DrVideo: Document Retrieval Based Long Video Understanding

Supplementary Material

1. More Implementation Details

Experiments Compute Resources. All experiments are conducted on single NVIDIA RTX 4090 GPU. The minimal GPU memory requirement is 24GB. We set the temperature to 0 for all experiments using GPT-3.5 [8], GPT-4 [1], and DeepSeek [4].

Prompt Details. We provide detailed prompts for all agents (planning agent, interaction agent, and answering agent) in the EgoSchema benchmark [7]. Planing agent is to determine whether the video captions are sufficient for answering the question. Below is the planning agent prompt:

User
You are given some language descriptions of a first-person view video along with a question about the video.
1.The video is 3 minutes long, containing a total of 90 frames.
2. Each sentence in these language descriptions represents the text description for a single frame.
3. The format of each sentence is frame id, description. The frame id indicates the temporal position of the frame, ranging from 1 to 90.
Here are the original descriptions of this video: Documents
Here is the question: Question
Here is the memory: Memory
Your task is to determine whether these descriptions above can answer the question accurately, reasonably, and without contradiction.
If your answer is yes, please give me an reasonable explanation. the output will be as follows: **{"confidence": "1", "explanation": ["xxxx"]}**
If your answer is no, the confidence is 0, indicating the provided information is insufficient. Please give me a reasonable explanation for what frame is missing. For each frame identified as potentially relevant, provide a concise description focusing on essential visual elements(e.g., objects, humans, interactions, actions, and scenes) in the explanation. The output will be as follows: **{"confidence": "0", "explanation": ["xxxx"]}**
You must not provide any other response or explanation.

Assistant
"confidence": "0/1", "explanation": ["xxxx"]

Interaction agent is used to find potential missing key frames with different types of information. The interaction agent prompt in the EgoSchema benchmark is shown as below:

User
You are given some language descriptions of a first-person view video along with a question about the video.
1.The video is 3 minutes long, containing a total of 90 frames.
2. Each sentence in these language descriptions represents the text description for a single frame.
3. The format of each sentence is frame id, description. The frame id indicates the temporal position of the frame, ranging from 1 to 90.
Here are the original descriptions of this video: Documents
Here are the memory: Memory
To answer the following question: Question
Theses descriptions are insufficient and cannot answer this question accurately, reasonably, and without contradiction.
Your task is to determine which frame needs which type of information and can answer this question accurately, reasonably, and without contradiction.
The two types of information are as follows:
A: Given an image, get a detailed description of the image (image caption, just like what is shown in this image?)
B: Given an image, get a response to the above question (visual question answering)
Please note that frame selections range from 1 to 90. These frames (type_A) already have type A information and these frames (type_B) already have type B information, please note not to repeatedly select this type of information from these frames. Please note that the the key of frame only one number. The output must be as follows: **["frame": "1/2/3/.../90", "type": "A/B"]**

Assistant
["frame": "1/2/3/.../90", "type": "A/B"]

Finally, the answering agent is used to predict the answer once the video captions are sufficient. The answering agent is shown as below:

User

You are individual C, with others represented as O. Your task is to answer a question related to this video, choosing the correct option out of five possible answers. You are given some language descriptions of a first person view video along with a question about the video.
1.The video is 3 minutes long, containing a total of 90 frames.
2. Each sentence in these language descriptions represents the text description for a single frame.
3. The format of each sentence is frame id, description. The frame id indicates the temporal position of the frame, ranging from 1 to 90.
Here are the descriptions of this video: Documents
Please answer the following question: Question
Here are the choices. A: option1 B: option2 C: option3 D: option4 E: option5
The question has 5 choices, labeled as A, B, C, D, E. Please think step by step and write the best answer index. Note your final answer must be one of the letters (A, B, C, D, or E), the confidence must be one of the letters (1, 2, 3), please provide a concise one-sentence explanation for your chosen answer. the output must be the following format. You must not provide any other response or explanation.
{"final_answer": "xxx", "confidence": "xxx", "explaination": "xxx"}

Assistant

{"final_answer": "xxx", "confidence": "xxx", "explaination": "xxx"}

Details of LaViLa. For the experiments on EgoSchema [7], we utilize LaViLa [15] as the captioner, a CLIP-based captioning model. LaViLa processes input clips with a resolution of $4 \times 336 \times 336$ and is trained on the Ego4D dataset [3]. The original LaViLa training set contains 7,743 videos with 3.9 million video-text pairs, while the validation set includes 828 videos with 1.3 million video-text pairs. Since the EgoSchema dataset is cropped from Ego4D and designed for zero-shot evaluation, using the original LaViLa model could lead to overlap with EgoSchema videos, resulting in an unfair comparison. To address this, we use the retrained LaViLa model which same as LLoVi [14] that do not have any overlap with EgoSchema videos to avoid unfair comparison with other methods. The checkpoints are available at https://drive.google.com/file/d/1AZ5I4eTUAUBX31rL8jLB00vNrlFGWiv8/view.

Details of LLaVA-NeXT. For the experiments on MovieChat-1K [9] and Video-MME [2], we utilize LLaVA-NeXT [6] as the captioner, which is a frame-based captioning model. LLaVA-NeXT used in our framework has 7B parameters and the checkpoints are available at https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf. We also use the same LLaVA-NeXT model to augment key frames with different prompts in both datasets.

VideoAgent Reproduce Details. To evaluate the performance of VideoAgent [11] on MovieChat-1K benchmark [9] and make a fair comparison, we select the LLaVA-NeXT model as the captioner and preprocess videos by simply sampling them at 0.5 FPS. GPT-4 [1], *i.e.,* gpt-4-1106-preview, is used as the agent to predict answer, self reflect, and find missing information. The prompt for each agent is same as the original paper. The max interaction rounds is set to 3 and the initial sampled frames for whole video is set to 5, which is the default setting in the original paper. The EVA-CLIP-8B-plus model [10], a state-of-the-art CLIP model that includes a vision encoder with 7.5 billion parameters and a text encoder with 0.7 billion parameters, is used for frame retrieval to align with the original paper [11]. The above setting is used to the global mode and breakpoint mode to ensure reproducibility. The checkpoints of LLaVA-NeXT model are available at https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf. The website of GPT-4 API is available at https://openai.com/index/gpt-4-api-general-availability/. The checkpoints of EVA-CLIP-8B-plus model are available at https://huggingface.co/BAAI/EVA-CLIP-8B-448.

To evaluate the performance of VideoAgent [11] on Video-MME benchmark [2] and make a fair comparison, we preprocess videos by simply sampling them at 0.2 FPS and DeepSeek, *i.e.,* DeepSeek V2.5, is used as the agent to predict answer, self reflect, and find missing information. Except for those, the other settings are the same as the MovieChat-1K benchmark. The website of DeepSeek API is available at https://api-docs.deepseek.com/. Note that for the experiment with subtitles, we select the subtitles corresponding to the sampled frames to add to the LLM, rather than using all the subtitles. This is why the performance of VideoAgent is significantly lower than the other LLM-based methods under the *w subs* setting.

LLoVi Reproduce Details. To evaluate the performance of LLoVi [14] on Video-MME benchmark and make a fair comparison, we select the same LLaVA-NeXT model [6] as the captioner and preprocess videos by simply sampling them at 0.2 FPS. DeepSeek [4], *i.e.,* DeepSeek V2.5, is used as the agent to summary the sampled captions and predict the final answer. The summaries words are set to 500 and the prompt of each prompt is same as the original paper [14]. The website of DeepSeek API is available at https://api-docs.deepseek.com/. The checkpoints of LLaVA-NeXT model are available at https:

Sampling Rate	Accuracy (%)
1 FPS	61.6
0.5 FPS	**62.6**
0.25 FPS	58.8

Table 1. Performance of different sampling rate on EgoSchema.

//huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf.

Only Subs Reproduce Details. In Section 4.3 of our paper, we conduct an experiment where only the subtitle, question and options are input into the LLM (*i.e.,* DeepSeek [4]) to predict the answer. Specifically, the subtitles used in the LLM are aligned with the sampled frames, while we preprocess videos by simply sampling them at 0.2 FPS. The prompt for the LLM is shown in below:

User
This video's subtitles are listed below:
subtitle_1
subtitle_2
...
subtitle_n
Select the best answer to the following multiple-choice question based on the subtitles and summary. Respond with only the letter (A, B, C, or D) of the correct option.
Question
OptionA
OptionB
OptionC
OptionD

Assistant
A/B/C/D

where n represents the number of subtitles under the 0.2 FPS sampling.

2. More Ablation Studies

In this paper, we choose LaViLa [15] as the captioning model to convert videos into documents. we preprocess videos by simply sampling them at 0.5 FPS for EgoSchema. To evaluate the impact of different sampling rates on the model's performance. We conduct the experiments with different sampling rates (1 FPS, 0.5 FPS, and 0.25 FPS) to convert videos into documents on EgoSchema subset. Besides, we use GPT-3.5 [8] as the planning agent, interaction agent, and answering agent for the below comparisons.

Algorithm 1: DrVideo

Input: V, Q
▶ **Video-Document Conversion Module**;
$\text{Doc}_{init} \leftarrow \text{getDoc}(V, \phi_{vlm}, \mathcal{P})$;
$\mathcal{E}_{\text{doc}} \leftarrow \text{getEmbedding}(\text{Doc}_{init}, \phi_{emb})$;
▶ **Document Retrieval Module**;
Initialize $\mathcal{RT}$ as Q;
$\mathcal{E}_{\mathcal{RT}} \leftarrow \text{getEmbedding}(\mathcal{RT}, \phi_{emb})$;
$\text{topk\_doc} \leftarrow \text{Retrieval}(\mathcal{E}_{\mathcal{RT}}, \mathcal{E}_{\text{doc}}, K)$;
▶ **Document Augmentation Module**;
Initialize $\mathcal{AP}$;
$\mathcal{AD}_0 \leftarrow \text{Augment}(\text{topk\_doc}, \text{Doc}_{init}, \mathcal{AP})$;
$\mathcal{H} \leftarrow \text{addToMemory}(\text{topk\_doc})$;
▶ **Multi-Stage Agent Interaction**;
Initialize $i \leftarrow 0$;
while $i \leq I$ **do**
  Planning Agent:;
  $\mathcal{S}, R \leftarrow \text{checkSufficient}(\mathcal{AD}_i, \mathcal{H}_i, Q)$;
  if $\mathcal{S} == 1$ **then**
    break;
  else
    $\mathcal{H} \leftarrow \text{addToMemory}(\mathcal{H}, R)$;
    Interaction Agent:;
    $\mathcal{M} \leftarrow \text{FindMissInfo}(\mathcal{AD}_i, \mathcal{H}, Q)$;
    $N, \mathcal{AP} \leftarrow \mathcal{M}$;
    $\mathcal{H} \leftarrow \text{addToMemory}(\mathcal{H}, N)$;
    $i \leftarrow i + 1$;
    $\mathcal{AD}_i \leftarrow \text{Augment}(N, \mathcal{AD}_i, \mathcal{AP})$;
▶ **Answering Module**;
$P \leftarrow \text{GetAnswer}(\mathcal{AD}_i)$;
return P;

Table 1 presents the results and we have the following observations: (i) When the sampling rate is set to 1 FPS, the performance is lower than the sampling rate is set to 0.5 FPS. It indicates that a higher sampling frequency is not always better, as higher frequencies introduce more redundant information. (ii) When the sampling rate is set to 0.25 FPS, the performance is also lower than the sampling rate is set to 0.5 FPS. This suggests that a low sampling frequency leads to the loss of significant information, which in turn causes a decline in performance. (iii) When the sampling rate is set to 0.5 FPS, DrVideo achieves the best performance. This highlights the importance of selecting an appropriate sampling rate to reduce redundant information while retaining critical details for DrVideo. It also leaves room for future improvements by adaptively sampling the video based on the question.

Design two retrieval mechanisms. Leveraging text-semantic similarity to find keyframes is straightforward, which is however insufficient to identify some keyframes

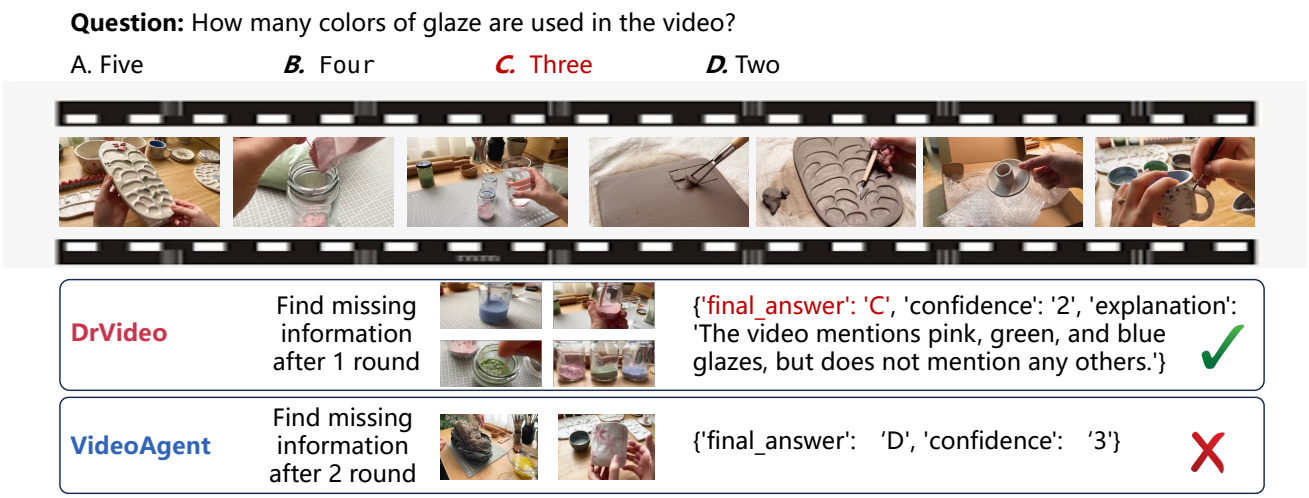

Figure 1. Case study on an instance from Video-MME. Long Case of DrVideo. This video contains 33 minutes.

that require contextual reasoning. Thus, we design the interaction loop to discover potential keyframes through iterative reasoning. Tab. 4 of the main paper presents the effectiveness of combining both retrieval modules and their complementarity.

More comparisons We compare two concurrent and unpublished works (VideoTree [13] and LifelongMemory [12]) on the EgoSchema and MovieChat-1k benchmarks. As reported below, DrVideo still performs better especially on MovieChat-1k, showing the effectiveness of the document retrieval method for processing long videos.

Method	LLM	EgoSchema		MChat-Glob.		MChat-Break.	
		Sub.	Full.	Acc.	Score	Acc.	Score
LifelongMemory	GPT-4	64.1	58.6	57.3	3.88	34.7	2.51
VideoTree	GPT-3.5	57.6	-	-	-	-	-
VideoTree	GPT-4	66.2	**61.1**	73.3	3.75	40.1	2.39
DrVideo (ours)	GPT-3.5	62.6	-	-	-	-	-
DrVideo (ours)	GPT-4	**66.4**	61.0	**93.1**	**4.41**	**56.4**	**2.75**

Table 2. Comparisons with LifelongMemory and VideoTree.

Effects of iterative rounds. In Fig. 3 of the main paper, to be consistent with VideoAgent, we manually control the number of iterative rounds in DrVideo and remove the planning agent. As for examining the effectiveness of the planning agent, we set different maximum numbers of iterative rounds (*i.e.,* 2, 5, and 10) and obtain accuracies of 62.6%, 62.6%, and 62.4%, respectively. This shows that the planning agent is robust and can properly terminate the loop.

3. Detailed Algorithm

In Algorithm 1, we present the algorithm behind DrVideo to give the reader a clearer understanding. The definition of the symbols has already been provided in the main text.

4. More Case Studies

Fig. 1 shows how DrVideo can accurately solve hour-long videos from the long split of Video-MME benchmark. The question is about figuring out the colors of glaze used in the video, which not only requires the model to have a comprehensive understanding of the video but also a clear understanding of its local details. DrVideo accurately identifies the necessary information and predicts the answer correctly, outperforming state-of-the-art models like VideoAgent. This highlight the potential of our document retrieval method in handle longer videos.

Besides, we also provide a failure case to explore the limitations of our DrVideo as shown in Fig. 2. The question is determine the overarching theme of the video, considering the activities performed by both characters. DrVideo, after undergoing key frame retrieval enhancement and multi-stage interaction loop continuation retrieval enhancement, evaluates whether the information collected so far can accurately answer the question. However, due to the inability of the VLM to accurately describe the video content, DrVideo makes incorrect judgments, ultimately leading to a wrong answer. This indicates that DrVideo heavily relies on the capabilities of both LLMs and VLMs. We believe DrVideo will be improved with the development of better LLMs and VLMs in the future.

5. Compared with LangRepo

Both LangRepo [5] and DrVideo convert raw videos into all-textual representations. LangRepo transforms raw video into multi-scale structured repository (*e.g.,* captions, entries, summaries) without keyframes retrieval

Question: What is the overarching theme of the video, considering the activities performed by both characters?
A. The overarching central theme presented in the video is that individuals can be both sociable and independent simultaneously. the visual content demonstrates that it is entirely possible to be both connected to others meaningfully and to savor solitary moments, emphasizing that it is crucial to find a harmonious balance between these two aspects.
B. The overarching theme of the video is that people can be both engaged in challenging activities and enjoying leisurely activities at the same time. the video shows that it is possible to be both productive and relaxed, and that it is important to find a balance between the two.
C. The primary, overarching theme presented in the video emphasizes that individuals can truly be both creative and practical simultaneously. the enlightening video demonstrates the realistic possibility of being both highly imaginative and remarkably efficient, while stressing the significance of discovering an equilibrium between these two essential aspects.
D. The overarching theme of the video is that people can be both ambitious and humble. the video shows that it is possible to be both driven and modest, and that it is important to find a balance between the two.
E. The primary overarching theme presented in the video is that individuals can simultaneously possess and exhibit both intelligence and emotional aspects. effectively, the video demonstrates that the coexistence of rational and intuitive qualities is feasible, emphasizing the significance of establishing equilibrium between these two crucial elements..

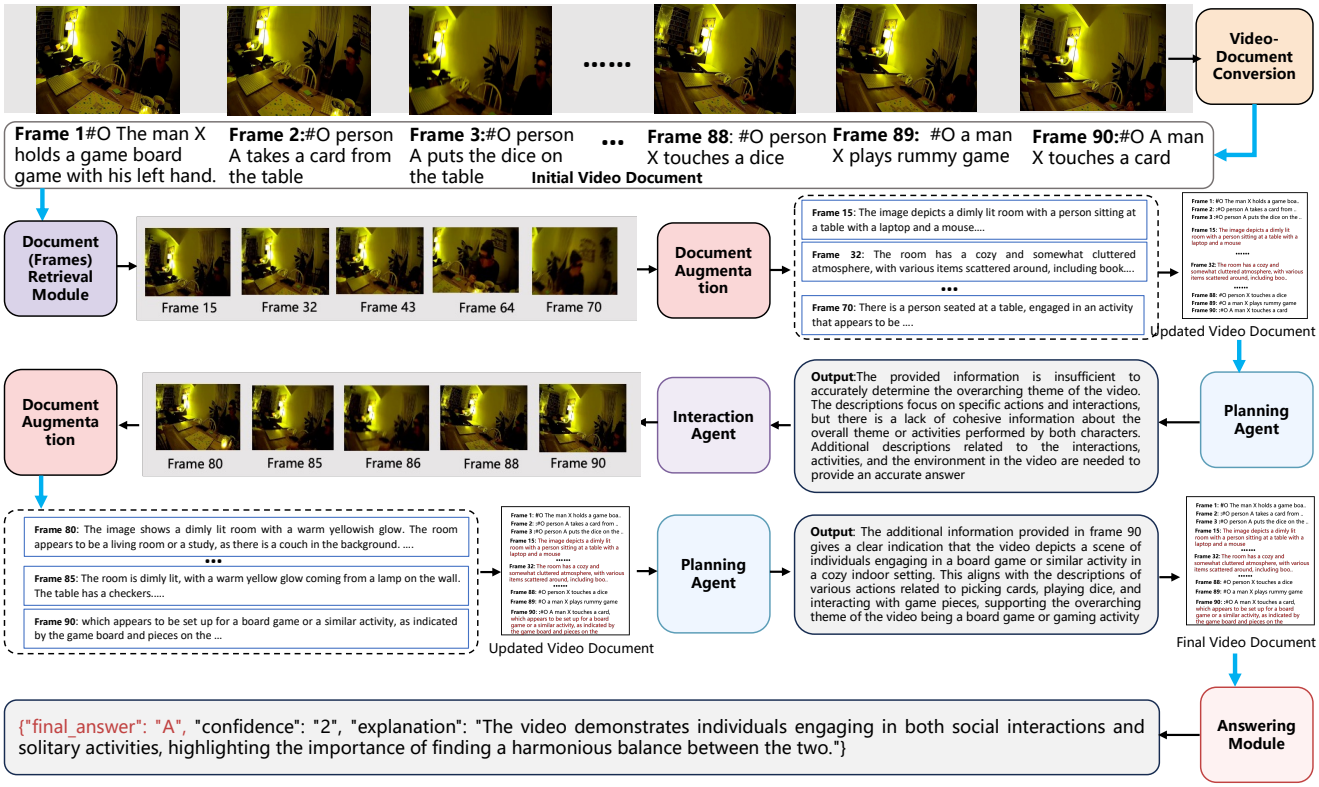

Figure 2. Case study on an instance from EgoSchema. Failure Case of DrVideo.

and augmentation. Instead, DrVideo designs two retrieval mechanisms to find keyframes and augment them with different types of information.

6. Limitation

Although Drvideo achieves impressive results on different long video benchmarks, it has several limitations. (i) The longest video DrVideo can handle depends on the maximum token length of the LLM, which is the bottleneck for much longer videos than the existing benchmarks, *e.g.,* 10 hours. (ii) It has space for further improvement on how to generate sufficient information while minimizing irrelevant information in the video document.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever compre-

hensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2

[3] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 18995–19012, 2022. 2

[4] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. 1, 2, 3

[5] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*, 2024. 4

[6] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024. 2

[7] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 1, 2

[8] OpenAI. GPT3.5. https://platform.openai.com/docs/models/gpt-3-5, 2021. 1, 3

[9] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 2

[10] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 2

[11] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024. 2

[12] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023. 4

[13] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024. 4

[14] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 2

[15] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6586–6597, 2023. 2, 3