Scene-Clipping Long Video For Better Understanding

Ziyu Zhao
2024213688
School of Software

Jin Wang
2024213687
School of Software

Jinsong Xiao
2024210747
Department of Electrical Engineering

1 Introduction

Recent advances in large-scale video-language models, such as GPT-4o and Gemini-1.5-Pro, have showcased their remarkable ability to understand long video content, due to their support for long context length. These models exhibit impressive potential for deep comprehension of video content, particularly in tasks that require real-time analysis by processing ongoing sequences and retrieving information from long-term memory. However, training such foundational models at this scale remains out of reach for most academic researchers because of the immense computational resources needed to handle the high-dimensional complexity of long-video data. Many current open-source large multimodal models concatenate the query embeddings of each frame along the time axis and input them into the LLM. Although this approach has shown promising results, particularly with short videos, it faces significant challenges when applied to long videos. Consequently, this design becomes impractical for longer videos, as the inherent context length limitations of LLMs and the high GPU memory consumption severely restrict the number of frames that can be processed. For instance, LLaMA has a context length limitation of 2048 tokens, while large multimodal models like LLaVA(1) and BLIP-2(2) can only process 256 and 32 tokens per image respectively.

To address these challenges, there has been a growing interest in developing efficient Video-LLMs that can efficiently process long video sequences despite restricted context length. VideoChat(3), Video-LLaVA(4) and Video-Llama(5) convert a fixed number of sampled frames into a small number of embeddings, regardless of the video's duration, resulting in inadequate information for effectively representing long videos. Both MA-LMM(6) and MovieChat(7) utilize memory-augmented mechanisms to extend the context window for processing long-form video content, allowing them to retain and reference information over extended time periods. However, this memory-averaging approach can lead to a gradual reduction in the richness of the retained information, as it tends to compress and dilute details over time. This can result in an uneven representation, where earlier frames or key moments lose significance, making it challenging to maintain a balanced and detailed understanding of the entire video. TimeChat(8) and LVCHAT(9) group the original video frames and then apply specific aggregation techniques to reduce the number of tokens, achieving more efficient compression. However, the group size must be predetermined and remains fixed, limiting the model's ability to adapt to the unique characteristics of each video. Additionally, the video content within the same group may vary significantly, which can lead to a substantial loss in representational quality after aggregation, hindering the model's ability to accurately capture critical details. Chat-UniVi(10) and VideoLLaMB(11) reduce information loss during aggregation by segmenting the video into distinct segments based on scene changes, helping to preserve semantic coherence. However, these methods still require predefined segmentation ratios or a fixed number of segments, limiting their flexibility in adapting to different video content. Moreover, Chat-UniVi's DPC-KNN-based scene segmentation algorithm can disrupt the original temporal sequence, potentially affecting the natural flow of events within the video.

In light of these challenges, we propose our scene-clipping long video LLM, a novel approach that aggregates spatiotemporal context across extended temporal horizons. To address the limitations of the aforementioned scene segmentation algorithms, we propose a dynamic scene-clipping algorithm that partitions the original video into clips based on the specific scene distribution, eliminating the need to pre-specify the number of clips. This approach ensures semantic consistency within each

Preprint. Under review.

clip. Subsequently, we utilize video-Qformer to extract features from each clip while incorporating temporal encoding information, enabling the LLM to comprehensively understand the spatio-temporal content of the long video.

2 Related Work

2.1 Video-LLMs

Recent Video-LLMs have made strides in improving the understanding of temporal dynamics in video content. For instance, Video-Llama(5) enhances the BLIP-2 architecture by introducing an additional video-querying transformer to explicitly model temporal relationships. Similarly, Video-ChatGPT(12), built on LLaVA, employs a simple average pooling of frame-level features across spatial and temporal dimensions to generate a unified video-level representation. Meanwhile, VideoChat(3) employs perception models to generate action and object annotations, which are then processed by LLMs for higher-level reasoning. Building on these advances, VideoChat2(13) introduced a multi-stage bootstrapping technique focused on modality alignment and instruction tuning, allowing the collection of high-quality video data for fine-tuning instruction-driven tasks. Video-LLaVA(4) enhances modality integration by using a pre-aligned encoder adaptable to both images and videos which enables shared projections and synergistic training across image and video tasks. Although these models represent significant advances, they are predominantly designed for short videos. Longer videos present considerable challenges due to the inherent limitations of LLM context length and the high memory demands on GPUs. These factors restrict the ability of current models to scale effectively for long-term video understanding.

2.2 Long-term Video-LLMs

Long-term Video-LLMs aim to capture extended patterns in videos that typically exceed 30 seconds in duration. Long videos pose challenges due to high computational complexity and memory demands, prompting long-term video LLMs to adopt advanced temporal modeling techniques for improved efficiency. MovieChat(7) introduced a novel memory-based mechanism that strategically merges similar frames to reduce computational load and memory usage. Chat-UniVi(10) proposed a unified approach to processing images and videos by dynamically merging similar spatial and temporal tokens to improve efficiency. LLaMA-VID(14) condensed video representations by representing each frame with only two tokens, separating context and content tokens for more efficient compression. For long video QA, Xu *et al.*(15) explore selectively using frames or clips from long videos using retrieval-based methods. This approach aims to focus on the most relevant video segments, improving efficiency and effectiveness in answering questions based on extended video content. TimeChat and LVCHAT group the original video frames and apply specific aggregation techniques to reduce the number of tokens, thus achieving more efficient compression.

3 Method

We propose a fine-tuning approach that leverages a frozen video LLM integrated with a Video-Qformer, pre-trained on short video clips, to adapt it for long video content. Given a video V with n frames, we first extract frames to obtain a complete sequence of frame representations $F = \{f_1, f_2, ..., f_n\}$ using the pre-trained image encoder. Next, we apply our entropy-based scene-clipping algorithm to frame embeddings F to generate k clips. The frame embeddings within each clip are then fed into the Video-Qformer to obtain clip embeddings $C = \{c_1, c_2, ..., c_k\}$, incorporating temporal position information, which are then concatenated to produce the final representation. This approach enables the LLM to comprehensively understand the spatiotemporal content of long videos.
Datasets Long videos data from VideoChat2(13) and ShareGPT4Video(16). Long video dataset ActivityNet Captions(17).
Baselines Video-Llama(5), VideoLLaMB(11), Chat-Univi(10), MovieChat(7) and Timechat(8).
Benchmarks MVBench(13) for short video understanding. EgoSchema(18) for long video QA. LongVideoBench(19) and MLVU(20) for multi-task long video understanding.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[3] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[4] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[5] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[6] B. He, H. Li, Y. K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, and S.-N. Lim, "Ma-lmm: Memory-augmented large multimodal model for long-term video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 504–13 514.

[7] Z. Song, C. Wang, J. Sheng, C. Zhang, G. Yu, J. Fan, and T. Chen, "Moviellm: Enhancing long video understanding with ai-generated movies," *arXiv preprint arXiv:2403.01422*, 2024.

[8] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "Timechat: A time-sensitive multimodal large language model for long video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 313–14 323.

[9] Y. Wang, Z. Zhang, J. McAuley, and Z. He, "Lvchat: Facilitating long video comprehension," *arXiv preprint arXiv:2402.12079*, 2024.

[10] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 700–13 710.

[11] Y. Wang, C. Xie, Y. Liu, and Z. Zheng, "Videollamb: Long-context video understanding with recurrent memory bridges," *arXiv preprint arXiv:2409.01071*, 2024.

[12] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[13] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo *et al.*, "Mvbench: A comprehensive multi-modal video understanding benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 195–22 206.

[14] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *European Conference on Computer Vision*. Springer, 2025, pp. 323–340.

[15] J. Xu, C. Lan, W. Xie, X. Chen, and Y. Lu, "Retrieval-based video language model for efficient long video question answering," *arXiv preprint arXiv:2312.04931*, 2023.

[16] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *arXiv preprint arXiv:2406.04325*, 2024.

[17] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[18] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 212–46 244, 2023.

[19] H. Wu, D. Li, B. Chen, and J. Li, "Longvideobench: A benchmark for long-context interleaved video-language understanding," *arXiv preprint arXiv:2407.15754*, 2024.

[20] J. Zhou, Y. Shu, B. Zhao, B. Wu, S. Xiao, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu, "Mlvu: A comprehensive benchmark for multi-task long video understanding," *arXiv preprint arXiv:2406.04264*, 2024.