-T, and CLIPBERT use additional large-scale data for pre-training. The results show that our method achieves the best performance without the use of additional pre-training data.

Architecture	#params	MSRVTT (Acc.)	
		QA	MC
Existing methods	257M	35.2 ± 0.32	85.4 ± 0.36
ClipBERT	215M	37.8 ± 0.28	88.7 ± 0.29
Ours*	222M	$\mathbf{38.5 \pm 0.14}$	$\mathbf{89.1 \pm 0.16}$

Table 2: **Our proposed architecture vs. Previous mainstream architectures.** All architectures use the same language and vision encoder. For the architecture of the existing methods, we use standard Transformer Encoder as the module of Cross-Modal Encoding. For fairness, we remove the video scene graph from our model and only do single-step reasoning (Ours*). The rest of the hyper-parameters are the same as each other.

Architecture	Visual Feat.	MSRVTT (Acc.)	
		QA	MC
Existing methods	Image	35.2 ± 0.32	85.1 ± 0.36
	Image + Caption	36.3 ± 0.26	85.9 ± 0.24
	Image + Scene Graph	37.1 ± 0.16	86.4 ± 0.18
ClipBERT	Image	37.8 ± 0.28	88.7 ± 0.29
	Image + Caption	38.4 ± 0.22	89.2 ± 0.19
	Image + Scene Graph	39.2 ± 0.21	89.8 ± 0.18
Ours	Image	38.5 ± 0.14	89.1 ± 0.16
	Image + Caption	39.0 ± 0.12	89.4 ± 0.11
	Image + Scene Graph	$\mathbf{39.5 \pm 0.14}$	$\mathbf{90.2 \pm 0.11}$

Table 3: **Impact of the video scene graph.**

sion, providing a basis for introducing video scene graphs and dynamic multistep reasoning.

Analysis of Video Scene Graph We introduce the video scene graph to get a structural visual semantic representation which is better for reasoning. We use all architectures shown in Figure 2 as the benchmarks to evaluate the effect of the video scene graph. We use an image captioning model to extract captions for each clip. Then we use the scene parser (Schuster et al., 2015) to convert each caption sentence into a semantic sub-graph and integrate the same nodes of each sub-graph to obtain a video scene graph. We use a 2-layer Graph Attention Network (Veličković et al., 2017) with 12 heads to learn the representations of the scene graph. As shown in Table 3, adding the clip-level captions improves performance. And the video scene graph brings further improvement. The video scene graph contains video-level semantic information in a structural form. Therefore, it can better represent key semantic information and reduce redundant information.

Analysis of Dynamic Reasoning As Table 4 shows. We first evaluate the simple static reasoning mechanism adopted by most previous works. Then we evaluate the performance of the 3.3 ***Evaluator*** and 3.3 ***Integrator***. Row 0 shows the result of the single-step reasoning, achieving 39.5% accuracy on MSRVTT-QA. As expected, the static multistep reasoning mechanism (row 1-4) performs better and achieves the best performance around $Step = 3$. But the performance does not increase when setting a larger number of steps. It means that the average complexity of all questions is moderate, which can be handled well with three-step reason-

	Step Strategy	Integration Strategy	Step	MSRVTT (Acc.)	
				QA	MC
0	Single-Step	-	1	39.5 ± 0.14	89.9 ± 0.11
1	Static Multistep	-	2	40.2 ± 0.12	90.4 ± 0.09
2			3	40.6 ± 0.12	90.6 ± 0.12
3			4	40.4 ± 0.13	90.4 ± 0.10
4			5	40.2 ± 0.11	90.2 ± 0.10
5	Evaluator	-	≤ 5	40.8 ± **0.05**	90.7 ± **0.04**
6	Static Multistep	MeanPooling	2	39.8 ± 0.11	90.3 ± 0.10
7			3	40.2 ± 0.10	90.4 ± 0.12
8			4	40.1 ± 0.13	90.4 ± 0.09
9			5	39.8 ± 0.12	90.2 ± 0.08
10		MaxPooling	2	40.3 ± 0.08	90.4 ± 0.07
11			3	40.9 ± 0.09	90.5 ± 0.08
12			4	40.6 ± 0.07	90.7 ± 0.09
13			5	40.8 ± 0.08	90.6 ± 0.07
14		Integrator	2	40.4 ± 0.05	90.5 ± 0.04
15			3	41.0 ± 0.04	90.7 ± 0.04
16			4	40.8 ± 0.06	90.8 ± 0.06
17			5	40.7 ± 0.04	90.6 ± 0.05
18	Evaluator	Integrator	≤ 5	**41.6** ± **0.02**	**91.4** ± **0.03**

Table 4: **Impact of dynamic reasoning strategy.** When using a static reasoning strategy, the model will execute multistep reasoning with a fixed number of steps. When using a dynamic reasoning strategy, the model will dynamically determine the number of steps of multistep reasoning according to the problem's difficulty (***Evaluator***) and dynamically get the weight sum of intermediate reasoning results as the input of answer decision (***Integrator***).

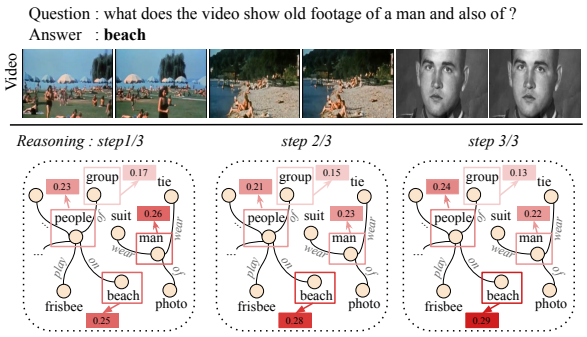

Figure 4: **Visualization for Dynamic Multistep Reasoning.** We trace the attention weight distribution of the ***Reader*** to visualize the dynamic multistep reasoning process. For this case, the output of the ***Evaluator*** is $S = 3$, which means that three steps of reasoning are required. With the reasoning process, the attention weight of 'beach' is increasing. On the contrary, the attention weight of the irrelevant nodes reduced gradually. It indicates that the reasoning mechanism is gradually deepening the understanding of the question.

ing. The ***Evaluator*** can dynamically decide the number of reasoning steps (≤ 5) for each question. Row 5 shows that using different numbers of reasoning steps for questions of different complexity can perform better than a static multistep reasoning mechanism. At the same time, we believe that each step of reasoning pays attention to different parts of the video content and gradually deepens the understanding of the question. Therefore, the intermediate reasoning results are also helpful in choosing an answer. Row 6-17 show that the ***Integrator*** can further improves the performance and has better integration ability than the traditional pooling method. Row 18 shows that the Evaluator and Integrator can promote each other and achieve the best performance.

5.3 Interpretability

Our model is interpretable by visualizing the attention weight of the ***Reader***. As Figure 4 shows, we conclude with the following two conclusions: (1) **Video scene graph can better represents the visual semantic information.** The Video Scene Graph contains the global semantic information of the video, and the structured form can not only highlight the critical semantic elements but also reduce the interference of redundant information. (2) **Dynamic multistep reasoning mechanism can deepen the understanding of the question iteratively.** With the reasoning process, the semantic nodes related to the question obtain more and more attention. On the contrary, the attention weight of the irrelevant nodes reduced gradually. It shows that the dynamic multistep reasoning mechanism can deepen the understanding of the question.

6 Conclusion

We propose a dynamic reasoning mechanism based on video scene graph for video QA to alleviate the drawback of existing methods, that is, lack of deep understanding and multistep reasoning. Experiments show that our method significantly surpasses previous methods on multiple video QA datasets due to better understanding and reasoning mechanisms and achieves much better interpretability by backtracing along with the attention mechanism to the video scene graph. In addition, different from the conventional manners that perform classification after crossmodal feature encoding, the model realizes the decoupling of question understanding and video understanding and the decoupling of understanding and decision-making, thus providing more possibilities for improvement. On the one hand, we can optimize video QA by introducing external knowledge, designing more effective reasoning mechanisms, defining and constructing better video scene graphs. On the other hand, we can also jointly model multimodal QA, such as QA based on videos, tables, texts, and graphs.

Acknowledgements

This work is supported by the Beijing Natural Science Foundation (Z190020). This work is also supported by Baidu and CASICT Joint Project. We would like to thank the anonymous reviewers for their valuable feedback.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, and Liang Lin. 2018. Visual question reasoning on general dependency tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7249–7257.

Ruisheng Cao, Lu Chen, Zhi Chen, Yanbin Zhao, Su Zhu, and Kai Yu. 2021. Lgesql: Line graph enhanced text-to-sql model with mixed local and non-local relations. *arXiv preprint arXiv:2106.01093*.

Ronald P Carver and James V Hoffman. 1981. The effect of practice through repeated reading on gain in reading ability using a computer-based instructional system. *Reading Research Quarterly*, pages 374–390.

Aman Chadha, Gurneet Arora, and Navpreet Kaloty. 2020. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*.

Anna Chang and Sonia Millett. 2013. Improving reading rates and comprehension through timed repeated reading.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. *arXiv preprint arXiv:1905.05733*.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007.

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

Noa Garcia and Yuta Nakashima. 2020. Knowledge-based video question answering with unsupervised scene descriptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 581–598. Springer.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10826–10834.

Greta Gorsuch and Etsuo Taguchi. 2008. Repeated reading for developing reading fluency and reading comprehension: The case of efl learners in vietnam. *System*, 36(2):253–278.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. *arXiv preprint arXiv:1905.08205*.

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

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. 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.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Hyounghun Kim, Zineng Tang, and Mohit Bansal. 2020a. Dense-caption matching and frame-selection gating for temporal localization in videoqa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4812–4822.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. 2020b. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. *arXiv preprint arXiv:2009.08043*.

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

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. 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.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Tianhao Li and Limin Wang. 2020. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

Jiangming Liu, Matt Gardner, Shay B Cohen, and Mirella Lapata. 2020a. Multi-step inference for reasoning over paragraphs. *arXiv preprint arXiv:2004.02995*.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020b. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Kwabena Nuamah. 2021. Deep algorithmic question answering: Towards a compositionally hybrid ai for algorithmic reasoning. *arXiv preprint arXiv:2109.08006*.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2021. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. *arXiv preprint arXiv:2012.14610*, 54:57–60.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language (VL15)*, Lisbon, Portugal. Association for Computational Linguistics.

Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. 2021. Attend what you need: Motion-appearance synergistic networks for video question answering. *arXiv preprint arXiv:2106.10446*.

Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230:107408.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2019. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. *arXiv preprint arXiv:1911.04942*.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2021. Video as conditional graph hierarchy for multi-granular question answering. *arXiv preprint arXiv:2112.06197*.

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

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

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. 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.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018a. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018b. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2016. Video captioning and retrieval models with semantic attention. *arXiv preprint arXiv:1610.02947*, 6(7).

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. *arXiv preprint arXiv:1801.04726*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.