VideoAgent: Long-form Video Understanding with Large Language Model as Agent

Xiaohan Wang* Yuhui Zhang* Orr Zohar Serena Yeung-Levy
Stanford University
{xhanwang,yuhuiz,orrzohar,syyeung}@stanford.edu

Abstract

Long-form video understanding represents a significant challenge within computer vision, demanding a model capable of reasoning over long multi-modal sequences. Motivated by the human cognitive process for long-form video understanding, we emphasize interactive reasoning and planning over the ability to process lengthy visual inputs. We introduce a novel agent-based system, VideoAgent*, that employs a large language model as a central agent to iteratively identify and compile crucial information to answer a question, with vision-language foundation models serving as tools to translate and retrieve visual information. Evaluated on the challenging EgoSchema and NExT-QA benchmarks,* VideoAgent *achieves 54.1% and 71.3% zero-shot accuracy with only 8.4 and 8.2 frames used on average. These results demonstrate superior effectiveness and efficiency of our method over the current state-of-the-art methods, highlighting the potential of agent-based approaches in advancing long-form video understanding.*

1. Introduction

Understanding long-form videos, ranging from minutes to hours, poses a significant challenge in the field of computer vision. This task demands a model capable of processing multi-modal information, handling exceedingly long sequences, and reasoning over these sequences effectively.

Despite numerous attempts [9, 16, 21] to address this challenge by enhancing these capabilities, existing models struggle to excel in all three areas simultaneously. Current large language models (LLMs) excel in reasoning and handling long contexts [18, 20, 24], yet they lack the capability to process visual information. Conversely, visual language models (VLMs) struggle to model lengthy visual inputs [5, 6]. Early efforts have been made to enable VLMs' long context modeling capability, but these adaptations underperform in video understanding benchmarks and are inefficient in dealing with long-form video content [7].

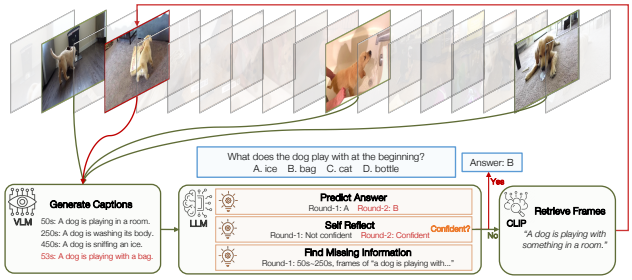

Figure 1. *A typical video agent framework example.* Given a video, the agent iteratively searches and aggregates key information to answer the question. The process is controlled by a large language model (LLM) as the agent, with the visual language model (VLM) and contrastive language-image model (CLIP) serving as tools.

Do we really need to feed the entire long-form video directly into the model? This diverges significantly from how humans achieve the long-form video understanding task. When tasked with understanding a long video, humans typically rely on the following interactive process to formulate an answer: The process begins with a quick overview of the video to understand its context. Subsequently, guided by the specific question at hand, humans iteratively select new frames to gather relevant information. Upon acquiring sufficient information to answer the question, the iterative process is concluded, and the answer is provided. Throughout this process, the reasoning capability to control this iterative process is more critical than the capacity to directly process lengthy visual inputs.

Drawing inspiration from how humans understand long-form videos, we present VideoAgent, a system that simulates this process through an agent-based system. We formulate the video understanding process as a sequence of states, actions, and observations, with an LLM serving as the agent controlling this process (Figure 1). Initially, the LLM familiarizes itself with the video context by glancing at a set of uniformly sampled frames from the video. During each iteration, the LLM assesses whether the current information (state) is sufficient to answer the question; if not, it identifies what additional information is required (action).

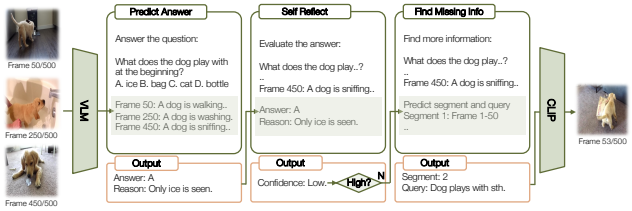

Figure 2. *Detailed view of VideoAgent's iterative process.* Each round starts with the state, which includes previously viewed video frames. The large language model then determines subsequent actions by answering prediction and self-reflection. If additional information is needed, new observations are acquired in the form of video frames.

Subsequently, it utilizes CLIP [13] to retrieve new frames containing this information (observation) and VLM to caption these new frames into textual descriptions, updating the current state. This design emphasizes the reasoning capability and iterative processes over the direct processing of long visual inputs, where the VLM and CLIP serve as instrumental tools to enable the LLM to have visual understanding and long-context retrieval capabilities.

Our work differs from previous works in two aspects. Compared to the works that uniformly sample frames or select frames in a single iteration [4, 23, 26], we select frames in a multi-round fashion, which ensures the information gathered to be more accurate based on the current need. Compared to the works that retrieve frames using the original question as the query [23, 26], we rewrite the query to enable more accurate and fine-grained frame retrieval.

Our rigorous evaluation of two well-established long-form video understanding benchmarks, EgoSchema [10] and NExT-QA [22], demonstrates *VideoAgent*'s exceptional effectiveness and efficiency compared to existing methods. *VideoAgent* achieves 54.1% and 71.3% accuracy on these two benchmarks, respectively, outperforming concurrent state-of-the-art method LLoVi [27] by 3.8% and 3.6%. Notably, *VideoAgent* only utilizes 8.4 frames on average to achieve such performance, which is 20x fewer compared to LLoVi. Our ablation studies highlight the significance of the iterative frame selection process, which adaptively searches and aggregates relevant information based on the complexity of the videos. Additionally, our case studies demonstrate that *VideoAgent* generalizes to arbitrarily long videos, including those extending to an hour or more.

In summary, *VideoAgent* represents a significant stride for long-form video understanding, which embraces the agent-based system to mimic human cognitive process and underscores the primacy of reasoning over long-context visual information modeling. We hope our work not only sets a new benchmark in long-form video understanding but also sheds light on future research in this direction.

2. Method

In this section, we introduce the method of *VideoAgent*, which is inspired by the human cognitive process of understanding long-form videos. Given a video with the question, a human will first glance at several frames to understand its context, then iteratively search additional frames to obtain enough information to answer the question, and finally aggregate all the information and make the prediction.

We formulate the process into a sequence of states, actions, and observations $\{(s_t, a_t, o_t)|1 \leq t \leq T\}$, where the state is the existing information of all the seen frames, action is whether to answer the question or continue to search new frames, observation is the new frames seen in the current iteration, and T is the maximum number of iterations.

We leverage large language model (LLM) GPT-4 [12] as an agent to perform the above process (Figure 1). LLMs have been demonstrated to have memory, reasoning and planning, and tool-use capability [14], which can be used to model states, actions, and observations.

2.1. Obtaining the Initial State

To start the iterative process, we first familiarize the LLM with the context of the video, which can be achieved by glancing at N frames uniformly sampled from the video. Since the LLM has no capability for visual understanding, we leverage vision-language models (VLMs) to convert the visual content to language descriptions. Specifically, we caption these N frames with the prompt "describe the image in detail" and feed the captions to the LLM. This initial state s_1 records a sketch of the content of the video.

2.2. Determining the Next Action

Given the current state s_t that stores the information of seen frames, there are two possible options for the next action a_t:

- **Action 1: answer the question.** If the information in state s_t is enough to answer the question, we should answer the questions and exit the iterative process.
- **Action 2: search new information.** If the current information in s_t is insufficient, we should decide what further information is required to answer the question and continue searching for it.

To decide between actions 1 and 2, we need the LLM to reason over the question and existing information. This is achieved by a three-step process. First, we force the LLM to make a prediction based on the current state and question via chain-of-thought prompting. Second, we ask the LLM to self-reflect and generate a confidence score based on the state, question, prediction and its reasoning process generated by step 1. The confidence score has three levels: 1 (insufficient information), 2 (partial information), and 3 (sufficient information). Finally, we choose action 1 or 2 based on the confidence score. This process is illustrated in Fig-

ure 2. We propose to use a three-step process over a single-step process that directly chooses action as direct prediction always decides to search for new information (Action 2). This self-reflection process is motivated by [15].

2.3. Gathering a New Observation

Suppose the LLM determines insufficient information to answer the question and chooses to search for new information. In that case, we further ask the LLM to decide what extra information is needed so that we can leverage tools to retrieve them (Figure 2). Since some piece of information could occur multiple times within a video, we perform segment-level retrieval instead of video-level retrieval to enhance the temporal reasoning capability. For example, suppose the question is "What is the toy left on the sofa after the boy leaves the room?" and that we have seen the boy leave the room at frame i. If we retrieve with the query "a frame showing the toy on the sofa," there may be frames before frame i containing "toy on the sofa", but they are irrelevant to answering the question.

To perform segment-level retrieval, we first split the video into different segments based on the seen frame indices, and ask the LLM to predict what segments to retrieve with the query texts. For example, if we have seen frames i, j, and k of a video, one valid prediction is segment 2 (frame i to j) with the query "a frame showing the toy on the sofa".

We leverage CLIP [13] to obtain this additional information given LLM output. Specifically, given each query and segment, we return the image frame with the highest cosine similarity with the query in that segment. These retrieved frames are served as observations to update the state.

2.4. Updating the Current State

Finally, given the new observations (i.e., retrieved frames), we leverage VLMs to generate captions for each frame, and then simply sort and concatenate the new captions with old frame captions based on frame index, and ask the LLM to generate next-round predictions.

A question one may posit is why we leverage the multi-round process, as some existing works use all or uniformly sampled frames as the state in a single step [4, 27]. There are many advantages of our approach over these baselines. First, using too many frames introduces extensive information and noise, which leads to performance degradation since LLMs suffer from long contexts [8]. Furthermore, it is computationally inefficient and hard to scale up to hour-long videos due to the LLM context length limit [12]. On the opposite, using too few frames may not capture relevant information. Our adaptive selection strategy finds the most relevant information and requires the lowest cost to answer questions at different difficulty levels.

Methods		Val			
		Acc@C	Acc@T	Acc@D	Acc@All
Supervised					
ATP [1]	[CVPR2022]	53.1	50.2	66.8	54.3
SeViT [3]	[arXiv2023.1]	54.0	54.1	71.3	56.7
HiTeA [25]	[ICCV2023]	62.4	58.3	75.6	63.1
Zero-shot w/o Agent					
SeViLA [26]	[NeurIPS2023]	61.3	61.5	75.6	63.6
LLoVi [27]	[arXiv2024.2]	69.5	61.0	75.6	67.7
Zero-shot w/ Agent					
ViperGPT [17]	[ICCV2023]	-	-	-	60.0
AssistGPT [2]	[arXiv2023.6]	60.0	51.4	67.3	58.4
MoReVQA [11]	[CVPR 2024]	-	-	-	69.2
VideoAgent [19]	[arXiv2024.4]	**72.7**	**64.5**	**81.1**	**71.3**

Table 1. *Comparison of the agent-based methods with others on NExT-QA.* C, T, and D are causal, temporal, and descriptive subsets, respectively.

3. Experiments

In this section, we introduce the datasets, results, and case studies of *VideoAgent*.

3.1. Datasets and Metrics

In our experiments, we use two well-established datasets to benchmark our model's performance, namely EgoSchema [10] and NExT-QA [22], with a particular focus on zero-shot understanding capabilities. Since each dataset features multiple-choice questions, we utilized accuracy as our evaluation metric.

VideoAgent sets new benchmarks, achieving state-of-the-art (SOTA) results on the EgoSchema and NExT-QA datasets, surpassing previous methods significantly while requiring only a minimal number of frames for analysis.

NExT-QA. In Table 1, we show that *VideoAgent* achieves a 71.3% accuracy on the NExT-QA full validation set, surpassing the former SOTA, LLoVi [27], by 3.6%. With an average of merely 8.2 frames used per video for zero-shot evaluation, *VideoAgent* consistently outperforms previous supervised and zero-shot methods across all subsets by a large margin, including those testing the model's causal, temporal, and descriptive abilities. Importantly, *VideoAgent* achieves remarkable performance improvements on the more challenging subsets, ATP-hard [1], demonstrating its adeptness at addressing complex long-form videos. These results underscore *VideoAgent*'s exceptional effectiveness and efficiency in processing and understanding complex questions from long-form videos. The results on EgoSchema are in appendix.

4. Conclusion

In this work, we introduce *VideoAgent*, a system that employs a large language model as an agent to mirror the human cognitive process for understanding long-form videos. *VideoAgent* effectively searches and aggregates information through a multi-round iterative process. It demonstrates exceptional effectiveness and efficiency in long-form video understanding, as evidenced by both quantitative and qualitative studies on various datasets.

References

[1] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022.

[2] 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] Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. Semi-parametric video-grounded text generation. *arXiv preprint arXiv:2301.11507*, 2023.

[4] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021.

[5] 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] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[7] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

[8] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024.

[9] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023.

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

[11] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] OpenAI. Gpt-4 technical report, 2023.

[13] 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.

[14] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *Advances in neural information processing systems*, 35:38032–38045, 2022.

[17] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[19] 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.

[20] 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. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.

[22] 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.

[23] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Retrieval-based video language model for efficient long video question answering, 2023.

[24] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416, 2023.

[26] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, 2023.

[27] 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.