Semantic-aware Dynamic Retrospective-Prospective Reasoning for Event-level Video Question Answering

Chenyang Lyu[†] **Tianbo Ji**[‡*] **Yvette Graham**[¶] **Jennifer Foster**[†]
[†] School of Computing, Dublin City University, Dublin, Ireland
[‡] Nantong University, China
[¶] School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
chenyang.lyu2@mail.dcu.ie, ygraham@tcd.ie, jennifer.foster@dcu.ie
jitianbo@ntu.edu.cn

Abstract

Event-Level Video Question Answering (EVQA) requires complex reasoning across video events to obtain the visual information needed to provide optimal answers. However, despite significant progress in model performance, few studies have focused on using the explicit semantic connections between the question and visual information especially at the event level. There is need for using such semantic connections to facilitate complex reasoning across video frames. Therefore, we propose a semantic-aware dynamic retrospective-prospective reasoning approach for video-based question answering. Specifically, we explicitly use the Semantic Role Labeling (SRL) structure of the question in the dynamic reasoning process where we decide to move to the next frame based on which part of the SRL structure (agent, verb, patient, etc.) of the question is being focused on. We conduct experiments on a benchmark EVQA dataset - TrafficQA. Results show that our proposed approach achieves superior performance compared to previous state-of-the-art models. Our code is publicly available at https://github.com/lyuchenyang/Semantic-aware-VideoQA.

1 Introduction

This paper focuses on one specific variant of Video Question Answering (VQA) (Xu et al., 2016; Yu et al., 2018; Zhong et al., 2022), namely Event-level VQA (EVQA) (Xu et al., 2021). In general, the objective of the VQA task is to provide an answer to a visual-related question according to the content of an accompanying video. Despite significant recent progress in VQA, EVQA still remains one of the most challenging VQA-based tasks since it requires complex reasoning over the *events* across video frames (Sadhu et al., 2021; Zhong et al., 2022; Liu et al., 2022). To tackle the challenges in EVQA, a number of approaches have been proposed (Xu et al., 2021). Luo et al. (2022) propose a temporal-aware bidirectional attention mechanism for improving event reasoning in videos, while Zhang et al. (2022) propose a novel model named Energy-based Refined-attention Mechanism (ERM), which obtains better performance compared to previous approaches with a smaller model size. Liu et al. (2022), on the other hand, incorporate visual-linguistic causal dependencies based on Graph Convolutional Networks (Kipf and Welling, 2017) for enhancing cross-modal event reasoning for EVQA.

Despite recent advances, conventional EVQA approaches generally fail to take into account the explicit semantic connection between questions and the corresponding visual information at the event level. Therefore, we propose a new approach that takes advantage of such semantic connections, using the Semantic Role Labeling (SRL) (Màrquez et al., 2008; Palmer et al., 2010; He et al., 2017) structure of questions. The model uses SRL information to learn an explicit semantic connection between the text-based questions and visual information in videos. Additionally, we carry out a multi-step reasoning mechanism over video frames to avoid adapting to spurious correlation and shortcuts by explicitly learning the reasoning process itself (Yi et al., 2018; Zhang et al., 2021; Picco et al., 2021; Hamilton et al., 2022; Zhu, 2022).

Specifically, in each reasoning step, the model should explicitly decide which frame should be focused on by predicting the reasoning direction (*retrospective* or *prospective*). In terms of the question, in each reasoning step, we focus on one or more specific SRL arguments with high attention weights, and model its connection with the visual information (i.e., video frames) contained within the corresponding video. For example, for a question such as *[**ARG1**: How many cars] were [Verb: involved] [**ARG2**: in the accident?]*, the model con-

[*] corresponding author

Proceedings of the 61st Annual Meeting of the
Association for Computational Linguistics - Student Research Workshop, pages 50–56
July 10-12, 2023 ©2023 Association for Computational Linguistics

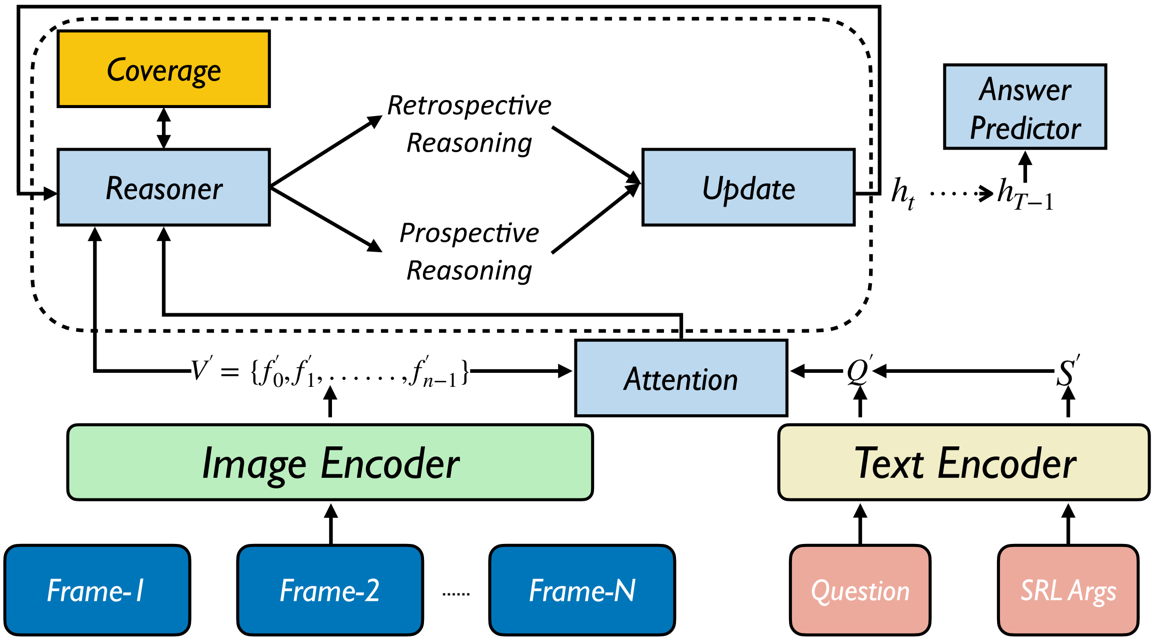

Figure 1: Overview of our approach for multi-step visual reasoning. In each reasoning step, the model predicts the reasoning direction (either *retrospective* or *prospective*) and focuses on a specific SRL argument with high attention weights. A *coverage mechanism* is employed to improve the coverage of SRL arguments in the question.

centrates on the ***ARG2*** when locating the accident, before determining how many cars were involved in the accident (***ARG1***). In a specific reasoning step, t, we inject the relevant visual information based on the semantic connection between the question and video frames by updating a hidden vector. This vector is ultimately expected to contain the necessary information for predicting the correct answer. In the reasoning process, we employ a *coverage mechanism* (Tu et al., 2016) to improve the coverage of the SRL arguments of question. Namely, instead of simply focusing on a small number of specific arguments, the model is capable of including a large range of arguments.

To investigate the effectiveness of the proposed approach, we conduct experiments on a benchmark EVQA dataset: TrafficQA. Results reveal the model to achieve performance superior to that of existing baselines for a range of reasoning types (e.g., counterfactual, prospective).

2 Methodology

An overview of our approach is shown in Figure 1. Suppose the input of our model consists of a video V composed of n image frames sampled from it: $V = \{f_0, f_1, \ldots\ldots, f_{n-1}\}$, and a corresponding question $Q = \{w_0, w_1, \ldots\ldots, w_{m-1}\}$ with associated SRL arguments $S = \{S_0, S_1, \ldots\ldots, S_{N-1}\}$ where $S_i = \{w_i, w_{i+1}, \ldots\ldots, w_k\}$. All frames $V = \{f_0, f_1, \ldots\ldots, f_{n-1}\}$ are fed into an Image Encoder followed by temporal attention modeling to produce temporal-aware frame representations $V' = \{f_0', f_1', \ldots\ldots, f_{n-1}'\} \in \mathbf{R}^{n \times d}$. Meanwhile, we use a Text Encoder to obtain the representations of the question with its corresponding SRL arguments: $Q' \in \mathbf{R}^{1 \times d}$ and $S' \in \mathbf{R}^{N \times d}$. We then perform multi-step reasoning in which we iteratively update the hidden state vector h with the visual information from frame representations based on the attention weights between them and the SRL arguments of the question. h is updated from the initial step h_0 to the final step h_{T-1} where T is the total number of reasoning steps. Finally, we predict the most probable answer a based on h_{T-1}.

2.1 Multi-step Reasoning

Before the first reasoning step, we initialize:

$$h_0 = Attn(Q', V', V') \tag{1}$$

$$j = argmax(AttnWeights(Q', V', V')) \tag{2}$$

where $Attn$ serves as the q, k, v *attention*[1] modeling (Vaswani et al., 2017) and j represents the

[1] In this work, we use a low temperature τ in the *softmax* to encourage the model to assign more attention weights to the most relevant frame.

index of the frame with the highest attention weight. In each specific reasoning step t, we firstly use h_{t-1} as the *attention key* to obtain the relevant SRL argument: $S_t^{'} = Attn(h_{t-1}, S^{'}, S^{'})$. Subsequently, we infer the next focused frame by:

$$V^{focus} = Attn(r_t, V^{'}, V^{'}) \tag{3}$$

where $r_t = g(h_{t-1}, S_t^{'})$. Finally, we update the hidden state vector h_{t-1} based on the currently focused frame (the frame with the largest attention weight):

$$h_t = \delta(h_{t-1}, V^{focus}) \tag{4}$$

2.2 Retrospective-Prospective Reasoning

We propose a *Retrospective-Prospective Reasoning* mechanism for Eq.3 in order to explicitly decide whether the model should move to future frames (*prospective reasoning*) or move back to previous frames (*retrospective reasoning*). We obtain the *retrospective frame* V^{retro} and *prospective frame* V^{prosp} by:

$$V^{retro} = \psi(g(h_{t-1}, S_t^{'}), V^{'}, RetroMask(j)) \tag{5}$$

$$V^{prosp} = \phi(g(h_{t-1}, S_t^{'}), V^{'}, ProspMask(j)) \tag{6}$$

where ψ and ϕ are MASKED ATTENTION that are used to obtain *retrospective* and *prospective* frames, $g(h_{t-1}, S_t^{'})$ and $V^{'}$ serve as *query* and *key, value* respectively. $RetroMask(j)$ means all frames after j ($f_{i>j}$) will be masked whereas $ProspMask(j)$ means that all frames before j ($f_{i<j}$) will be masked. After obtaining V^{retro} and V^{prosp} we generate a probability:

$$p = \sigma(\lambda(V^{retro}, V^{prosp})) \tag{7}$$

If p is larger than a pre-defined threshold α, we update $h_t = \delta(h_{t-1}, V^{retro})$,otherwise we update $h_t = \delta(h_{t-1}, V^{prosp})$ as in Eq. 4. The index for the next-focused frame j is also updated accordingly. The reasoning process is shown in Algorithm 1.

2.3 Coverage Mechanism

We additionally propose to employ a *coverage mechanism* (Tu et al., 2016) to encourage the model to include as many SRL arguments as possible in the reasoning process. Specifically, we track the attention distribution $C_t \in \mathbf{R}^{1\times N}$ of h_{t-1} on all SRL arguments S

$$C_t = C_{t-1} + \frac{AttnWeights([h_{t-1}; C_{t-1}], S^{'}, S^{'})}{\chi} \tag{8}$$

Algorithm 1: Multi-step dynamic retrospective-prospective reasoning with coverage mechanism

```
$V^{'} = \{f_0, f_1, ......, f_{n-1}\}$: representations of video
  frames
$Q^{'}$: question
$S^{'}$: SRL representations of $Q$
$T$: reasoning steps
$\chi$ : normalization factor
$\alpha$: threshold of the probability for using retrospective
  frame
$h_0 = Attn(Q^{'}, V^{'}, V^{'})$
$j = argmax(AttnWeights(Q^{'}, V^{'}, V^{'}))$
$C_0 = 0$
for $i$ in $T$ do
    $S_i^{'} = Attn(h_{i-1}, S^{'}, S^{'}, C_{i-1})$
    $C_i = C_{i-1} + \frac{AttnWeights(h_{i-1}, S^{'}, S^{'}, C_{i-1})}{\chi}$
    $V^{retro} = \psi(g(h_{t-1}, S_t^{'}), V^{'}, RetroMask(j))$
    $V^{prosp} = \phi(g(h_{i-1}, S_i^{'}), V^{'}, ProspMask(j))$
    $p = \sigma(f(V^{retro}, V^{prosp}))$
    if $p > \alpha$ then
        $h_i = \delta(h_{i-1}, V^{retro})$
        $j = argmax(\psi(g(h_{t-1}, S_t^{'}), V^{'}, RetroMask(j)))$
    else
        $h_i = \delta(h_{i-1}, V^{prosp})$
        $j = argmax(\phi(g(h_{i-1}, S_i^{'}), V^{'}, ProspMask(j)))$
```

where χ represents the normalization factor.[2] We obtain the weighted $S_t^{'}$ by $S_t^{'} = Attn([h_{t-1}; C_{t-1}], S^{'}, S^{'})$ where we concatenate C_{t-1} to h_{t-1} as an additional input to the *Attn* function for the purpose of informing the model to assign more attention weights to previously less-focused SRL arguments, in order to improve the coverage for all SRL arguments.

2.4 Training Objective

For the answer prediction, we encode all answer options $A = \{a_0, ..., a_{M-1}\}$ separately and then select the one with the highest similarity with h_{T-1}. We optimize our model parameters θ using *Cross Entropy* loss:

$$J(\theta) = -\sum_i \sum_k log \frac{e^{F(a_k, h_{T-1})}}{\sum_{j=0}^{M-1} e^{F(a_j, h_{T-1})}} y_{i,k} \tag{9}$$

where F is the function measuring the similarity between answer candidate and h_{T-1}, and $y_{i,k}$ represents the answer label for the $i-$th example - if the correct answer for the $i-$th example is the $k-$th answer then $y_{i,k}$ is 1 otherwise it is 0.

[2] In this work, we use the number of SRL arguments of the corresponding question as the normalization factor.

Models	Setting-1/4	Setting-1/2
Q-type (random) (Xu et al., 2021)	25.00	50.00
QE-LSTM (Xu et al., 2021)	25.21	50.45
QA-LSTM (Xu et al., 2021)	26.65	51.02
Avgpooling (Xu et al., 2021)	30.45	57.50
CNN+LSTM (Xu et al., 2021)	30.78	57.64
I3D+LSTM (Xu et al., 2021)	33.21	54.67
VIS+LSTM (Ren et al., 2015)	29.91	54.25
BERT-VQA (Yang et al., 2020)	33.68	63.50
TVQA (Lei et al., 2018)	35.16	63.15
HCRN (Le et al., 2020a)	36.49	63.79
Eclipse (Xu et al., 2021)	37.05	64.77
ERM (Zhang et al., 2022)	37.11	65.14
TMBC (Luo et al., 2022)	37.17	65.14
CMCIR (Liu et al., 2022)	38.58	N/A
Ours	**43.19**	**71.63**

Table 1: Evaluation results on TrafficQA dataset.

3 Experiments

3.1 Dataset

We employ a benchmark dataset for EVQA - TrafficQA (Xu et al., 2021) which contains 62,535 QA pairs and 10,080 videos. We follow the standard split of TrafficQA – 56,460 pairs for training and 6,075 pairs for evaluation. We further sample 5,000 examples from training data as the dev set.

3.2 Experimental Setup

We use CLIP ViT-B/16 (Radford et al., 2021) [3] to initialize our image encoder and text encoder. We evenly sample 10 frames from each video in the TrafficQA dataset. The SRL parser employed in the experiments is from AllenNLP (Gardner et al., 2018; Shi and Lin, 2019). We train our model over 10 epochs with a learning rate of 1×10^{-6} and a batch size of 8. The optimizer is AdamW (Loshchilov and Hutter, 2019). We set the maximum reasoning step T to 3 and we use a temperature τ of 0.2 in *Attention* modeling. The hyper-parameters are empirically selected based on the performance on dev set. There are two experimental settings for TrafficQA (Xu et al., 2021): 1) Setting-1/2, this task is to predict whether an answer is correct for a given question based on videos; 2) Setting-1/4: this task follows the standard setup of multiple-choice task in which the model is expected to predict the correct the answer from the four candidate options.

3.3 Results

The experimental results on the test set of TrafficQA are shown in Table 1, where we also include the previous baseline models for EVQA.[4] The results show that our proposed approach obtains accuracy of 43.19 under the multiple-choice setting, which surpasses previous state-of-the-art approaches including Eclipse (Xu et al., 2021), ERM (Zhang et al., 2022), TMBC (Luo et al., 2022) and CMCIR (Liu et al., 2022) by at least 4.5 points. Furthermore, our approach achieves an accuracy of 71.63 under Setting 1/2, outperforming previous strong baselines by at least 6 points. The results show the effectiveness of our proposed multi-step reasoning approach for event-level VideoQA.

Ablation Study We conduct experiments on the dev set of TrafficQA, investigating the contribution of both the *retrospective-prospective reasoning* and *coverage mechanism* on the performance of our proposed EVQA approach. The results are shown in Table 3, which reveals that multi-step reasoning is critical in terms of model performance while the *coverage mechanism* can provide additional, albeit less substantial, improvements.

Results by Question Type We take a closer look at model performance on different question types, e.g. reverse reasoning, counterfactual reasoning, etc. The results are shown in Table 2. They reveal that our proposed approach outperforms previous state-of-the-art models on all individual question types by a large margin with large improvements seen for *introspection*, *reverse* and *counterfactual* questions.

Effect of Reasoning Steps We study the effect of varying reasoning steps. The results are shown in Table 4. Increasing reasoning steps improves performance, especially from 1 step to 3 steps. Additionally, the performance (both Setting 1/4 and 1/2) is stable with reasoning steps exceeding three.

4 Conclusion and Future Work

In this paper, we propose a multi-step dynamic retrospective-prospective approach for EVQA. Our approach employs a multi-step reasoning model that explicitly learns reasoning based on the semantic connection of the SRL structure of a question and corresponding video frames. We additionally proposed a *coverage mechanism* to improve the coverage of SRL arguments in the reasoning process. Experimental results show that the proposed

[3] https://openai.com/blog/clip/

[4] Some of the baseline results are taken from Xu et al. (2021).

Method	Question Type						
	Basic	Attribution	Introspection	Counterfactual	Forecasting	Reverse	All
HCRN (Le et al., 2020b)	34.17	50.29	33.40	40.73	44.58	50.09	36.26
VQAC (Kim et al., 2021)	34.02	49.43	34.44	39.74	38.55	49.73	36.00
MASN(Seo et al., 2021)	33.83	50.86	34.23	41.06	41.57	50.80	36.03
DualVGR (Wang et al., 2021)	33.91	50.57	33.40	41.39	41.57	50.62	36.07
CMCIR (Liu et al., 2022)	36.10	52.59	38.38	46.03	48.80	52.21	38.58
Ours	**37.05**	**52.68**	**43.91**	**50.81**	**54.26**	**55.52**	**43.19**

Table 2: Results by various *question type* on the dev set of TrafficQA. The highest performance are in bold.

Models	Setting-1/4	Setting-1/2
Model w/o MR and CM	42.53	69.61
Model w/o CM	46.15	74.97
Model	47.38	75.83

Table 3: Ablation study results on TrafficQA dev set, where *MR* represents *Multi-step Reasoning* and *CM* represents *Coverage Mechanism*. MR and CM are coupled in our approach.

Reasoning Steps	Setting-1/4	Setting-1/2
Model w/ 1 step	41.57	71.46
Model w/ 2 steps	44.21	74.95
Model w/ 3 steps	47.38	75.83
Model w/ 4 steps	47.23	75.96
Model w/ 5 steps	47.15	75.87

Table 4: The effect of various reasoning steps.

approach obtains superior performance compared to that of state-of-the-art EVQA models.

Acknowledgements

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183). We thank the reviewers for helpful feedback.

Limitations

This papers focuses on a variety of VideoQA - event-level VideoQA, we only incorporate *event* information from the question (textual) side as we think that parsing video frames is inaccurate and could introduce unexpected errors, we should also explore how to inject *event-level* information from visual side in the future with more competitive visual parsing models. Our experiments are only conducted on one dataset due to resource constraint, we should also conduct experiments on more datasets to verify the effectiveness of our approach.

References

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2022. Is neuro-symbolic ai meeting its promise in natural language processing? a structured review. *arXiv preprint arXiv:2202.12205*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Nayoung Kim, Seong Jong Ha, and Je-Won Kang. 2021. Video question answering using language-guided deep compressed-domain video feature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1708–1717.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020b. 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, 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.

Yang Liu, Guanbin Li, and Liang Lin. 2022. Cross-modal causal relational reasoning for event-level visual question answering. *arXiv preprint arXiv:2207.12647*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yuanmao Luo, Ruomei Wang, Fuwei Zhang, Fan Zhou, and Shujin Lin. 2022. Temporal-aware mechanism with bidirectional complementarity for video q&a. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3273–3278. IEEE.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Gabriele Picco, Thanh Lam Hoang, Marco Luca Sbodio, and Vanessa Lopez. 2021. Neural unification for logic reasoning over natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3939–3950, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2021. Video question answering with phrases via semantic roles. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2460–2478, Online. 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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6167–6177, Online. Association for Computational Linguistics.

Peng Shi and Jimmy J. Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

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.

Li Xu, He Huang, and Jun Liu. 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888.

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1556–1565.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

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

Fuwei Zhang, Ruomei Wang, Fan Zhou, and Yuanmao Luo. 2022. Erm: Energy-based refined-attention mechanism for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2021. Explicit cross-modal representation learning for visual commonsense reasoning. *IEEE Transactions on Multimedia*, 24:2986–2997.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.

Zihao Zhu. 2022. From shallow to deep: Compositional reasoning over graphs for visual question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8217–8221. IEEE.