DrVideo: Document Retrieval Based Long Video Understanding

Ziyu Ma[1,2]*, Chenhui Gou[2]*, Hengcan Shi[1,2], Bin Sun[1]†, Shutao Li[1],
Hamid Rezatofighi[2], Jianfei Cai[2]
[1] College of Electrical and Information Engineering, Hunan University,
[2] Data Science & AI Department, Faculty of IT, Monash University
{maziyu, sunbin611, shutao_li}@hnu.edu.cn
{chenhui.gou, hengcan.shi, hamid.rezatofighi, jianfei.cai}@monash.edu

Abstract

Most of the existing methods for video understanding primarily focus on videos only lasting tens of seconds, with limited exploration of techniques for handling long videos. The increased number of frames in long videos poses two main challenges: difficulty in locating key information and performing long-range reasoning. Thus, we propose DrVideo, a document-retrieval-based system designed for long video understanding. Our key idea is to convert the long-video understanding problem into a long-document understanding task so as to effectively leverage the power of large language models. Specifically, DrVideo first transforms a long video into a coarse text-based long document to initially retrieve key frames and then updates the documents with the augmented key frame information. It then employs an agent-based iterative loop to continuously search for missing information and augment the document until sufficient question-related information is gathered for making the final predictions in a chain-of-thought manner. Extensive experiments on long video benchmarks confirm the effectiveness of our method. DrVideo significantly outperforms existing LLM-based state-of-the-art methods on EgoSchema benchmark (3 minutes), MovieChat-1K benchmark (10 minutes), and the long split of Video-MME benchmark (average of 44 minutes). Code is available at https://github.com/Upper9527/DrVideo.

1. Introduction

Video understanding is a challenging task in computer vision, requiring the processing of spatio-temporal information and advanced reasoning abilities. Previous works have successfully processed short videos lasting around tens of seconds [8, 27, 32, 53, 69, 70]. However, how to deal with long video understanding remains unclear. Recent advancements in large language models (LLMs) [6, 10, 34, 41, 45, 48, 56] demonstrate strong abilities in language understanding and reasoning across long text sequences. These advancements have inspired the development of video-language models (Video-LMs) [25, 38, 68, 77] to address long video understanding issues.

Video-LMs typically encode a video as a sequence of visual tokens, which are then concatenated with language tokens into a long sequence and LLM is then used to understand this long sequence. Although Video-LMs improve video understanding, they have the following limitations. (i) They cannot process an entire video as input to a Video-LM (*e.g.,* the visual encoder OpenAI's CLIP-L-14 outputs 24×24 tokens for each image, while the max length of LLaVA-NeXT-Video [77] is only 8192). (ii) They typically perform uniform or random frame sampling at large strides to handle long videos without considering the content (*e.g.,* 16 for PLLaVA [68] and 6 for IG-VLM [25]), leading to potential key information loss. (iii) The simple concatenation of visual tokens increases the difficulty for the LLM to locate question-related (key) visual information within the long video sequence, complicating long-range reasoning across the vision token sequences.

Apart from the studies of encoding a video as a sequence of visual tokens, another line of research converts raw videos into captions and then utilities the long-range reasoning abilities of LLMs to predict the final answer. A pioneering work - LLoVi [74] uses a short-term video model [78] in conjunction with an LLM [40] to solve the long video understanding task. Given a long video, LLoVi first divides it into multiple short clips and converts them into short textual descriptions. Afterward, these short textual descriptions are summarized by the LLM, and the summary is finally used by LLoVi to answer the given question. Inspired by LLoVi, VideoAgent [61] designs an agent-based system to further leverage the reasoning abilities of LLMs and locate question-related video clips or images. Initially,

*Equal contribution.
†Corresponding author.

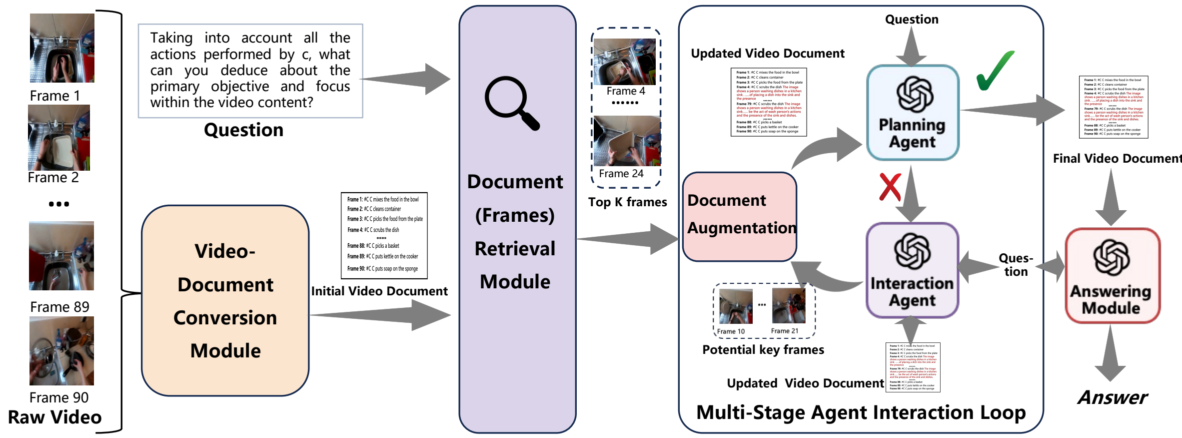

Figure 1. **Overview of our DrVideo framework.** It comprises five components: a video-document conversion module, a retrieval module, a document augmentation module, a multi-stage agent interaction loop and an answering module.

VideoAgent samples only a minimal number of frames from the video (*i.e.,* 5 frames), converting them into textual descriptions as input for the LLM. Then, through continuous feedback and interaction with the LLM [1], VLM [78], and CLIP [51], it gradually finds question-related frames in a coarse-to-fine manner via image-text similarity. Finally, VideoAgent answers the question based on these question-related frames. Despite these LLM-based methods overcoming the challenge of performing long-range reasoning and achieving promising results, they still suffer from the following problems.

- These methods struggle to accurately identify question-related frames. For instance, VideoAgent [61] performs sparse sampling on long videos (*e.g.,* selecting 5 frames) and subsequently searches for keyframes from the unsampled frames in a coarse-to-fine manner based on the initial sparse space via CLIP image-text similarly given the missing question-related information inferred by the LLM, which improves inference efficiency but unavoidably at the cost of lacking a holistic grasp of video content, particularly when a long video consists of multiple events (*e.g.,* over 100 events) with significant content variations, and since the CLIP text encoder is trained with simple language structures and limited context (the max length is 77), it struggles to understand complex video semantics, limiting its retrieval accuracy.
- These methods convert videos into captions, addressing long-range reasoning but overlooking critical information loss. Specifically, even if question-related frames are identified, the generated captions cannot cover all the essential information in the key frames. For example, the caption *"a woman is looking in the mirror"* does not contribute to answering the question *"what is the woman wearing when standing in front of the mirror in the video"*. In this case, LLM can only make an aimless and biased surmise to answer the question.

To solve the above problem, we turn to how human understands a long video. In particular, a human needs to gain a general understanding of the video, identifying events across segments. Then, the human locates question-related segments, examines them in detail, and answers the question based on the extracted information. Inspired by this, in this paper, we propose DrVideo, a document retrieval-based system for long video understanding that progressively finds key frames and generates a question-related augmented document for LLMs to make confident predictions. Our key idea is to convert long video understanding into long document understanding and leverage the impressive capabilities of LLMs in key information location [46] and long-range reasoning [12] on various NLP tasks. Note that long-document retrieval in our paper means passage retrieval within a single long document, not retrieval from a corpus consisting of long documents.

Specifically, DrVideo transforms raw long videos into long documents using a pretrained frame/clip-level visual captioner (*e.g.,* LLaVA-NeXT [33]) to make the LLM have a holistic grasp of video content. Different from VideoAgent [61], which performs sparse sampling and identifies key frames via image-text similarity without information augmentation, DrVideo performs dense sampling and proposes a retrieval module and a multi-stage agent interaction loop that dynamically finds the potential missing information through sequential steps of reasoning and then interacts with a document augmentation module to request augmentation of the information within language space. Once sufficient question-related information is gathered, it feeds into the answering module to get the final prediction.

DrVideo achieves significant performance gains (3.8% on EgoSchema subset [36], 8.3% on MovieChat-1K global mode, 24.8% on MovieChat-1K breakpoint mode [49], and 6.3% on the long split of Video-MME [15]) over the existing LLM-based state-of-the-art methods [14, 37, 61, 74] across different long video benchmarks, from 3 minutes to 10 minutes and longer than 1 hour. Surprisingly, combining the subtitles on the long split of Video-MME benchmark, DrVideo obtains 71.7% accuracy, outperforming many heavily-engineered large-scale proprietary models (*e.g.,* Gemini 1.5 Flash [55], GPT-4o mini [44], and GPT-4V [42]). This highlights the great potential of DrVideo. Moreover, DrVideo is a training-free framework and is researcher-friendly, as our results can be replicated on an RTX 4090 with a reasonable number of GPT accesses.

The contribution of this work can be summarized as:

- We propose DrVideo, the first document-retrieval-based system designed for the long video understanding task by converting it into a long-document retrieval and understanding task to effectively leverage the power of large language models.
- Different from previous works, we propose a retrieval module and a new multi-stage agent interaction loop that dynamically finds the potential missing information and augments these information within language space.
- DrVideo outperforms the existing LLM-based state-of-the-art methods across different long video benchmarks (*i.e.,* EgoSchema [36], MovieChat-1K [49], and the long split of Video-MME [15]) by a margin.

2. Related Work

Long Video Understanding. Modeling long videos, which are several minutes or more in length, generally requires advanced temporal modeling, resulting in complex model designs. LF-VILA [52] introduces a Temporal Window Attention (HTWA) mechanism to capture long-range dependencies in long videos. MeMViT [67] and MovieChat [49] employ a memory-based design to save question-related information from previous video segments. Some other approaches employ space-time graphs [21, 62] or relational space-time modules [71] to capture spatio-temporal dependencies from the raw long videos. Recently, S4ND [39], ViS4mer [22], and S5 [58] have used Structured State-Space Sequence (S4) [18] layers to capture long-range dependencies in videos. Different from these methods, our DrVideo does not design a complex module to perform long video understanding. Instead, we develop a document retrieval-based system with LLM for zero-shot long video understanding.

LLMs for Video Understanding. The recent rise of LLMs [73] and VideoChat [30] align the visual features extracted by pretrained visual models to LLMs and apply them to video understanding. Video ChatCaptioner [7] and ChatVideo [57] utilize LLMs to represent videos and engage users through dialogues, respectively. VidIL [64] applies the image-level models to video understanding tasks via few-shot learning. In addition to short-term video understanding, recent studies [4, 11, 32] have explored LLMs for long-range video modeling. For instance, GPT-4 is applied in various long-range video modeling tasks in [32], though quantitative evaluation is limited. Meanwhile, the research in [11] focuses on movie datasets with minimal visual analysis [36], relying largely on speech and subtitles. In contrast, DrVideo focuses on vision modality in multiple benchmarks.

LLM Agents. In parallel, the computer vision community has started exploring the use of LLMs as agents in various vision tasks such as GUI understanding and robot navigation [5, 13, 20, 53]. In the realm of long video comprehension, initial efforts have employed an agent-like approach, where LLMs interact with external tools or integrate additional functionalities [14, 17, 53, 61, 65, 72]. In contrast to these approaches, our method, DrVideo, reconceptualizes long video understanding as a process of document retrieval, augmentation, and understanding, to leverage the strong capability of LLMs.

3. Methodology

In this section, we detail our proposed document retrieval-based system for long video understanding. As illustrated in Figure 1, given a long video and a question about the video, DrVideo first translates the long video into a long document, referred to as the initial video document. Then, the retrieval module identifies the top K key frames by calculating the similarity between the question and the initial video document. The document augmentation module enriches the information of these key frames and adds it to the initial video document, creating a new video document (named the updated video document), which is used as the starting point for the subsequent multi-stage agent interaction loop.

The loop contains two distinct agents: the planning agent and the interaction agent. The planning agent judges whether the updated video document is sufficient for answering the question. If not, the updated video document is fed into the interaction agent to dynamically find missing key frames. The interaction agent interacts with the document augmentation module to request the required information of these key frames. The output of the document augmentation module is added to the current document to obtain the latest updated video document. The new video document is then sent to the planning agent again to further judge whether the current information is sufficient. This searching and interaction loop continues until the planning agent considers the current information sufficient for answering the question or the maximum iteration is reached. After the loop ends, the final video document and the related question are given to the answering module to get the final prediction.

3.1. Video-Document Conversion Module

By translating each single video frame V_t (or a short video clip) into a short description S_{V_t}, we convert an input long video into the initial video document Doc_{init}:

$$Doc_{init} = \{\{1, S_{V_1}\}, \{2, S_{V_2}\}, \ldots, \{T, S_{V_T}\}\} \quad (1)$$

where T is the length of the video. Specifically, S_{V_t} is generated by a large vision-language captioning model ϕ_{vlm} (*e.g.,* LLaVA-NeXT [33]) , *i.e.*, $\mathrm{S}_{V_t} = \phi_{vlm}(\mathcal{P}, V_t)$, where $\mathcal{P}$ is the prompt for requesting a short description about the image, *e.g.*, *describe the picture in no more than 50 words*. The captioning model used in DrVideo can be replaced by other captioning models (*e.g.,* LaViLa [78], BLIP-2 [29]). The experimental results with different captioning models can be found in Table 6.

3.2. Document (Frames) Retrieval Module

After obtaining Doc_{init}, we introduce a document retrieval module to identify the top question-related frames by calculating the similarity between the question and the whole document. Specifically, we use the OpenAI embedding model ϕ_{emb} [43] to obtain the vector representation of the initial document, *i.e.*, $\mathcal{E}_{\text{doc}} = \phi_{emb}(Doc_{init})$. Given a specific retrieval text $\mathcal{RT}$, *i.e.*, the question Q, the retrieval module first computes the embedding $\mathcal{E}_{\mathcal{RT}} = \phi_{emb}(\mathcal{RT})$, and then retrieves the top K frames based on the cosine similarity between $\mathcal{E}_{\mathcal{RT}}$ and $\mathcal{E}_{\text{doc}}$:

$$topk\_doc = \arg\operatorname{top}_t K \cos(\mathcal{E}_{\mathcal{RT}}, \mathcal{E}_{\text{doc}_t}) \quad (2)$$

where $\mathcal{E}_{\text{doc}_t}$ is the embedding of the t-th description S_{V_t}.

3.3. Document Augmentation Module

For each $t' \in topk\_doc$, we use the LLaVA-NeXT model [33] with different prompts $\mathcal{AP}$ (augmented prompts) to generate a detailed description $L_{V_{t'}}$, *i.e.*, $L_{V_{t'}} = \phi_{vlm}(\mathcal{AP}_{t'}, V_{t'})$. The initial $\mathcal{AP}_{t'}$ is a general prompt: *If there are factual errors in the question, provide a precise description of the image; if not, proceed to answer the question:* $\{Q\}$. The updated video document $\mathcal{AD}$ becomes

$$\mathcal{AD} = \left\{ \begin{cases} \{t, S_{V_t}, L_{V'_t}\} & \text{if } t = t' \\ \{t, S_{V_t}\} & \text{otherwise} \end{cases} \middle| t = 1, \ldots, T \right\}. \quad (3)$$

Note that K is significantly smaller than the total number of frames (*e.g.*, $K = 5$ compared to $T = 90$), so the additional descriptions will not substantially increase the overall length of the document.

3.4. Multi-Stage Agent Interaction Loop

Apart from the retrieved key frames, we further introduce a multi-stage agent interaction loop to dynamically find other potential key frames and augment them with different types of information for the final answer. We employ two distinct agents: planning and interaction agents, each tailored for a specific task. Both agents utilize the same LLM, *i.e.,* GPT-3.5, but are given different prompts and reason varying types of information. The detail is shown in Fig. 2

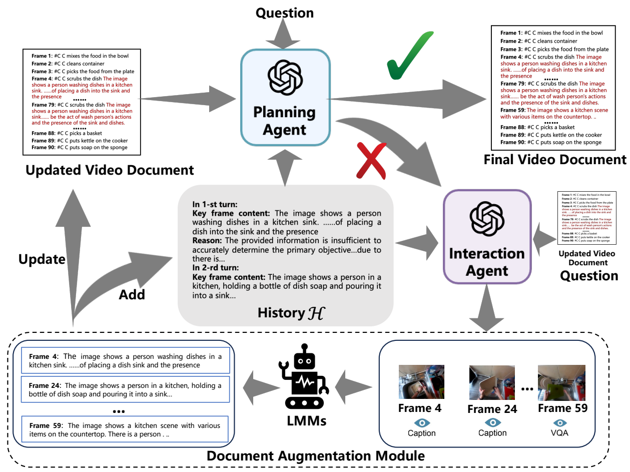

Figure 2. **Illustration of the multi-stage agent interaction loop and answering module.** There are two agents in the multi-stage agent interaction loop: a planning agent to plan the next step and an interaction agent to dynamically find missing information and interact with the document augmentation module.

Planning Agent. Given the question Q and the updated video document $\mathcal{AD}_i$ (where i denotes the iteration step), along with the analysis history $\mathcal{H}$ from all previous steps (initialized as empty $\{\}$), the planning agent first determines whether the updated video document is sufficient to generate a confident answer. If the information is sufficient, the process moves to the answering module to get the final answer. If not, the agent provides an analysis of why the updated video document $\mathcal{AD}_i$ is inadequate. The agent then updates $\mathcal{H}_i$ with its analysis and passes the updated $\mathcal{H}_i$ to the interaction agent.

Interaction Agent. Given the current video document $\mathcal{AD}_i$ and the updated $\mathcal{H}_i$, the interaction agent finds N frames with missing details crucial for answering Q:

$$\forall n \in N, n \in T, n \notin \text{topk\_doc}, \text{and} N < K. \quad (4)$$

In addition to identifying the missing N key frames, the interaction agent also determines the augmented information type for each $n \in N$. To achieve it, we give a task-specific prompt to the interaction agent, *i.e., 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).* After obtaining N

Method	(M)LLM	Subset	Fullset
MoReVQA [37]	PaLM-2	-	51.7
Vamos [60]	GPT-4	51.2	48.3
ProViQ [9]	GPT-3.5	57.1	-
LLoVi [74]	GPT-3.5	57.6	50.3
IG-VLM [25]	GPT-4V	59.8	-
VideoAgent [61]	GPT-4	60.2	54.1
MVU [47]	Mistral-13B	60.3	37.6
LLoVi [74]	GPT-4	61.2	-
VideoAgent [14]	GPT-4	<u>62.8</u>	<u>60.2</u>
DrVideo (ours)	GPT-3.5	**62.6**	-
DrVideo (ours)	GPT-4	**66.4**	**61.0**

Table 1. Results on EgoSchema compared to existing LLM-based state-of-the-art methods.

Method	Global		Breakpoint	
	Acc.	Score	Acc.	Score
Open-Source MLLM				
Video LLaMA [75]	51.7	2.72	39.1	2.11
Video Chat [30]	57.8	3.08	46.1	2.32
Video-ChatGPT [35]	47.6	2.89	48.0	2.43
MovieChat [49]	68.3	3.15	48.3	2.46
MovieChat+ [50]	71.2	<u>3.51</u>	<u>49.6</u>	<u>2.62</u>
Based on Open-source Captioners and Proprietary LLMs				
MM-VID [32]	58.6	2.86	10.4	0.56
LLoVi [74]	58.3	2.87	17.8	1.03
VideoAgent* [61]	65.4*	3.12*	31.6*	2.05*
Sullam Jeoung, et al [23]	<u>84.8</u>	-	-	-
DrVideo (ours)	**93.1**	**4.41**	**56.4**	**2.75**

Table 2. Performance comparison on the MoiveChat-1K [49] benchmark against state-of-the-art methods. * represents re-implemented results, implementation details refer to appendix.

frames along with their required information type, the agent interacts with the document augmentation module to enrich the information of these frames to update the current video document for the next step.

3.5. Answering Module

Given the final video document $\mathcal{AD}_{final}$, we employ another agent to provide a prediction using a zero shot chain-of-thought (CoT) approach [26, 66]. The answering module outputs the corresponding answer, the confidence score, and the reasoning behind the answer. Besides improving prediction accuracy, the CoT approach allows us to trace inference steps, ensuring transparency and explainability in the decision-making process.

4. Experiments

In this section, we first introduce the datasets and implementation details and then present the results and ablations of our proposed DrVideo.

4.1. Datasets and Metrics

In our experiments, we evaluate our DrVideo using three well-established datasets, emphasizing its zero-shot long video understanding capabilities.

EgoSchema. EgoSchema dataset [36] consists of 5000 multiple-choice questions sourced from 5000 three-minute egocentric videos. This dataset only has a subset of 500 questions with publicly accessible labels, while the full set is evaluated on the leaderboard.

MovieChat-1K. MovieChat-1K [49] is a longer video understanding benchmark, which contains 1000 videos from movies and TV shows, each video lasting approximately 10 minutes. This dataset has two modes: global mode and breakpoint mode. The global mode refers to analyzing the entire video to understand its overall content and context, while the breakpoint mode focuses on analyzing specific frames or scenes.

Video-MME. Video-MME [15] is a recently proposed benchmark for comprehensive video analysis evaluation. We evaluate DrVideo on the ”long-term videos” split of this dataset (long split), where video lengths vary from 30 to 60 minutes, with an average duration of 44 minutes.

Since EgoSchema and Video-MME are multi-choice tasks, we use accuracy as the evaluation metric for both datasets. MovieChat-1K focuses on open-ended questions, we use GPT-assisted evaluation to assess both the accuracy (true/false) and quality (score 0-5) of the models. We select Gemini-Pro [54] as the evaluation assistant and use the same prompt [38] to conduct a fair comparison.

4.2. Implementation Details

For EgoSchema [36] dataset, we choose LaViLa [78] as the captioning model to convert videos into documents. Note that the training data of our LaViLa model does not include EgoSchema videos, which is the same as LLoVi [74]. For MovieChat-1K and Video-MME benchmarks, LLaVA-NeXT [33] is used as our captioning model to generate brief descriptions for each frame. We preprocess videos by simply sampling them at 0.5 FPS for EgoSchema [36] and MovieChat-1K [49] and 0.2 FPS for Video-MME [15]. LLaVA-NeXT [33] is chosen to augment key frames in both datasets via different prompts. In the comparison experiments, GPT-4 [1], *i.e.*, gpt-4-1106-preview, is used as the agent to evaluate the performance of our DrVideo on EgoSchema and MovieChat-1K. DeepSeek [19], *i.e.*, DeepSeek V2.5, is used as the agent to evaluate the performance on Video-MME. For the ablation study, we use GPT-3.5 [40], *i.e.*, gpt-3.5-turbo-1106, to evaluate the effectiveness of our DrVideo due to the API cost.

4.3. Main Results

Comparative results on EgoSchema. Table 1 compares our DrVideo with existing LLM-based state-of-the-

Method	Frames	w/o Subs	w Subs
Proprietary MLLM			
Claude 3.5 Sonnet [2]	20	51.2	54.7
GPT-4V [42]	10	53.5	56.9
GPT-4o mini [44]	250	58.6	63.4
Gemini 1.5 Flash [55]	0.5 FPS	61.1	68.8
GPT-4o [44]	384	65.3	72.1
Gemini 1.5 Pro [55]	0.5 FPS	**67.4**	**77.4**
Open-Source MLLM			
LongVA [76]	128	46.2	47.6
VITA-8x7B [16]	32	48.6	50.9
LLaVA-OneVision-72B [28]	32	60.0	62.4
Qwen2-VL-72B [3]	768	62.2	74.3
Based on Open-source Captioners and Proprietary LLMs			
VideoAgent* [61]	7.6	40.2*	44.4*
LLoVi* [74]	0.2 FPS	45.4*	67.7*
DrVideo (ours)	0.2 FPS	51.7	71.7
-only subs	0.2 FPS	-	68.5

Table 3. Video-MME long split results. *w/o subs* represents the model without any subtitles. *w subs* represents that the subtitles corresponding to the sampled frames are both fed into the LLM to evaluate the performance. Our DrVideo outperforms a strong proprietary MLLM (Claude 3.5 Sonnet) and some open-source MLLMs (*e.g.,* VITA-8x7B). * represents re-implemented results, implementation details refer to appendix.

RM	MSAIL	CoT	Acc.(%)
✓	✓	✓	**62.6**
✓	✓	✗	62.2
✓	✗	✓	60.6
✗	✓	✓	59.4
✗	✗	✓	57.4

Table 4. Ablation results on different combinations of the retrieval module (RM), the multi-stage agent interaction loop (MSAIL), and the CoT on EgoSchema.

VQA	Caption	Acc. (%)
✓	✓	**62.6**
×	✓	60.4
✓	×	61.8

(a) Performance with different types of information

Top-K	Acc. (%)
5	**62.6**
10	61.4
20	60.6

(b) Performance with different Top-K frames

Table 5. Ablation results with different settings on EgoSchema.

art methods [9, 14, 25, 37, 47, 60, 61, 74] on EgoSchema benchmark [36]. Specifically, compared with LLoVi [74] and VideoAgent [61] that leverage the same visual captioner (LaViLa [78]) and LLM (GPT-4), DrVideo significantly outperforms these methods by 5.2% and 6.2% respectively. Compared with VideoAgent [14] which uses video-specific models (ViCLIP from InternVid [63] and Video-LLaVA [31]), DrVideo still performs better and achieves 3.6% higher accuracy on the subset evaluation. Moreover, DrVideo significantly outperforms IG-VLM [25] by 6.6%, which uses strong MLLM, *i.e.,* GPT-4V. These results confirm our idea of converting the long-video understanding task into a long-document understanding task is feasible and our DrVideo can find the question-related key frames more accurately.

Comparative results on longer video benchmarks. To further emphasize the advantage of our method on longer videos, we evaluate our DrVideo on the other two video benchmarks, *i.e.,* MoiveChat-1K (10 minutes) [49], and the long split of Video-MME (average 44 minutes) [15]. Table 2 presents the comparative results of our DrVideo and the current state-of-the-art methods [30, 38, 49, 74, 75] on the MovieChat-1K benchmark. Table 3 shows the comparative results of our DrVideo on the Video-MME benchmark, with comparisons against three types of methods: proprietary MLLMs [2, 42, 44, 55], open-source MLLMs [3, 16, 28, 76], and methods based on open-source captioners and proprietary LLMs [61, 74].

Compared to methods like LLoVi [74] and VideoAgent [61] that are based on open-source captioners (*e.g.,* LLaVA-NeXT [33]) and proprietary LLMs (*e.g.,* GPT-4 [1]), DrVideo achieves a 8.3% improvement in global mode and a 24.8% improvement in breakpoint mode on the MovieChat-1K benchmark in Table 2. It also achieves at least 6.3% improvement under the w/o subtitle setting on the Video-MME benchmark in Table 3. This indicates that DrVideo can accurately identify question-related frames and avoid critical information loss through augmentation, highlighting the potential of document retrieval methods for processing longer videos. Compared to proprietary MLLMs and open-source MLLMs, DrVideo outperforms Claude 3.5 Sonnet [2] by 0.5% and some open-source MLLMs (*e.g.,* VITA-8x7B [16]) under the w/o subtitle setting on Video-MME benchmark. It also achieves a 21.9% improvement in global mode and a 6.8% improvement in breakpoint mode on MovieChat-1K benchmark.

Moreover, combining the subtitle on the Video-MME benchmark, DrVideo achieves 71.7% accuracy and outperforms many heavily-engineered large-scale proprietary models (*e.g.,* Gemini 1.5 Flash [55], GPT-4o mini [44], and GPT-4V [42]). To better understand this significant improvement, we conduct an experiment where only the subtitle, question and options are input into the LLM (*i.e.,* DeepSeek [19]) to predict the answer and the accuracy is 68.5%. This suggests: (i) subtitles contain extensive question-related information, playing a crucial role in understanding long videos; (ii) compared with LLoVi and VideoAgent, DrVideo enhances visual information via document retrieval and augmentation, effectively integrating subtitles and resulting in a 3.2% improvement.

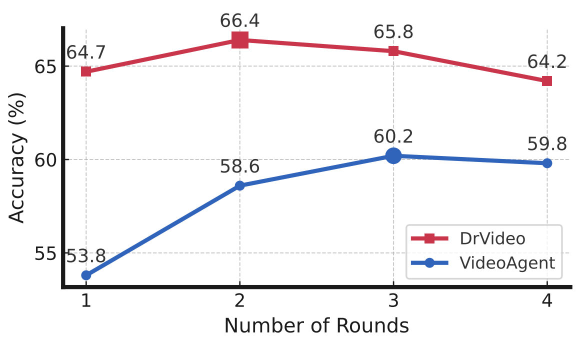

Figure 3. Performance of different rounds with DrVideo and VideoAgent [61] on EgoSchema. To align with VideoAgent, GPT-4, *i.e.,* gpt-4-turbo-1106-preview, is used as the LLM agents.

Captioning Model	Type	Acc.(%)
LaViLa [78]	Clip-based	**62.6**
LLaVA-NeXT [33]	Frame-based	61.2
BLIP-2 [29]	Frame-based	59.6

Table 6. Performance of different VLMs on EgoSchema.

4.4. Ablation Studies

Unless otherwise specified, we use GPT-3.5 as the default setting in the below experiments on EgoSchema subset [36].

Effects of different components of DrVideo. We conduct experiments with different combinations of the individual components in DrVideo, including the retrieval module, the multi-stage agent interaction loop, and the CoT used in the answering module. The results are presented in Table 4. We can see that: (i) without the retrieval module and the agent loop, the performance of our DrVideo is similar to LLoVi [74]; (ii) By adding the retrieval module and then the agent loop, the performance of DrVideo improves from 57.4% to 60.6% and then to 62.6%. This demonstrates that the retrieval module can identify key frames via measuring semantic similarity, while the agent loop can locate additional key frames through contextual inference; (iii) the additional CoT contributes to the accuracy improvement from 62.2% to 62.6%, demonstrating its effectiveness.

Effects of different types of augment information. To better enhance the required information for each potential key frame, we define two types of augment information, namely VQA (question-related information) and caption (general detailed information). Table 5a presents the performance of DrVideo with different types of information. We can see that when only one type of information (VQA or caption) is used for augmentation, the model's performance declines. This suggests the necessity of adaptively augmenting different types of information for each potential key frame.

Effects of different initial top K key frames. By default, the number of the initial key frames is set as $K = 5$. Intuitively, more frames should lead to better performance at the cost of increasing the complexity. However, our ablation in Table 5b shows an opposite trend, *i.e.*, more key frames resulting in worse performance. This suggests that enhancing more frames is not necessarily beneficial, as it can introduce a large amount of noisy and irrelevant information that interferes with the LLM's judgment.

LLMs	Type	Size	Acc.(%)
Mistral-8x7B [24]	Open-Source	8x7B	47.6
DeepSeek [19]	Proprietary	N/A	61.2
GPT-3.5 [40]	Proprietary	N/A	62.6
GPT-4 [1]	Proprietary	N/A	**66.4**

Table 7. Performance of different LLMs on EgoSchema.

Effects of iterative rounds. We also examined the impact of the iterative rounds I on model performance. Fig. 3 shows the performance of different iterative rounds with DrVideo and VideoAgent [61]. We have the following observations. (i) DrVideo consistently outperforms VideoAgent, reaching the peak performance at $I = 2$, while VideoAgent peaks at $I = 3$. This indicates that DrVideo can locate the information needed to answer questions in fewer interactions and with higher quality, underscoring the superiority of our document retrieval-based framework. (ii) When the number of interactions exceeds two, DrVideo's performance declines. This is likely due to more noisy and irrelevant information being introduced into the LLM with the increase of the iterations. This further confirms that excessive information is not necessarily beneficial.

Effects of different foundation models. DrVideo incorporates two types of foundational models — large language model (LLM), and visual language model (VLM), where VLM is used as the frame captioner while LLM is used as the planning agent, the interaction agent, and the answering agent. To assess the impact of the captions produced by different VLMs, we examine three state-of-the-art VLMs: frame-based BLIP-2 [29], LLaVA-NeXT [33], and clip-based LaViLa [78], as presented in Table 6. We observe that LaViLa achieves the best performance, while BLIP-2 is the worst. It indicates the superiority of the clip-based model which can capture motion information in videos and is more suited for video understanding.

To evaluate the effectiveness of DrVideo with different LLMs, Table 7 shows the results of DrVideo using open-source LLM (Mistral-8x7B [24]) or Proprietary LLM (DeepSeek [19], GPT-3.5 [40], or GPT-4 [1]) as the planning, interaction and answering agents. From the results, we can see that: (i) different LLMs in DrVideo exhibit significant performance gaps and the proprietary LLM significantly outperforms the open-source LLM, indicating DrVideo's dependency on the choice of LLM; (ii) as the LLM's capability increases, the performance of DrVideo

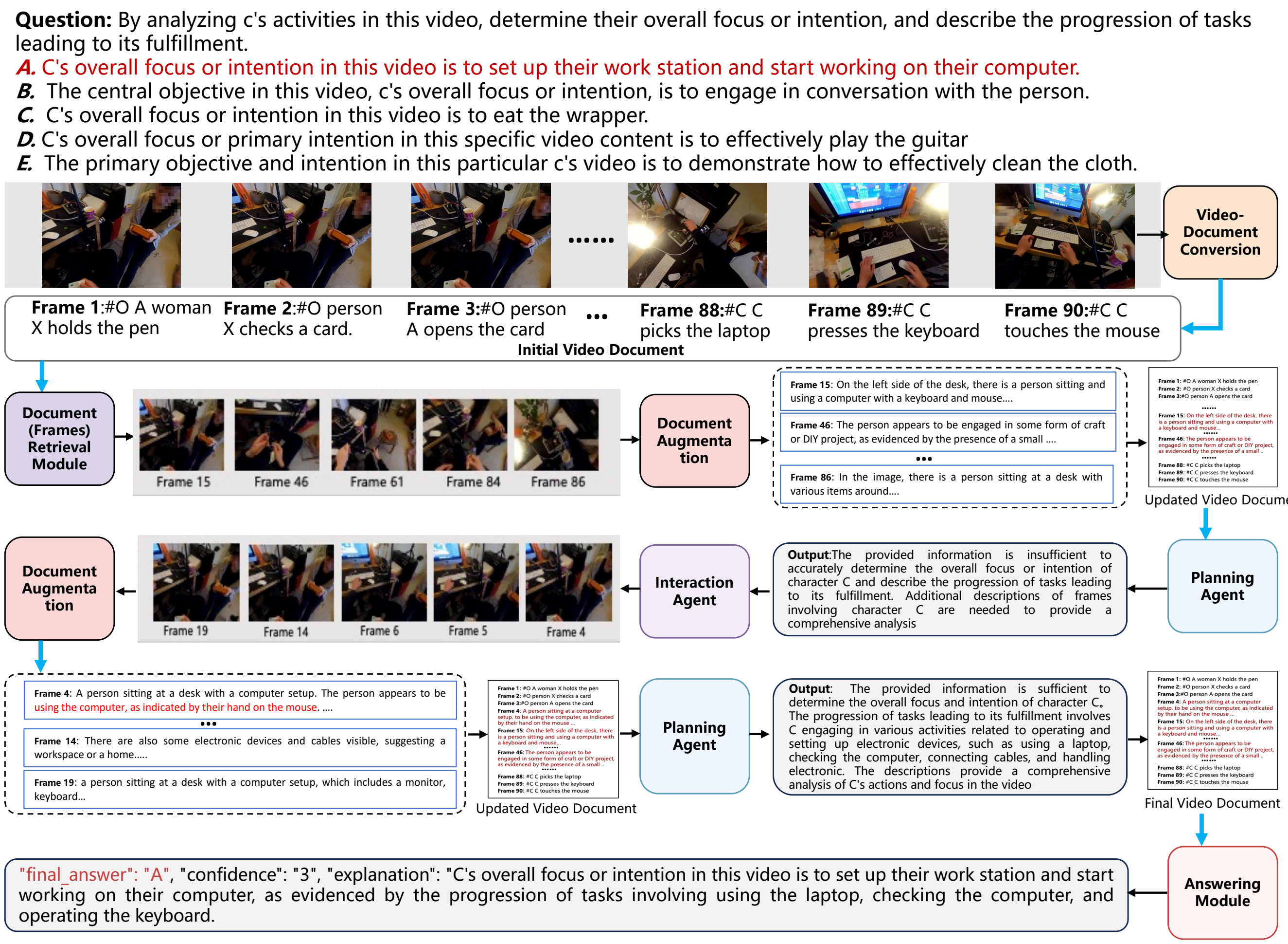

Figure 4. Case study on an instance from EgoSchema. DrVideo accurately identifies key frames and chooses the correct answer.

improves, demonstrating the great potential of DrVideo. It's worth noting that our main contribution is the introduction of a framework for translating the long video understanding into a document retrieval and understanding task to effectively leverage the power of large language models, rather than the employment of any specific models. Our DrVideo can be further improved by combining with better LLM and VLM models (*e.g.,* tarsier [59]).

Case study. Here we give a detailed example from EgoSchema [36] in Fig. 4 to qualitatively demonstrate the capability of our DrVideo. Specifically, the question is to determine the overall focus or intention and describe the progression of tasks leading to its fulfillment. DrVideo first converts the raw video into an initial video document, which is then processed by the frame retrieval module to obtain key frames (frames 15, 46, 61, 84, and 86) and is augmented with detailed key frame descriptions. The updated video document is then fed into the planning agent to assess if sufficient information is available for answering, *i.e.*, the provided information is sufficient to infer character C's main focus or intent. This updated document, along with the rationale, is input to the interaction agent to identify the additional key frames (frames 4, 5, 6, 14, and 19 here). After further augmentation, the updated document is sent back to the planning agent, for which the agent confirms its sufficiency. Finally, the answering module takes in the final video document and predicts the correct answer.

5. Conclusion

We have proposed DrVideo, a document retrieval-based system for long video understanding. Different from previous LLM-based methods, DrVideo proposes to adapt long-video understanding to long-document understanding and searches for missing information via document retrieval and multi-stage agent loop. Extensive comparative experiments and ablation studies on the three challenging datasets, *i.e.*, EgoSchema, MovieChat-1K, and Video-MME, demonstrate the effectiveness of DrVideo. We believe our work can serve as a strong baseline to stimulate further developments on the topic of long video understanding.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62221002, 62171183, U24A20327, 62401202), the Joint Fund of the Ministry of Education of China, and the China Scholarship Council (CSC) under Grant No. 202306130148.

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. 2, 5, 6, 7

[2] Anthropic. Claude 3.5 Sonnet: Enhanced Conversational AI. https://www.anthropic.com/index/claude-3.5-sonnet/, 2024. 6

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 6

[4] Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. *arXiv preprint arXiv:2305.09758*, 2023. 3

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 3

[6] 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. *Adv. Neural Inform. Process. Syst.*, pages 1877–1901, 2020. 1

[7] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023. 3

[8] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 10739–10750, 2023. 1

[9] Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. Zero-shot video question answering with procedural programs. *arXiv preprint arXiv:2312.00937*, 2023. 5, 6

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[11] Jiwan Chung and Youngjae Yu. Long story short: a summarize-then-search method for long video question answering. In *Brit. Mach. Vis. Conf.*, 2023. 3

[12] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024. 2

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[14] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024. 2, 3, 5, 6

[15] 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 comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2, 3, 5, 6

[16] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 6

[17] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023. 3

[18] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 3

[19] 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. 5, 6, 7

[20] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023. 3

[21] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. *arXiv preprint arXiv:1905.05143*, 2019. 3

[22] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *Eur. Conf. Comput. Vis.*, pages 87–104, 2022. 3

[23] Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024. 5

[24] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 7

[25] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-

shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024. 1, 5, 6

[26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Adv. Neural Inform. Process. Syst.*, pages 22199–22213, 2022. 5

[27] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7331–7341, 2021. 1

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. Int. Conf. Mach. Learn.*, 2023. 4, 7

[30] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 3, 5, 6

[31] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6

[32] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023. 1, 3, 5

[33] 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, 4, 5, 6, 7

[34] Ziyu Ma, Shutao Li, Bin Sun, Jianfei Cai, Zuxiang Long, and Fuyan Ma. Gerea: Question-aware prompt captions for knowledge-based visual question answering. *arXiv preprint arXiv:2402.02503*, 2024. 1

[35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5

[36] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 2, 3, 5, 6, 7, 8

[37] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 13235–13245, 2024. 2, 5, 6

[38] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023. 1, 5, 6

[39] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Adv. Neural Inform. Process. Syst.*, pages 2846–2861, 2022. 3

[40] OpenAI. GPT3.5. `https://platform.openai.com/docs/models/gpt-3-5`, 2021. 1, 5, 7

[41] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, 2022. 1

[42] OpenAI. GPT-4V(ision) System Card. `https://openai.com/index/gpt-4v-system-card/`, 2023. 3, 6

[43] OpenAI. New embedding models and api updates. `https://openai.com/index/new-embedding-models-and-api-updates/`, 2024. 4

[44] OpenAI. GPT-4o Mini: Advancing Cost-Efficient Intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024. 3, 6

[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.*, pages 27730–27744, 2022. 1

[46] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, et al. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*, 2023. 2

[47] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. Understanding long videos in one multimodal language model pass. *arXiv preprint arXiv:2403.16998*, 2024. 5, 6

[48] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1

[49] 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, 3, 5, 6

[50] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 5

[51] 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

[52] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *Adv. Neural Inform. Process. Syst.*, 35:38032–38045, 2022. 3

[53] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Int. Conf. Comput. Vis.*, pages 11888–11898, 2023. 1, 3

[54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[55] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3, 6

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[57] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*, 2023. 3

[58] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6387–6397, 2023. 3

[59] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 8

[60] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023. 5, 6

[61] 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. 1, 2, 3, 5, 6, 7

[62] Yang Wang, Gedas Bertasius, Tae-Hyun Oh, Abhinav Gupta, Minh Hoai, and Lorenzo Torresani. Supervoxel attention graphs for long-range video modeling. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 155–166, 2021. 3

[63] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 6

[64] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Adv. Neural Inform. Process. Syst.*, pages 8483–8497, 2022. 3

[65] 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. 3

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Process. Syst.*, pages 24824–24837, 2022. 5

[67] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 13587–13597, 2022. 3

[68] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1

[69] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Int. Conf. Comput. Vis.*, pages 1686–1697, 2021. 1

[70] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Adv. Neural Inform. Process. Syst.*, 35:124–141, 2022. 1

[71] Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. Relational space-time query in long-form videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6398–6408, 2023. 3

[72] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024. 3

[73] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3

[74] 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. 1, 2, 5, 6, 7

[75] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 5, 6

[76] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 6

[77] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. 1

[78] 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. 1, 2, 4, 5, 6, 7