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

Abstract

In recent years, the demand for effective long video understanding has surged, driven by the increasing volume of video content across various platforms. However, existing models primarily designed for short video clips struggle to capture the complex spatiotemporal dynamics inherent in longer videos. To address this challenge, we propose a novel scene-clipping long video LLM that dynamically segments videos based on scene distribution without pre-specifying the number of clips, ensuring semantic consistency. Our method segments videos into clips, extracts frame representations using a pre-trained image encoder, and employs an entropy-based scene-clipping algorithm to generate clip embeddings through the Video-Qformer while incorporating temporal position information. Our approach enables the LLM to comprehensively understand the spatiotemporal content of long videos, paving the way for enhanced applications in video summarization, question answering, and interactive video analysis. We train our proposed approach on long video QA and caption datasets and demonstrate its effectiveness on zero-shot long video understanding benchmarks, where it out performs state-of-the-art video-LLMs in absolute accuracy across most tasks.

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

Preprint. Under review.

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 clip. Subsequently, we utilize Clip Q-former 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. Finally, we use Video Q-former to merge the content of each clip to enhance the in-depth understanding of the entire long video. We develop our method by fine-tuning the Video-LLaMA model, originally pre-trained on short videos, using long video data. Experimental results demonstrate that our fine-tuning approach enhances the original model's performance on long video understanding tasks and outperforms several SOTA methods. In summary, our contributions are as follows:

- We propose a dynamic scene-clipping algorithm that segments the video based on its inherent scene distribution, eliminating the need to predefine the number of clips. This approach ensures semantic consistency within each clip and enhances adaptability to diverse video content.
- We introduce a multi-level feature extraction strategy: the Clip Q-former extracts spatiotemporal features from individual clips while incorporating temporal encoding to model local relationships, and the Video Q-former aggregates the clip-level features to achieve a comprehensive understanding of the long video's global context.
- Building on the Video-LLaMA model pre-trained on short videos, we fine-tune the model on long video data, enhancing its performance for long video understanding tasks. Experimental results demonstrate that our approach surpasses several SOTA methods, validating its effectiveness and superiority.

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

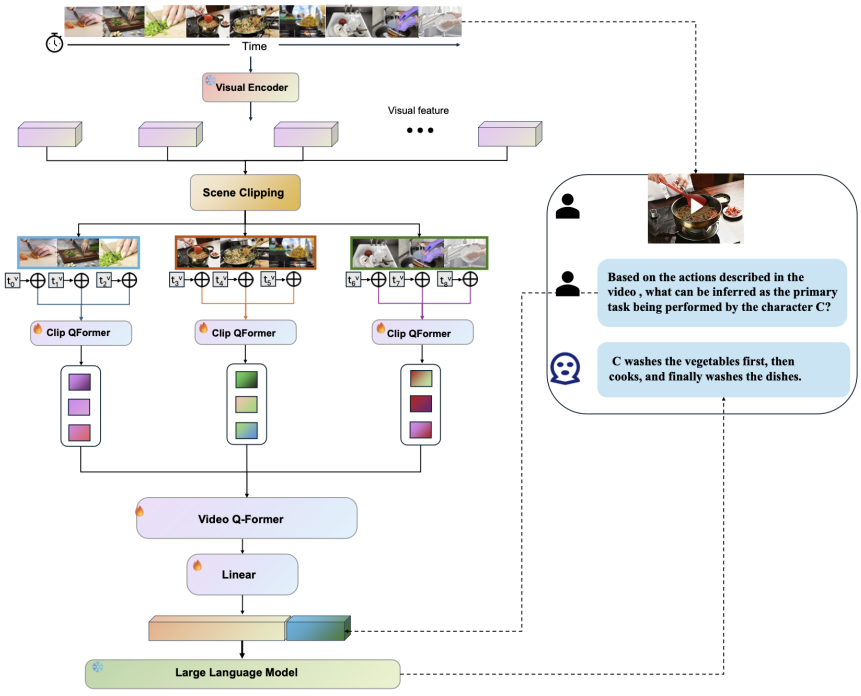

Figure 1: Overview of our full approach.

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 Clip Q-former to obtain clip embeddings $C = \{c_1, c_2, ..., c_k\}$, which are finally fed into Video Q-former and a linear layer to produce the video representation. This approach enables the LLM to comprehensively understand the spatiotemporal content of long videos. In this

section, we first introduce Scene-Clipping algorithm in detail, and then describe how our structure is better suited for long video understanding.

3.1 Scene-Clipping

Scene segmentation along video temporal sequences has long been recognized as a crucial task, as it preserves the non-linear structure of context and significantly contributes to compressing extensive contextual information(16; 17). Inspired by information entropy, Scene-Clipping divides the entire video sequence into semantically distinct segments, ensuring coherence between segments by considering the overall internal similarity, rather than focusing solely on adjacent changes.
Given a sequence of n frame features $\{F_1, F_2, ..., F_n\}$, the Scene-Clipping algorithm is as follows.

1. Compute the cosine similarity between any two frame feature pairs, then take the negative logarithm of this value. The larger the value, the greater the difference in content between the two frames. We have similarity entropy matrix $SE = (se_{ij})n \times n, se_{ij} = -\log\left(\frac{\mathbf{F_i} \cdot \mathbf{F_j}}{\|\mathbf{F_i}\|\|\mathbf{F_j}\|}\right)$.

2. Define the overall entropy of the video to be $Sum(SE)$, $Sum(SE_i^j)$ (SE_i^j represents $SE[i:j, i:j]$) is the entropy of a clip of the video which includes frames i to j. The goal of the optimization is to divide the original video into several clips connected end to end, and make the total entropy of these clips as small as possible.

3. We use beam search to search for the split point and stop when the total entropy is less than the threshold.

3.2 Architecture

We adjusted the structure of the Video-LLaMA(5) pre-trained on the short video dataset to adapt it for long video content. It is composed of a frozen pre-trained image encoder to extract features from video frames, a position embedding layer to inject temporal information into video frames, a clip Q-former to aggregate frame-level representations, a video Q-former to aggregate clip-level representations and a linear layer to project the output video representations into the same dimension as the text embeddings of LLMs. Given that a video consists of N frames, the image encoder will first map each frame into K_f image embedding vectors, generating video frame representations $V = \{v_1, v_2, ..., v_N\}$ where v_i is the set of d_f-dimensional image embeddings corresponding to the i-th frame.
We first use the Scene-Clipping algorithm on V to get m clips frame representations. Follow Video-LLaMA, since the frame representations v_i from the frozen image encoder are computed without considering any temporal information, we further apply position embeddings as the indicator of temporal information to the representations from different frames. Then, we feed the position-encoded frame representations to Clip Q-former, which shares the same architecture with Video Q-former in Video-LLaMA, to obtain k_C clip embedding vectors of dimension d_c as the representation $\hat{c}$ of the clip. In this step, since the semantics within each clip are highly consistent, Clip Q-former can fully fuse the representation information. Finally, we feed clip representations to Video Q-former, which inherits from the Video Q-former in Video-LLaMA, to extract all embedding vectors into the entire video embedding vectors.
To adapt the video representations to the input of LLMs, the linear layer is to transform the video embedding vectors into the video query vectors. The video query vectors are of the same dimension as the text embeddings of LLMs. In the forward pass, they will be concatenated to text embeddings as a video soft prompt and guide the frozen LLMs to generate text conditioned on video content. Follow Video-LLaMA, we utilize the pre-trained vision component of BLIP-2(2) as the frozen visual encoder, which includes a ViT-G/14 from EVA-CLIP(18), a pre-trained Q-former. The remaining components, including the position embedding layer, Clip Q-former, Video Q-former, and Linear layer are initialized from the pre-trained Video-LLaMA and optimized to well connect the output of the frozen visual encoder to frozen LLMs.

4 Experiments

Datasets. We train our approach on video longer than 30 seconds from VideoChat2(13), ShareGPT4Video(19) and ActivityNetQA(20). We evaluate our approach on zero-shot long video benchmark MLVU(21) for multi-task long video understanding.
Implementation details. We build our approach off the publicly available Video-LLama model and train for 3 epochs on the above datasets with 64 frames per video.

4.1 Results on MLVU

The MLVU (Multi-task Long Video Understanding Benchmark) is a comprehensive dataset designed to evaluate long video understanding (LVU) performance, addressing challenges like insufficient video lengths, limited diversity in video types, and a lack of varied evaluation tasks by including diverse genres (e.g., movies, surveillance, egocentric videos, cartoons, and games) and multiple evaluation tasks to benchmark the key capabilities of multimodal large language models (MLLMs). We report the results of our zero-shot evaluation on the MLVU benchmark in Table 1. Notice that since we don't have the OpenAI API, we didn't test generation tasks like Video Summary and Sub-Scene Captioning. Overall, our method outperforms current long video language models trained on similar data, demonstrating robust performance compared to other approaches and confirming its efficacy. Our method has significant improvements over Video-LLaMA, which shows that fine-tuning on long videos with our architecture can enhance the pre-trained video LLM's ability to understand long videos.

Methods	**Holistic**		**Single Detail**			**Multi Detail**		**M-Avg**
	TR	AR	NQA	ER	PQA	AO	AC	
MovieChat(7)	29.5	25.0	24.2	24.7	25.8	28.6	22.8	25.8
TimeChat(8)	23.1	27.0	24.5	28.4	25.8	24.7	32.0	30.9
Chat-Univi(10)	33.8	34.5	30.1	34.7	36.5	22.9	27.8	35.2
Video-LLama(5)	31.9	35.5	42.1	38.9	45.8	25.1	24.3	33.4
Our method	38.5	39.5	44.0	42.6	44.5	27.2	25.8	39.5

Table 1: Results on MLVU benchmark. (TR: Topic Reasoning, AR: Anomaly Recognition), the single-detail LVU tasks (NQA: Needle QA, ER: Ego Reasoning, PQA: Plot QA), and multi-detail LVU tasks (AO: Action Order, AC: Action Count). M-Avg: the average performance of multiple-choice tasks

4.2 Ablation Study

Method	M-Avg	Δ
frames avg pooling in clip	38.8	-0.7
clips avg pooling	35.2	-4.3
clips concat	38.3	-1.2
uniform clipping	31.2	-8.3
current method	39.5	

Table 2: Ablated results

In this section, we present an ablation study of our method, focusing on method for processing frame embedding after scene-clipping. First, we evaluate the effectiveness of the Clip Q-former. To this end, we replace it with a mean pooling strategy. Then we evaluate the effectiveness of Video Qformer by directly concat or average pooling the clip embeddings. In addition, we also adopted a uniformed clipping method to verify the effectiveness of scene-clipping. We analysis our method on MLVU M-Avg metric. The corresponding results are detailed in Table 2. Compared to a uniform clipping approach, our scene-clipping method is more adept at dividing videos into semantic segments. This segmentation results in a more efficient preservation of information, mitigating the information loss typically associated with sampling strategies. Video Q-former excels at capturing and representing the content across multiple frames in a video, enabling a more comprehensive understanding of temporal and spatial information. However, applying pooling techniques can significantly reduce memory consumption and computational time by aggregating frame-level features.

5 Conclusion

In this work, we address the challenges of long video understanding by proposing a novel framework that effectively processes extended temporal sequences while maintaining semantic coherence and spatiotemporal consistency. By introducing a dynamic scene-clipping algorithm and a hierarchical feature extraction strategy using Clip Q-former and Video Q-former, our method effectively models both local and global video content. Fine-tuning the Video-LLaMA model on long video data, we achieve significant performance improvements, surpassing state-of-the-art baselines and providing a robust solution for long video comprehension tasks.

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] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid, "Shot contrastive self-supervised learning for scene boundary detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9796–9805.

[17] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, "A local-to-global approach to multi-modal movie scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 146–10 155.

[18] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 358–19 369.

[19] 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.

[20] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[21] 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.