HARK: Hierarchical Agentic Retrieval with Keyframing for Video Understanding (Student Abstract)

Jingcheng Li[1], **Ye Qiao**[2], **Sitao Huang**[2]
[1]University of California, San Diego, La Jolla, CA, 92092, USA
[2]University of California, Irvine, Irvine, CA, 92697, USA
[1]jil458@ucsd.edu
[2]{yeq6, sitao}@uci.edu

Abstract

Current video understanding models struggle with temporal reasoning and efficient processing while balancing detail preservation with computational efficiency. We propose a hierarchical memory system that segments videos into action and scene units, combined with question-aware agentic keyframe selection. Our method achieves 70.3% overall accuracy on VideoMME short video benchmarks.

Introduction

Video understanding faces challenges in temporal reasoning, particularly in identifying when specific events occur within video sequences. Existing approaches like MovieChat (Song et al. 2024) use naive similarity-based merging that causes information loss through arbitrary frame consolidation. The challenge is further compounded by the fact that models specialized for temporal grounding tasks often degrade in general question-answering capabilities.

We propose a hierarchical memory architecture that segments videos into semantically meaningful action and scene units while maintaining temporal information through timestamp preservation. Our contributions include: (1) semantic boundary-aware hierarchical segmentation preserving temporal grounding, (2) agentic keyframe selection dynamically focusing on query-relevant content, and (3) empirical analysis showing 70.3% accuracy on VideoMME (Fu et al. 2024).

Proposed Framework

Our approach operates through a hierarchical memory architecture combined with agentic keyframe selection. Inspired by Video ReCap's hierarchical processing (Islam et al. 2024) and ViLAMP's differential distillation principle (Cheng et al. 2025), we develop a unified framework in Fig 1 that operates through four main stages: (1) semantic boundary detection for memory population, (2) hierarchical memory organization, (3) agentic segment retrieval, and (4) question-aware keyframe selection for reasoning.

Semantic Action Boundary Detection

Unlike MovieChat's (Song et al. 2024) naive similarity-based frame merging that suffers from information loss through arbitrary adjacent frame consolidation, our approach detects semantically meaningful action boundaries. We process video frames using a frozen ViT encoder to extract visual features:

$$f_i = \text{ViT}(\text{frame}_i)$$

For action boundary detection, we employ adaptive thresholding rather than fixed similarity thresholds. We maintain a sliding window W of recent similarity scores and compute:

$$\tau_{adaptive} = \max(\mu_W - k \cdot \sigma_W, \tau_{min})$$

where μ and τ are the mean and standard deviation of recent frame-to-frame similarities within window W , k controls sensitivity, and τ_{min} prevents over-segmentation. When cosine similarity between consecutive frames drops below $\tau_{adaptive}$:

$$\text{sim}(f_i, f_{i+1}) = \frac{f_i \cdot f_{i+1}}{\|f_i\| \cdot \|f_{i+1}\|} < \tau_{adaptive}$$

we detect an action boundary. This approach ensures that segmentation aligns with semantic content changes rather than arbitrary similarity thresholds. The accumulated frames are processed through a video language model to generate a descriptive caption, creating an Action Unit that contains the caption, timestamps, and caption embedding.

Hierarchical Scene Organization

We also adapt the hierarchical principle for detecting scene boundaries through semantic similarity between action captions. Scene transitions are identified when:

$$\text{sim}(\text{embedding}_{scene\_start}, \text{embedding}_{current}) < \tau_{scene}$$

This semantic approach ensures scene boundaries align with meaningful content transitions rather than visual similarity alone. When a scene boundary is detected, all actions within that scene are compressed into a Scene Unit that contains the summary of the actions, time range, and the indices of action unit.

Agentic Reasoning with Differential Keyframe Selection

Inspired by ViLAMP's differential distillation principle (Cheng et al. 2025), we apply CLIP-based keyframe selection to frames from semantically relevant segments, making

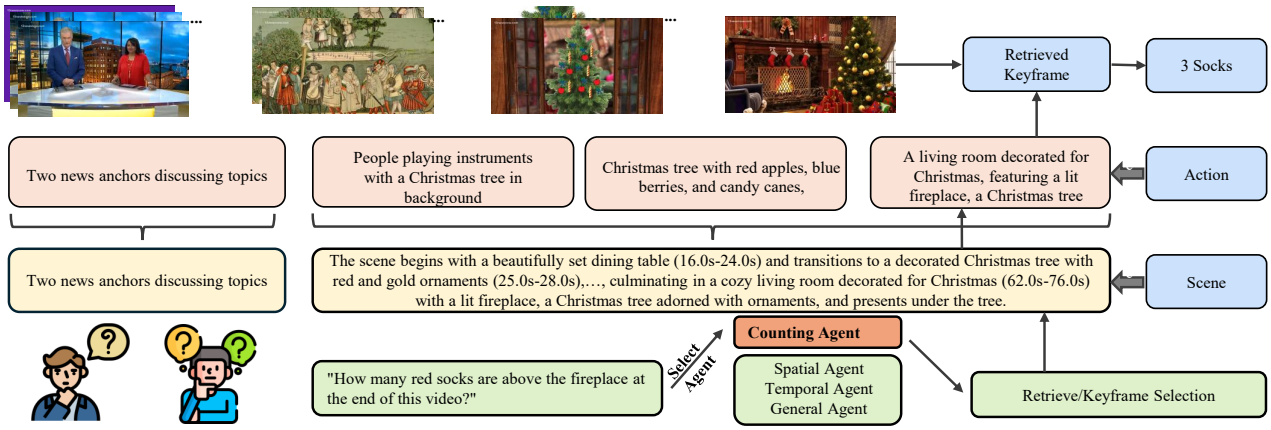

Figure 1: Overview of the Proposed Hierarchical Memory Approach

the selection more focused and efficient. For each frame f_i in the selected segments, we compute a differential saliency score:

$$D(f_i) = R(f_i, Q) - \lambda \cdot T(f_i, \mathcal{N}(f_i))$$

where $R(f_i, Q)$ refers to query relevance; $T(f_i, \mathcal{N}(f_i))$ refers to temporal redundancy; $\mathcal{N}(f_i)$ contains frames within a sliding window of size. The weighting parameter λ is question-adaptive, the higher it is, the higher the redundancy penalty is. We select keyframes using a greedy algorithm that prioritizes high differential scores while preventing redundancy. Our system then employs skill-specialized agents, each optimized for different question types with task-specific confidence thresholds and segment selection strategies. For example, the GeneralAgent is aimed at answering high-level questions like "what is the main theme of the video," and in this case we can use radical keyframe selection because the details are less important; while the CountingAgent should not ignore too many frames as counting requires better video grounding and cannot afford information loss.

Reasoning Process

Our agentic reasoning system operates through a multi-stage pipeline. Given an input question Q, we first classify its task type using pattern matching on question content and structure. Each agent computes a confidence score for handling the question based on both task type matching and keyword analysis, and we select the one with the highest score. The selected agent first identifies relevant scenes, then drills down to specific actions within those scenes. From the retrieved segments, we apply our differential keyframe selection algorithm to identify the most informative frames while avoiding redundancy. The process operates on frames extracted from relevant memory units. For each frame, we compute the differential saliency score. The algorithm greedily selects frames with high differential scores, and the selected keyframes are processed through the specialized agent using task-specific prompts. Each agent employs distinct instructions optimized for its domain. This hierarchical reasoning process ensures that computational resources are allocated efficiently based on question while maintaining high accuracy through specialized agent expertise and targeted keyframe selection.

Method	**Accuracy (%)**
LLaVA-Video-7B-Qwen2 (baseline)	63.3
Ours (Hierarchical + Agentic)	**70.3**

Table 1: Performance on VideoMME short videos.

Experiment and Results Analysis

We evaluate our approach on VideoMME((Fu et al. 2024)), a benchmark that spans multiple task categories. Our system employs LLaVA-Video-7B-Qwen2((Zhang et al. 2024)) as the primary video-language model for caption generation. Videos are processed at 1 FPS to balance temporal coverage with computational efficiency. We employ selective 8-bit quantization for LLaVA and Llama models while maintaining ViT in higher precision (bfloat16) for feature stability. All experiments are conducted on a system with 4 NVIDIA GeForce RTX 3090 GPUs. No task-specific fine-tuning is performed—all results represent zero-shot performance to demonstrate the generalizability of our approach. Table 1 presents our main results on VideoMME short videos. Our hierarchical memory approach with agentic keyframe selection achieves 70.3% overall accuracy, demonstrating competitive performance across diverse video understanding tasks.

Conclusion

We present a hierarchical memory approach for video understanding that addresses key limitations in existing methods through semantic boundary detection and agentic keyframe selection. The zero-shot evaluation without fine-tuning demonstrates the generalizability of our approach across diverse video understanding tasks.

References

Cheng, C.; Guan, J.; Wu, W.; and Yan, R. 2025. Scaling Video-Language Models to 10K Frames via Hierarchical Differential Distillation. *arXiv preprint arXiv:2504.02438*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Islam, M. M.; Ho, N.; Yang, X.; Nagarajan, T.; Torresani, L.; and Bertasius, G. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18198–18208.

Song, E.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.