Leveraging Textual Memory and Key Frame Reasoning for Full Video Understanding Using Off-the-Shelf LLMs and VLMs (Student Abstract)

Harsh Dubey[1], **Chulwoo Pack**[1]

[1]McComish Department of Electrical Engineering and Computer Science, South Dakota State University
Box 2222, Daktronics Engineering Hall 129, Brookings, SD, 57007, USA
harsh.dubey@jacks.sdstate.edu, chulwoo.pack@sdstate.edu

Abstract

To address the limitations of current Large-scale Video-Language Models (LVLMs) in fine-grained understanding, generalization on unseen-data, and long-term temporal memory, we propose a novel video understanding approach that integrates a Vision Language Model (VLM) and a Large Language Model (LLM) with a textual memory mechanism to ensure continuity and contextual coherence. In addition, we introduce a novel evaluation metric, VAD-Score (Video Automated Description Score), to assess precision, recall, and F1 scores for events, subjects, and objects. Our approach delivers competitive results on a diverse set of randomly sampled videos from the DREAM-1K dataset, spanning categories such as live-action, animation, shorts, stock, and YouTube, with a focus on fine-grained comprehension.

Introduction

Recent advancements in LVLMs have enhanced video comprehension, particularly in tasks like video question answering (Lin et al. 2023). While these open-source models perform well on benchmarks in a zero-shot manner, they struggle to generalize to real-world videos with multiple subjects, events, or non-standard perspectives. They also lack fine-grained understanding and long-term temporal memory, leading to incomplete analysis of complex videos.

To address these limitations, we propose a method that leverages a single VLM and a LLM. Our approach tackles the generalization issue by using a VLM instead of LVLMs, as VLMs have shown better performance in generalizing to unseen images. Since video content can be represented by frames, VLMs are well-suited to handle diverse video inputs, effectively capturing fine-grained details like recognizing multiple subjects, objects, predicates, and their attributes. In addition, we incorporate *textual memory* into the VLM prompt to resolve the lack of long-term temporal memory. By feeding the video summary as textual memory into the VLM, our model can maintain coherence across video frames, ensuring a more complete understanding of the temporal relationships throughout the video sequence. Furthermore, we introduce *VAD-Score* to assess the quality and granularity of video descriptions based on events, subjects, and objects. It calculates precision, recall, and F1 scores by comparing these components to ground truth data, then combines the F1 scores using a weighted average, prioritizing events over subjects and objects to generate the final score.

Our work makes the following key contributions: • We propose a method that leverages a single VLM and LLM to address generalization issues in video understanding by focusing on key frames. • We introduce a novel way of reasoning about key frames to capture fine-grained details such as objects, background, foreground, and attributes like color, shape, and size. • We introduce textual memory into the VLM prompt to maintain long-term temporal coherence across video sequences. • We introduce VAD-Score, a new evaluation metric that assesses the quality and granularity of video descriptions by analyzing events, subjects, and objects, and computing precision, recall, and F1 scores. • We demonstrate that our method achieves promising performance over state-of-the-art models on video understanding benchmarks, achieving significant improvements in video captioning and question answering tasks.

Methodology

Key Frame Processing and Attribute Reasoning. We employ LLAVA 1.6 for key frame processing and attribute reasoning, structured as follows:

1. Frame Description (FD): Each keyframe is processed by LLAVA, generating a detailed Frame Description (FD) FD_n for the n-th keyframe. This includes capturing all visual elements within the frame and producing a natural language description.

2. Query/Answer Pairs (QA): LLAVA generates six query/answer pairs (Q/A) for each keyframe, prompting the model to reason about the frame's content. The goal is to mimic the process of "thinking out loud," allowing the model to consider the "why" and "what" of the scene. This reasoning helps extract deeper attribute-specific information, including the objects, background, and scene elements. By generating these Q/A pairs, LLAVA is encouraged to think critically about what is happening in the frame, why it's happening, and what the various attributes are, enriching the frame description with contextual and semantic insights.

Combining FDs and Q/A Pairs. FD and the six Q/A pairs are combined using LLAMA 3.1, producing a Comprehensive Frame Description (CFD):

$$CFD_n = \text{LLM}(FD_n, Q/A_n) \quad (1)$$

where CFD_n represents the final attribute-rich, reasoned description of the n-th frame. This process aims to produce a FD that not only captures basic visual details but also includes a layer of reasoning about the scene's attributes.

Textual Memory and Contextual Continuity. To maintain continuity between frames, we introduce textual memory. The textual memory establishes a memory chain, ensuring that the description of each frame $n+1$ builds upon the previous frame n, thereby transferring information across the entire sequence of frames:

$$CFD_n \rightarrow \text{Textual Memory} \quad (2)$$

By using the CFD as memory, each frame is described as a continuation of the previous scene. This process ensures that the information from the first frame is carried through to the last frame.

Iterative Video Description Generation. The generated CFDs of each keyframe are combined iteratively to generate the final video description (vd_{0-n}). The process starts by combining the first two FDs CFD_0 and CFD_1, and iteratively incorporates each subsequent CFD:

$$vd_{0-1} = \text{LLM}(CFD_0, CFD_1) \quad (3)$$

$$vd_{0-2} = \text{LLM}(vd_{0-1}, CFD_2) \quad (4)$$

$$vd_{0-1-2-3} = \text{LLM}(vd_{0-1-2}, CFD_3) \quad (5)$$

This iterative combination ensures that the final video description vd_{0-n} incorporates information from all keyframes while respecting the memory limitations of LLAMA 3.1. This method guarantees that critical details from each frame are included in the final narrative, even when processing large amounts of data.

VAD-Score (Video Automated Description Score.) To evaluate the quality and granularity of video descriptions, we introduce VAD-Score, which is calculated as:

$$VAD = 0.5 \cdot F1_{\text{events}} + 0.3 \cdot F1_{\text{subjects}} + 0.2 \cdot F1_{\text{objects}} \quad (6)$$

where $F1_{\text{events}}$, $F1_{\text{subjects}}$, and $F1_{\text{objects}}$ are the F1 scores for events, subjects, and objects, respectively. The weights 0.5, 0.3, and 0.2 are set based on the priority of capturing events, subjects, and objects in the video, with events being the highest priority.

Experiment

We evaluated our model on 10 randomly sampled videos from the DREAM-1K dataset (Wang, Yuan, and Zhang 2024) across five categories: Movie Animation, Movie Live Action, Shorts, Stock, and YouTube. Each category had both short and long videos, except YouTube (only short). Long videos were over 17 seconds; short videos were under 17 seconds.

Human-Annotated Ground Truth. Due to the limited annotations in DREAM-1K, we manually created detailed descriptions, capturing events, subjects, and objects. This provided a more comprehensive ground truth.

Evaluation Metrics. We used VAD-Score as our primary metric, evaluating based on events, subjects, and objects.

Results by Video Length. Our model achieved better VAD-Score on short videos (0.692) than long videos (0.595). Short videos, with fewer events and simpler transitions, allowed for more coherence. Subject recognition was strong across both lengths, often achieving perfect F1 scores.

Comparison with Other Models. We compared our model against Video-LLAVA (Lin et al. 2023), MiniGPT-4VT (Zhu et al. 2023), LLAVA-NeXT-Video (Liu et al. 2024), VideoChat2 (Li et al. 2024), PLLaVA-34B (Xu et al. 2024), GPT-4V, and Gemini 1.5 Pro, which focused on event recognition only.

Live Action. Our model performed competitively with an event F1 score of 0.75, close to GPT-4V (0.75) and Gemini 1.5 Pro (0.77).

Movie Animation and YouTube. Our model lagged behind, with F1 scores of 0.29, trailing PLLaVA-34B and LLAVA-NeXT-Video.

Shorts and Stock. Our model achieved respectable F1 scores of 0.65 and 0.60, comparable to Tarsier-7B and MiniGPT-4VT.

Conclusion

Our approach of integrating textual memory into VLM in a zero-shot manner significantly improves the capture of fine-grained details and temporal dynamics, achieving promising improvements in granular quality measures over state-of-the-art video captioning and VQA models. Future work will involve in-depth analysis at scale.

References

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

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

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Wang, J.; Yuan, L.; and Zhang, Y. 2024. Tarsier: Recipes for Training and Evaluating Large Video Description Models. *arXiv preprint arXiv:2407.00634*.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.