Long Story Short: a Summarize-then-Search Method for Long Video Question Answering

Jiwan Chung
https://jiwanchung.github.io
Youngjae Yu
https://yj-yu.github.io/home

MIR Lab
Yonsei University
Seoul, Korea

Abstract

Large language models such as GPT-3 have demonstrated an impressive capability to adapt to new tasks without requiring task-specific training data. This capability has been particularly effective in settings such as narrative question answering, where the diversity of tasks is immense, but the available supervision data is small. In this work, we investigate if such language models can extend their zero-shot reasoning abilities to long multimodal narratives in multimedia content such as drama, movies, and animation, where the story plays an essential role. We propose Long Story Short, a framework for narrative video QA that first summarizes the narrative of the video to a short plot and then searches parts of the video relevant to the question. We also propose to enhance visual matching with `CLIPCheck`. Our model outperforms state-of-the-art supervised models by a large margin, highlighting the potential of zero-shot QA for long videos.

1 Introduction

Recent video QA models face challenges in handling long video narrative QA tasks [2, 13, 27] (*i.e.*, films, dramas, and YouTube web videos) due to the limitation in data and annotations. This results in an inability to comprehend the long video narratives beyond answering mainly visual questions on short video clip [16, 17, 30]. The sizes of such long video QAs are insufficient to train the models to fully comprehend the complex narrative structures within a video, yielding sub-optimal performances. [10] demonstrate that the supervised models rely more on language biases in the question than the narrative context: they can obtain similar performance even without seeing any video context. This highlights the necessity of multimodal reasoning capability beyond small task-specific supervision.

To address the challenge caused by low generalization, a zero-shot approach using pretrained Large Language Models (LLMs) can be an efficient alternative for tackling complex QA tasks [32], and text context summarization [8, 37]. Yet, is the narrative QA capability of such LLMs transferable to the video domain?

We propose Long Story Short (`LSS`), illustrated in figure 1, that translates video clips into text screenplay format inspired by Socratic Model [35]. Using GPT-3 [1], we first summarize

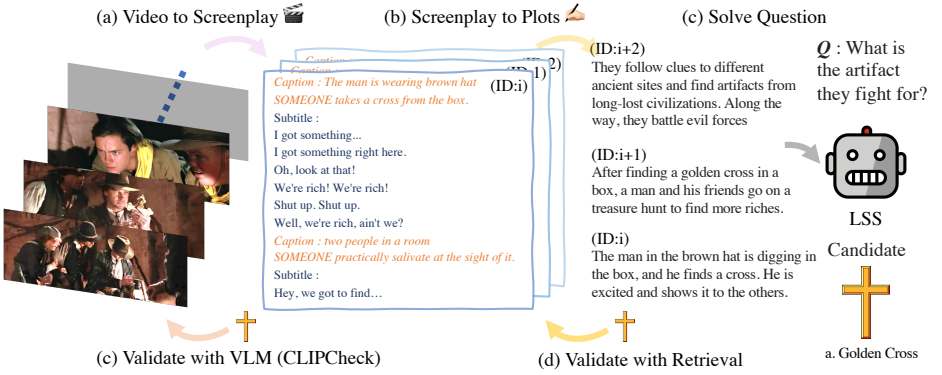

Figure 1: Long Story Short (LSS) uses Large Language Models (LLMs) (*i.e.*, GPT-3) to generate (a) Screenplay and summarized (b) Plots from video. Further details about data processing can be found in Section 2. When LSS answer questions about the video, the model (c) validate given raw video footage with Visual Language Model, CLIP, and (d) search further grounded scripts in a backward manner, which we call CLIPCheck in Section 2.3.

the long video into a list of plots and then navigate both the generated summary and the raw video context to resolve the given question. Our zero-shot method shows better results than state-of-the-art supervised methods in MovieQA and DramaQA dataset. Furthermore, we propose CLIPCheck, a visual-text matching method to enhance visual alignment of the reasoning results provided by GPT-3. To summarize, our main contributions are three-fold:

1. We present LSS, a framework that summarizes a long video narrative to a list of plots and retrieves the subplot relevant to the question.

2. We demonstrate the importance of considering visual alignment strength via CLIP-based matching in visual prompting.

3. Our zero-shot approach achieves state-of-the-art performance in MovieQA [27] and DramaQA [2], outperforming supervised baselines.

2 Method

We describe Long Story Short (LSS), a prompt-based strategy that divides and conquers narrative video QA by summarizing the plot and retrieving relevant information with the plot. Our objective is to predict the correct answer $a_{\hat{y}}$ from long video context $\mathbf{X}$. Since language models can only accept tokens of fixed length (*e.g.* 4096 tokens for GPT-3), they cannot process the rich context (*e.g.* subtitles or scene descriptions) of movies that typically spans two long hours.

Thus, we introduce a summarize-then-search method for video question answering. We first divide the long video context into clip segments and summarize each segment into a subplot using GPT-3 to get the list of subplots $\mathbf{S}_X = \{s_1, s_2 \cdots s_n\}$. Then, we retrieve video segments $\mathbf{X}_k = \{x_{k_1}, \cdots, x_{k_m}\}$ relevant to the question with the subplot list as the input. We use both the full context of the selected segments and the global plot information to derive

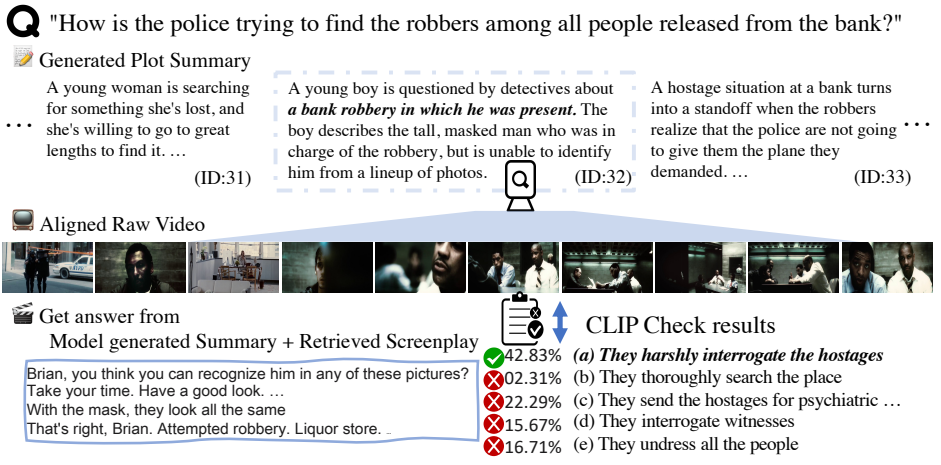

Figure 2: The qualitative result showing our proposed Long Story Short (LSS) model that generates and retrieves the index of raw video footage. When the model predicts the final answer from (i) the generated Summary and (ii) the retrieved text context, CLIPCheck validates each candidate's answers to revise the final answer for the question.

likelihood over the answer choices $p_\theta(a_1), \cdots, p_\theta(a_5)$. Finally, we apply CLIPCheck, to strengthen the visual grounding of the selected answer. We provide the prompt templates in the appendix.

2.1 Plot Generation

We use the ground-truth video partitions to segment the whole video into a set of shorter clips. Each long video $\mathbf{X} = \{x_1, \cdots, x_n\}$ consists of n clip segments $\mathbf{X_i}$, and each segment contains video v_i and the corresponding text t_i such as subtitle or ASR.

$$\mathbf{X} = \{(v_1, t_1), (v_2, t_1) \cdots (v_n, t_n)\}$$

Given a video $\mathbf{X}$, we first extract visual and text features of the video $\mathbf{X}'$ in natural language forms. As [32] transcribe image as language prompt for frozen GPT-3, we retrieve DVS caption [26] and image captions with pretrained BLIP [18] for video $v_i \in \mathbf{X}$, and concatenate the aligned text t_i as n generated scripts as illustrated in figure 1. We compose a prompt to summarize the context of a video segment into a plot piece of up to three sentences and use GPT-3 to build the plot list aligned with the segment indices $S_\mathbf{X} = \{s_1, \cdots, s_n\}$.

2.2 Narrative Search

Given the summarized narrative and the question, we wish to retrieve the relatively short clip relevant to the question from the long video. Language models generate open-ended text which is irregular and often noisy. To retrieve the exact part of the video, we drive the model to output indices of the plot rather than the text form.

We first assign consecutive indices to the list of summarized plot pieces $S_\mathbf{X} = \{s_1, \cdots, s_n\}$ aligned with the clip segmentation. Then, we prompt the language model to output the indices of the plot pieces $k = \{k_1, \cdots, k_m\}$ to lookup for.

The generated indices might still be noisy due to the open-ended nature of language models. When the model outputs an answer in text form, we use rouge-l [19] score to find plot piece candidates whose similarity with the generated sentence are above the specified threshold $\alpha \geq 0.5$.

Finally, we concatenate the plots $S_{\mathbf{X}}$, the visual and text representation of the selected segments $\mathbf{X'}_k = \{(v_{k_1}, t_{k_1}), \cdots, (v_{k_m}, t_{k_m})\}$, the question q, and the answer choices $A = \{a_1, \cdots, a_5\}$ to build the prompt input for question answering. We process the prompt with the language model with weights θ and use the index token likelihood as the answer choice score.

$$p_\theta(a_i) = p_\theta(i | S_{\mathbf{X}}, \mathbf{X'}_k, q, A)$$

2.3 Visual Checking

For a tighter visual-text matching, we introduce `CLIPCheck`, a method to conjoin CLIP visual distance [25] and the language model likelihood. We start from the selected video segments $\mathbf{X}_k = \{x_{k_1}, \cdots, x_{k_m}\}$, answer choices $A = \{a_1, \cdots, a_5\}$, and answer likelihoods $P_\theta = \{p_\theta(a_1), \cdots, p_\theta(a_5)\}$ of GPT-3.

First, we use the CLIP image encoder to encode each frame $x_{k_{i_j}}$ of the selected segments. When l is the number of frames within a segment:

$$\bar{x}_{k_{i_j}} = CLIP_V(x_{k_{i_j}})$$
$$\bar{x}_{k_i} = \{\bar{x}_{k_{i_1}}, \cdots, \bar{x}_{k_{i_l}}\}, \bar{\mathbf{X}}_k = \{\bar{x}_{k_1}, \cdots, \bar{x}_{k_m}\}$$

Then, we extract the CLIP text feature of each answer choice $\bar{a}_i = CLIP_L(a_i)$ and compare cosine similarity between the video input and the answers. We select the best-matched frame for each answer to derive the cosine similarity score.

$$c(a, \bar{\mathbf{X}}_k) = \max_{i \leq m, j \leq l} cossim(\bar{a}, \bar{x}_{k_{i_j}})$$

Then we apply the softmax function with temperature τ on the scores to get normalized visual likelihood over the answer candidates $P_c = \{p_c(a_1), \cdots, p_c(a_5)\}$. Lastly, we multiply the answer likelihood from the language model P_θ with the visual likelihood P_c to obtain the final likelihood. We simply select the answer with the maximum value as the model answer.

We choose to consider `CLIPCheck` only when the language model is not certain of its choice. Given the likelihoods of the top two answers $p_\theta(a_{h_1}), p_\theta(a_{h_2})$ from the language model, we measure the model certainty with binary entropy E' of the re-normalized probability. We only use the combined likelihood when the binary entropy is greater than the given threshold $E' \geq 0.4$. Otherwise, we do not apply `CLIPCheck` and just use the language model likelihood.

$$h_1 = \arg\max_{i \leq 5} p_\theta(a_i), h_2 = \arg\max_{j \leq 5, j \neq h_1} p_\theta(a_j)$$
$$E' = -p_\theta(a_{h_1}) \log p_\theta(a_{h_1}) - p_\theta(a_{h_2}) \log p_\theta(a_{h_2})$$

3 Experiments

For all experiments, we use GPT-3 [1] (`text-davinci-003`) as the backbone language model. Unless stated otherwise, we use the ground truth clip boundary to segment the videos. All `LSS` variants do not use any training data and thus are zero-shot methods.

	Model	Aligned	V + S	V Only	S Only
Supervised	A2A	✓	41.66	40.28	41.05
	PAMN	✓	43.34	42.33	42.56
	UniversalQA	✓	48.87	50.67	47.62
	DHTCN	✓	49.60	47.38	48.43
zeroshot	No Context	✗	36.36	34.28	38.07
	LSS	✓	53.44	49.83	56.42
	LSS-Search	✗	51.24	49.00	53.09
	LSS-Search+CLIPCheck	✗	51.49	49.55	53.09

Table 1: Evaluation on MovieQA validation split. The dataset provides GT alignment with 3 minutes of video clip on average: We also report Ours-search which searches the whole movie context without GT alignment. (V) indicates Video and (S) indicates Subtitle.

Model		Labels		Acc
		Plot	Aligned	
Supervised [13]		✓	✓	68.00
GPT3 w\o Context		✗	✗	36.90
LSS	Base	✓	✓	66.76
	+ Search	✓	✗	48.98
	+ Plot	✗	✓	65.80
	+ Plot + Search	✗	✗	53.34

Table 2: Evaluation on PororoQA validation split. The machine-generated plot (+Plot) performs close to the human annotations (Base).

Model	Level3	Level4
CharacterAttention	60.82	65.62
Kim *et al.* [14]	70.00	70.00
LSS	72.20	75.23
+Caption	73.54	75.68
+CLIPCheck	**75.78**	**79.28**
+Caption+CLIPCheck	75.34	77.93
+CLIPCheck-Shuffle	71.74	73.87

Table 3: Evaluation on the levels three and four of DramaQA validation split. CLIPCheck achieves state-of-the-art over the baselines and a prompt-based approach [35] of inputting image descriptions.

Model	Aligned	V + S
LSS	✓	53.44
LSS-Search	✗	51.24
LSS-Search+CLIPCheck	✗	51.49
LSS-Random	✗	48.92
LSS-Full	✗	48.57

Table 4: Ablation Study on MovieQA validation split.

3.1 Evaluating Long Story Short

MovieQA [27] is a large-scale QA dataset sourced from 408 movies. There are multiple sources of information in the dataset; subtitles, scripts, DVS, video clips, and plots. We report four state-of-the-art supervised baselines; A2A [20], PAMN [11], UniversalQA [10], and DHTCN [21].

Table 1 shows zero-shot LSS improves over previous supervised approaches. Also, Ours-search shows strong performance even without the ground-truth segment index label. CLIPCheck slightly improves the accuracy in the video split. However, the difference is marginal since MovieQA often requires character-based grounding rather than general visual matching. Finally, we experiment with the null hypothesis: No Context tests whether GPT-3 solves MovieQA by simply memorizing every fact. No Context performs worse than LSS, rejecting the null hypothesis.

PororoQA [13] is a video story QA dataset built from a cartoon series. The supervised baseline takes the human-generated plot and the ground truth video segment index, while LSS +Plot+Search takes neither.

Table 2 summarizes our result on the PororoQA dataset. When using both the ground-truth episode and plots, GPT-3 performs almost on par with the supervised baseline. Substituting a human-generated summary with a model-generated one results in only a marginal performance drop. Perhaps intriguingly, the search process works better when using model-generated plots. We attribute this result to the fact that the human annotations are not designed for episode discriminability.

3.2 Evaluating CLIPCheck

DramaQA [3] is video QA dataset that focuses on story understanding. The dataset is arranged with four levels of hierarchical difficulty, which follow the human cognitive-developmental stages. We evaluate LSS on the two high levels of DramaQA to test plot understanding. We report two latest baselines in level-wise DramaQA; CharacterAttention and Kim *et al*. [14].

We compare the effect of CLIPCheck and *Caption*, a prompt-based method of incorporating image frame descriptions extracted from BLIP [18] as inputs to GPT-3. Table 3 shows that CLIPCheck offers greater improvement than image descriptions. Also, while adding image captions improves LSS, the gain disappears when used jointly with CLIPCheck. We suspect that this is because frame captions provide similar information to CLIPCheck while being much noisier. Note that the automatic *Captions* here are not an integral component of LSS. As DramaQA has visually grounded annotations already, adding automatic image

Long story short

Harry Potter is being moved to a safe house on the 30th of the month, just before his **17th birthday**. However, Voldemort and his followers are aware of the move and plan to attack Harry en route. Snape volunteers to kill Harry, but due to the fact that their wands are twinned, he is unable to do so. Bellatrix Lestrange then volunteers and is given the task. The **Death Eaters have ambushed** Harry, Ron, and Fred, and they are nowhere to be found. Hagrid is the only one who made it back safely.

In the wake of Albus Dumbledore's death, Harry Potter, Ron Weasley, and Hermione Granger **search for the Horcruxes** that will allow them to destroy Lord Voldemort. Along the way, they face many challenges and make new allies, as they try to stay one step ahead of Voldemort and his forces. **To Harry James Potter, Dumbledore leaves the Snitch** he caught in his first Quidditch match at Hogwarts as a reminder of the rewards of perseverance and skill. He also leaves Harry the **sword of Godric Gryffindor**, which is a powerful historical artifact. However, the sword is missing and its whereabouts are unknown.

WikiPedia

At the beginning of the book, Harry is about to turn **seventeen** and will lose his deceased mother's protection. Members of the Order of the Phoenix relocate the Dursleys, and prepare to move Harry to The Burrow by flying him there, using Harry's friends as decoys. **Death Eaters attack them upon departure**, and in the ensuing battle, "Mad-Eye" Moody and Hedwig are killed while George Weasley is wounded. Voldemort arrives to kill Harry, but Harry's wand fends him off on its own.

Harry, Ron, and Hermione prepare to **hunt down Voldemort's four remaining Horcruxes**. They each are given an object in Dumbledore's will: **a Golden Snitch for Harry**, a Deluminator for Ron, and The Tales of Beedle the Bard, for Hermione. Harry is also bequeathed the **Sword of Godric Gryffindor**, but the Ministry prevents him from receiving it. During Bill Weasley and Fleur Delacour's wedding, the Ministry of Magic falls to Voldemort; Death Eaters attack the wedding reception. The trio flee to 12 Grimmauld Place, Sirius Black's family home that was left to Harry.

Figure 3: Comparison between the plot summary generated by LSS and the ground-truth summary from Wikipedia. Here, we only show the first two paragraphs of the entire plot because of the space limit.

Captions on top of that would not necessarily improve the model performance. Rather, we use the *Captions* to explicitly compare early vs. late visual alignment methods.

Finally, we check whether CLIPCheck exploits the dataset bias rather than understanding the visual context. To this end, we devise a variant of CLIPCheck with random visual context (CLIPCheck-Shuffle). CLIPCheck-Shuffle does not improve over LSS with no CLIPCheck, denying the bias hypothesis.

3.3 Ablation Study

Are both the summarization and search important for narrative understanding? Here, we evaluate LSS variants with full context without the narrative search (LSS-Full) or with the plot summary and random segment as inputs (LSS-Random). Table 4 shows that both LSS-Full and LSS-Random fall behind LSS-Search, indicating the importance of retrieval. Note that we could not employ the full context in LSS-Full due to the token length limitation. Instead, we use the longest prefix of the full context that GPT3 accepts (4000 tokens minus the length of the instruction).

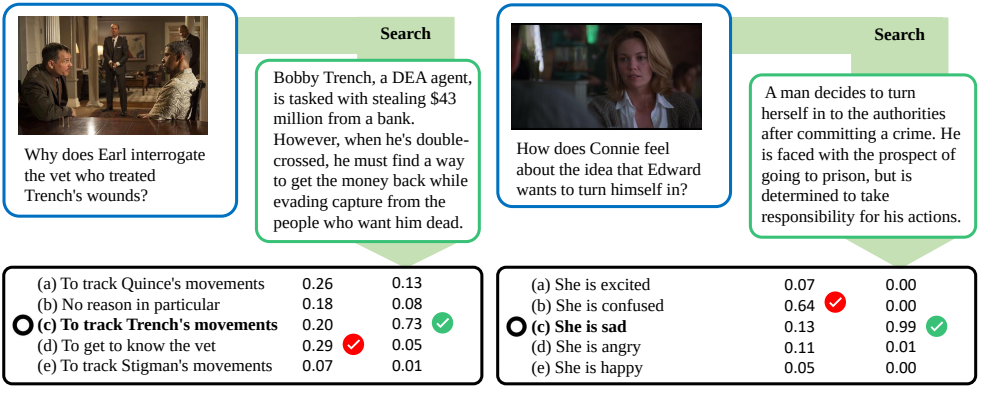

Figure 4: QA process samples in LSS. Conditioning on the searched plot piece has a substantial impact on the language model's answer likelihood distribution.

3.4 Qualitative Results

Figure 3 shows the automatic plot summary generated as an intermediate context of the long video QA using the language model in the LSS framework. As shown in the qualitative sample, the generated plots align well with the human-written plots from Wikipedia. For example, in the first scene of the movie "Harry Potter and the Deathly Hallows", the LSS summary correctly writes that Harry Potter is currently 17 years old and the main event in which the death eaters attack the protagonist.

Figure 4 depicts the connection between the searched plot piece and the answer likelihood. In the example on the left, the retrieved summary tells that Trench committed a crime and thus is on the run, suggesting that another character interested in him would be chasing him. The language model understands this context to modify the answer likelihood in the correct way. In the right example, the LSS plot piece suggests that Edward is confident in his decision. While this context does not offer a direct cue to the question, the language model sees it as information strong enough to alter the answer.

4 Related Work

Movie Summarization Movies are typical examples of long videos with clear narrative structures. Gorinski *et al*. [7]generate the shorter version of a screenplay as the task of finding an optimal graph chain of a movie scene. TRIPOD [23] is a screenplay dataset containing turning point annotations. In the same work, an automatic model to identify the turning point from movie narratives is proposed. Papalampidi *et al*. [24] later uses the TV series CSI to demonstrate the usefulness of turning points in automatic movie summarization. Lee *et al*. [15] further improves turning point identification with dialogue features and transformer architecture.

Long Video QA The task of video question answering has been studied extensively in the literature in the form of both Open-Ended QA [9] and Multi-Choice Problems [28, 29]. Several approaches have been proposed to address this task, starting from RNN-based attention networks [9, 30, 36, 38], to memory networks [12, 22, 27], and transformers [4, 6]. Recently, multimodal models pre-trained on large-scale video datasets (VideoQA [31], VIOLET [5],

and MERLOT [33] and MERLOT-Reserve [34]) shows promising performance in video question answering as well.

However, long video QA has received relatively less attention despite its importance. MovieQA [27] formulates QAs on the entire movies, which typically span two long hours. DramaQA [3] uses a single TV series as visual context, and tasks a solver to understand video clips of length from one to twenty minutes.

5 Conclusion

We introduced Long Story Short, a summarize-then-search method to understand both global narrative and the relevant details for video narrative QA. Our approach is effective when the context of QA is vast and a high-level interaction with such context is necessary to solve the said QA, which is the case in long video QAs. Also, we propose to further enhance the visual grounding of the model-generated answer by post-checking visual alignment with `CLIPCheck`. Our zero-shot method improves supervised state-of-art approaches in MovieQA and DramaQA benchmarks. We plan to release the code and the generated plot data to the public.

There are two possible research directions beyond this work: first, providing visual descriptions better aligned with the story with character re-identification and co-reference resolution improve input quality to GPT-3. Second, one can devise a more dynamic multi-hop search that combines global and local information in a hierarchical manner.

6 Limitations

Our study has some limitations, including:

1. We experiment with only videos with English subtitles. However, our method can be extended to include multi-lingual contexts given a strong multilingual language model.
2. The computation and memory requirement of our method is substantial due to its heavy reliance on the large language model, GPT-3.
3. We evaluate Long Story Short with only a single instance of LLM (GPT-3).

Potential Risk. Summarizing the long video context with GPT-3 carries on ethical risks related to the open-ended nature of the language model. GPT-3 may (a) hallucinate fake facts about the content, (b) generate toxic utterances, or (c) implicitly embed social biases into the summary and the answer likelihoods.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Seungchan Lee, Minsu Lee, and Byoung-Tak Zhang. DramaQA: character-centered video story understanding with hierarchical qa. *arXiv preprint arXiv:2005.03356*, 2020.

[3] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1166–1174, 2021.

[4] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 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, 2019.

[5] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.

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

[7] Philip John Gorinski and Mirella Lapata. Movie script summarization as graph-based scene extraction. In *NAACL*, 2015.

[8] Pengcheng He, Baolin Peng, Liyang Lu, Songhe Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. Z-code++: A pre-trained language model optimized for abstractive summarization. *ArXiv*, abs/2208.09770, 2022.

[9] 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.

[10] Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. Are we asking the right questions in movieqa? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019.

[12] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019.

[13] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022, 2017.

[14] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *AAAI*, 2021.

[15] Myungji Lee, Hong-Seok Kwon, Jaehun Shin, WonKee Lee, Baikjin Jung, and Jong-Hyeok Lee. Transformer-based screenplay summarization using augmented learning representation with dialogue information. In *NUSE*, 2021.

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[17] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[20] Chao-Ning Liu, Ding-Jie Chen, Hwann-Tzong Chen, and Tyng-Luh Liu. A2a: Attention to attention reasoning for movie question answering. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 404–419. Springer, 2019.

[21] Fei Liu, Jing Liu, Xinxin Zhu, Richang Hong, and Hanqing Lu. Dual hierarchical temporal convolutional network with qa-aware dynamic normalization for video story question answering. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4253–4261, 2020.

[22] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 677–685, 2017.

[23] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, November 2019.

[24] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

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

[26] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *IJCV*, 2017.

[27] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[28] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[29] 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.

[30] 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.

[31] 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.

[32] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021.

[33] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c6d4eb15f1e84a36eff58eca3627c82e-Paper.pdf.

[34] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[35] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. 2022.

[36] 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.

[37] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[38] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1050–1058, 2017.