

# HERMES: temporal-coHERent long-forM understanding with Episodes and Semantics

## Supplementary Material

### I. Supplementary Material

This Supplementary document is organized as follows:

- [A.1 Reproducibility Statement](#)
- [A.2 Implementation Details](#)
- [A.3 Model Details](#)
- [A.4 Extended Ablations](#)
- [A.5 HERMES vs. MA-LMM vs. MovieChat](#)
- [A.6 A Note on Latency](#)
- [A.7 More Qualitative Results](#)
- [A.8 Error Analysis: When does HERMES fail and why?](#)
- [A.9 How is our approach related to cognitive processes?](#)

#### I.1. Reproducibility Statement

To facilitate the reproducibility of our work, we will make our code, pretrained models, default hyperparameters, and preprocessed annotations publicly available. Detailed hyperparameters for each dataset are also provided in Table 9. Our model demonstrates efficient performance, completing inference on the MovieChat-1k test set in 13 minutes (22 FPS) using a single V100 GPU (32 GB), and training on the MovieChat-1k dataset in less than 12 minutes with 8x 32 GB GPUs. In contrast to recent LLM-based approaches that necessitate extensive and costly multi-stage pretraining on increasingly large datasets, our model is designed for accessibility, thereby lowering the barrier for researchers without access to high-end computing resources. We achieve high performance while maintaining accessibility by leveraging existing pretrained weights and implementing our training-free ECO and SeTR, resulting in a model where finetuning is optional. We also demonstrate the applicability of our modules to existing video models, and are planning to submit pull requests to integrate our modules into these models.

For fully-supervised results, QFormers and adapter are fine-tuned on the respective dataset’s training split. For plug-in experiments, ECO and SeTR are inserted into target architectures at inference time, with **zero additional training**, demonstrating true plug-and-play capability.

#### I.2. Implementation Details

To ensure the reproducibility of our results, we provide training and inference details in Table 9. These settings are mostly consistent across different datasets. In the table, LR is the learning rate, and Keep Ratio is the SeTR keep ratio. Episodes refer to the number of episodes to which we compress the input frames (i.e., the capacity of ECO). The number of frames (N) represents the quantity of frames retained from the original video to serve as input to the model.

These frames are selected by applying a regular stride over the original video’s frame sequence, where the stride length is determined by the ratio of original frame count to N. *Max Epoch = 20* means we run the program for 20 epochs, performing evaluation after each epoch, and then pick the model with the highest validation accuracy. MovieChat-1k (G) and MovieChat-1k (B) denote global and breakpoint modes, respectively. All models were trained on 8 V100 GPUs (32GB VRAM each). We test on VideoMME using the zero-shot setting by applying our modules to two different models, the same parameters were used across models for consistency.

#### I.3. Model Details

##### I.3.1. Details of our Episodic QFormer

The Episodic Q-Former, as visualized in Figure 7, extends the original QFormer architecture by inserting the Episodic Compressor (ECO) described in Section 4.2. It begins with a set of initial queries that undergo a self-attention process, enhancing internal query representations. These queries then interact with episodic visual features through cross-attention, allowing the incorporation of contextual visual information. The resulting enhanced queries are fed into our ECO module alongside existing query episodes, which represent previously processed queries grouped into episodes. ECO iteratively updates the query episodes, adding the new queries to the existing episodes. This Episodic QFormer allows the model to better handle long sequences or repeated queries by maintaining richer contextual knowledge across iterations.

To mitigate *temporal confusion* during merging, we apply positional encoding (PE) to frame features before ECO. This effectively discourages out-of-order merges by embedding temporal locality directly into similarity calculations. As an ablation, **removing PE reduces MovieChat-1k accuracy from 78.6 to 77.3** on MovieChat-1k, indicating its effectiveness in preserving temporal coherence despite compression. Other studies such as Transformer-XL [8] and Compressive Transformer [28], also report performance drops when positional biases are removed from their compression modules.

**ECO implicitly captures event frequency:** frequent events naturally occur across multiple frames and thus have higher likelihoods of being retained or merged into reinforced prototypes within the memory bank. This self-reinforcing mechanism ensures high-importance (and often high-frequency) events remain well-represented. Explicit event frequency tracking is an idea worth exploring, however, we believe it would be more computationally intensive

Dataset	Max Epochs	LR	Batch	Frames (N)	Episodes	Keep Ratio
MovieChat-1k (G)	1	1e-4	32	100	20	0.2
MovieChat-1k (B)	1	1e-4	32	40	10	0.5
LVU	20	1e-4	32	100	20	0.2
COIN	20	1e-4	32	100	20	0.2
Breakfast	20	1e-4	32	100	20	0.2
VideoMME (LongVA)	-	-	1	128	32	0.125
VideoMME (Llava-OV)	-	-	1	128	32	0.125

Table 9. Hyperparameters used for different datasets.

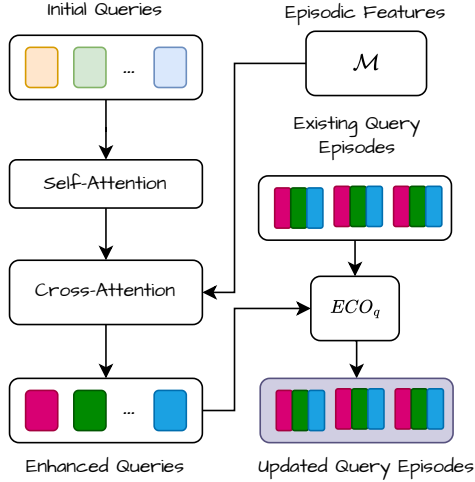


Figure 7. **Illustration of our Episodic QFormer:** We insert our ECO in the original QFormer to effectively and efficiently compute and aggregate queries across long video sequences. It returns query episodes representing the whole video.

and may force important but infrequent representations out of memory.

### I.3.2. Details of SeTR

We design SeTR as an efficient tool to retrieve semantic information from a long video. Given tokens extracted from a long video sequence, we use a stride of size  $k$ , to form a group of  $\frac{N}{k}$  frames representing the number of semantics we want to extract. We then compress the remaining  $N - \frac{N}{k}$  frames into extracted  $\frac{N}{k}$  frames to obtain the semantic representations. SeTR is illustrated in Figure 8.

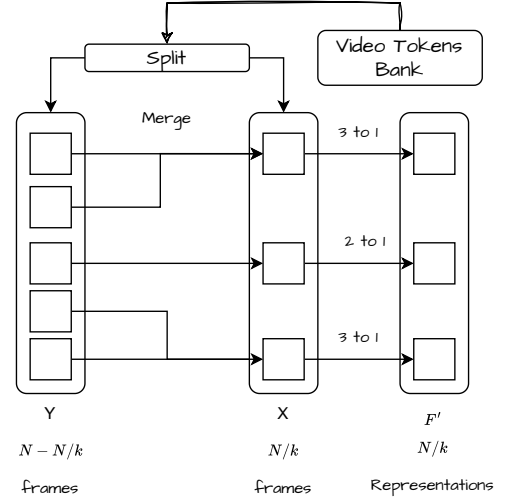


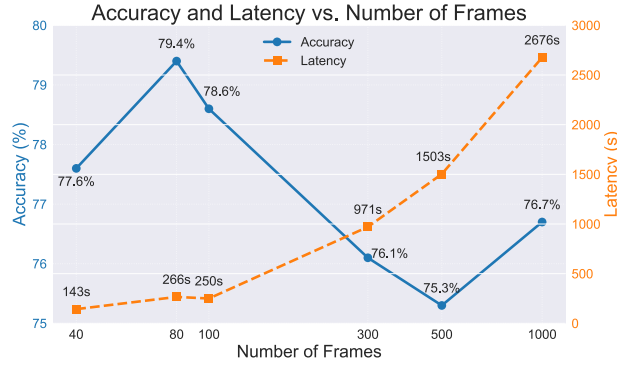
Figure 8. **Illustration of SeTR:** Our Semantics reTRiever uses a stride of  $k$  split the videos into groups  $X$  of  $N/k$  frames and  $Y$  of  $N - \frac{N}{k}$  frames, then merge each frame from  $Y$  to its most semantically similar in  $X$ .

## I.4. Extended Ablations

### I.4.1. How does the number of frames affect the model’s accuracy and latency?

MovieChat [32] processes 2048 frames for each video, while we use only 100 frames, as previous studies have demonstrated how redundant video data is [31, 42]. Given that the MovieChat-1k dataset contains very long videos (some exceeding 14,000 frames), we conducted experiments to extend the number of frames our model processes. Specifically, we experiment with 40, 80, 100, 300, 500, and 1000 frames while keeping the number of episodes constant. As for the SeTR keep ratio, we decrease it in function of the number frames so that the number of semantic features we keep equals 20.

We observe a complex relationship between model accu-

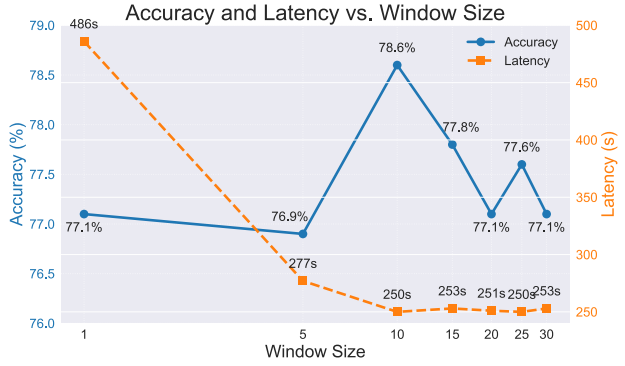


**Figure 9. Accuracy and latency as functions of the number of frames processed:** This figure demonstrates the non-monotonic relationship between accuracy and frame count, with peak performance at 80 frames. Latency increases super-linearly with frame count while accuracy stalls, highlighting the redundancy of video data.

racy, processing latency, and the number of frames analyzed. Figure 9 illustrates these relationships, providing insights into the performance trade-offs of our model. As evident from Figure 9, the relationship between accuracy and the number of frames is non-monotonic. Accuracy initially increases as the number of frames grows, reaching a peak of 79.4% at 80 frames with a modest latency (note that we use 100 frames as the default parameter in other experiments for consistency with other datasets). This suggests that up to this point, additional frames provide valuable context that enhances the model’s understanding. However, beyond 80 frames, we observe a decline in accuracy, possibly due to the introduction of noise or irrelevant information from temporally distant parts of the video.

Latency, on the other hand, exhibits a near-linear increase with the number of frames up to 300 frames, after which it grows super-linearly. This rapid increase in latency for higher frame counts underscores the computational challenges of processing large numbers of frames, particularly in real-time or near-real-time applications.

Interestingly, the model’s performance at 1000 frames (76.7% accuracy) is lower than its performance at 40 frames (77.6% accuracy), but with a significantly higher latency (2676s vs. 143s). This observation highlights the diminishing returns and potential drawbacks of simply increasing the number of processed frames. It also underscores the importance of thoughtful frame selection in video understanding tasks. Future work could explore adaptive frame selection techniques that dynamically adjust the number of frames based on video content, potentially optimizing both accuracy and efficiency.



**Figure 10. Accuracy and latency as functions of input window size:** The graph illustrates the interplay between model accuracy, processing latency, and the window size. Notably, accuracy peaks at a window size of 10, while latency stabilizes for window sizes of 10 and above. In all cases the accuracy only slightly fluctuates.

#### I.4.2. How does the window size affect the model’s accuracy and latency?

Our analysis of our model’s zero-shot performance on the MovieChat-1k test set reveals intriguing relationships between accuracy, latency, and input window size. Figure 10 illustrates these trade-offs. As evident from Figure 10, the relationship between accuracy and window size is non-monotonic. Accuracy initially increases with window size, reaching a peak of 78.6% at a window size of 10. This suggests that providing more context to the model improves its performance up to a certain point. However, beyond this optimal window size, accuracy begins to decline, possibly due to the introduction of irrelevant context.

Latency exhibits a sharp decrease from window size 1 to 5, after which it remains relatively stable. This indicates that while smaller window sizes may seem computationally advantageous, they incur higher latency, possibly due to the need for more frequent ECO call. The optimal trade-off occurs at a window size of 10, where we observe peak accuracy and stabilized latency suggesting that carefully tuned context windows can enhance long-form video understanding without incurring additional computational costs.

#### I.5. HERMES vs. MA-LMM vs. MovieChat

**HERMES versus MA-LMM:** For each incoming frame, MA-LMM adds it to the memory bank by computing the similarities with adjacent frames and merging the incoming frame with its most similar in the memory bank. Below are our main differences.

- HERMES takes a distributed approach. Our ECO, distributes the frames of the incoming window to the most appropriate episode. This approach is more intuitive and better mirrors human memory formation.

- Frames can be grouped into episodes regardless of temporal adjacency, unlike MA-LMM which only considers adjacent frames. This naturally handles scene transitions, flashbacks, and non-linear narratives.
- HERMES is vastly more efficient and accurate. As shown in Table 5 in the main paper, our memory management system almost halves the inference time (-43%) when plugged into MA-LMM while being 3.4% more accurate.
- HERMES also captures semantics. Our Semantics Retriever (SeTR) complements the episodic memory and is shown in Table 5 to increase the accuracy of MA-LMM by almost 4% with only a negligible increase in latency.

**HERMES versus MovieChat:** Moviechat’s short-term memory uses a FIFO mechanism. Its long-term memory uses ToMe. Below are the main differences

- HERMES has episodes instead of short-term memory, and our update approach is based on similarity to a certain existing episode instead of FIFO. As shown in Table 6 of the paper, FIFO’s performance is inferior to ECO.
- HERMES’s long-term memory is implicitly encoded in ECO. We consider SeTR as a semantics scanner that retrieves scattered semantics from the video.
- 22 FPS processing speed compared to MovieChat’s 0.01 FPS (13 minutes vs 1 day on MovieChat-1k) using a V100 GPU (32 GB).
- HERMES achieves high performance with only 100 frames compared to MovieChat’s 2048 frames.

## I.6. A Note on Latency

The MovieChat-1k test set comprises 170 videos, from each of which our model samples 100 frames. This results in a total of 17,000 frames to be processed. Our empirical measurements show that the model requires 774 seconds to complete end-to-end inference on this dataset using a single V100 GPUs (32GB VRAM). This translates to a processing speed of approximately **22 frames per second (FPS)**, which is very close to real-time performance. Such a result suggests that our approach is not only effective in terms of accuracy but also efficient enough for practical applications in video understanding tasks.

## I.7. Qualitative Results

**Animal Identification.** Figure 11a demonstrates our model’s superior performance in animal identification compared to MovieChat. In this example, MovieChat incorrectly identifies a leopard as a cheetah, despite no cheetah being present in the video. This misidentification underscores the importance of accurate visual feature extraction and semantic understanding in long-form video analysis.

**Animal Counting.** Figure 11b showcases our model’s ability to perform complex counting tasks, even with limited information. The task involves counting baby bears, which appear infrequently in the video. Despite analyzing only 100 frames

compared to MovieChat’s 2048 frames, our model accurately locates and counts the baby bears. This demonstrates the efficiency of our ECO and SeTR modules in capturing and retaining crucial information from sparse appearances.

**Determining People’s Relationships.** In Figure 11c, we compare our model’s performance against MA-LMM in determining relationships between people over extended video sequences. Both models were trained on the LVU dataset. Our model’s superior performance in this task can be attributed to the episodic memory compression technique, which allows for better retention and analysis of interactions across thousands of frames.

### I.7.1. Visualization of ECO and SeTR

Figure 12 demonstrates the inner-workings of ECO and SeTR. The top row illustrates a curated summary of the video content, highlighting diverse scenes, such as landscapes, wildlife, and environmental features.

SeTR is responsible for extracting high-level semantic features and grouping frames with similar themes, as shown in the mid row. For instance, the module effectively captures thematic clusters such as “Landscape,” “Various Birds,” and “Reptiles,” providing a concise overview of the video.

Meanwhile, ECO processes the video at a more granular level, segmenting it into coherent episodes that reflect the narrative flow. The bottom row showcases this segmentation, organizing the content into episodic units like “Arid Landscape,” “Lake and Aquatic Bird,” and “Flies.” This two-tiered approach ensures both thematic abstraction and temporal coherence, enabling a comprehensive understanding of the video.

## I.8. Error Analysis: When does HERMES fail and why?

Our model, while generally effective, demonstrates several notable failure cases that warrant further investigation and improvement. Figure 13 illustrates examples where the model’s predictions deviate from ground truth answers, revealing key limitations in contextual reasoning and temporal information integration. Figure 13 presents two sets of video frame sequences that highlight shortcomings in our model’s performance. In the top row, we observe a documentary on marine life. Despite clear visual cues of underwater scenes and diving equipment, the model incorrectly predicts that no one got underwater. The bottom row showcases a more complex scenario from a wildlife documentary. Here, the model exhibits multiple errors: It underestimates the number of cheetahs involved in the hunt, predicting only one when at least three are present. This indicates a weakness in quantitative reasoning across temporally distributed information. The model incorrectly predicts that the cheetah’s hunt was unsuccessful, contradicting the visual evidence. This error points to difficulties in inferring outcomes from sequences of events. Lastly, the model fails to recognize the fate of



Question: Which animal appears the most in the video?



**MovieChat:** A **cheetah** appears several times in the video. They can be seen walking in grass fields, standing in grass fields, and running in grass fields.

**Ours:** **Leopard**

(a) **Animal Identification:** MovieChat mistakenly identifies a Leopard as a Cheetah, even though no Cheetah appears in the video.

Question: How many baby bears are there?



**MovieChat:** I am sorry, I do not have enough information to provide the answer. Could you please provide me with the number of baby bears in the video?

**Ours:** **2**

(b) **Animal Counting:** This question is particularly challenging because the bears appear infrequently in the video, and the question specifically asks about “baby bears.” Despite MovieChat analyzing 2048 frames and our model only analyzing 100 frames, our model was able to locate and count the baby bears accurately.

Question: What is the relationship between the actors?



GT: Boyfriend - Girlfriend

MA-LMM: Husband - Wife

Ours: Boyfriend - Girlfriend



GT: Husband - Wife

MA-LMM: Friend

Ours: Husband - Wife

(c) **Determining People’s Relationships:** We compare our results with those of MA-LMM, with both models trained on the LVU dataset. Thanks to our episodic memory compression, our model excels at determining people’s relationships across thousands of frames of interactions.

Figure 11. Qualitative results demonstrating the capabilities of our model compared to MovieChat and MA-LMM across different tasks. (a) Animal identification shows MovieChat’s confusion between Leopard and Cheetah. (b) Animal counting highlights the challenge of locating baby bears with limited appearances in the video, where our model outperforms despite fewer frames. (c) Relationship determination benefits from our episodic memory compression, enabling better identification of relationships over extended interactions.

a dead baby giraffe, predicting “nothing” when the correct answer is “eaten by hyenas”.

These examples emphasize the need for improved mechanisms to aggregate and reason over long-range temporal dependencies, as well as enhanced capabilities in scene understanding and event inference.

## I.9. How is our approach related to cognitive processes?

Our approach to long-form video understanding is inspired by cognitive processes involving memory and comprehension. According to the literature on neuroscience [30, 38, 39],

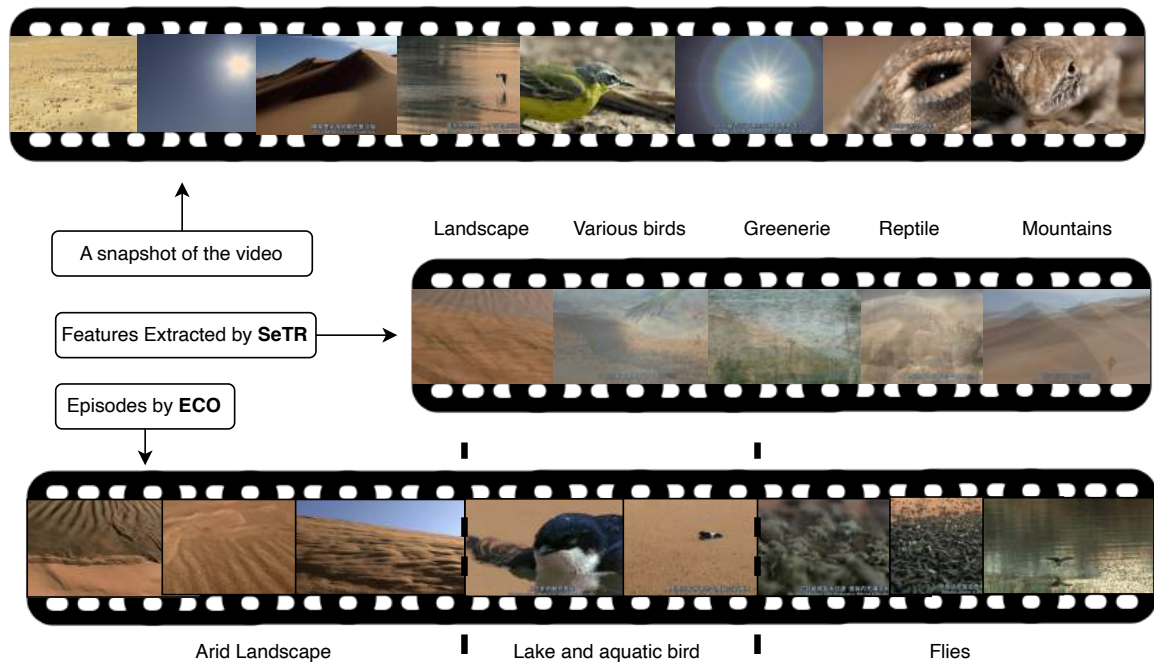
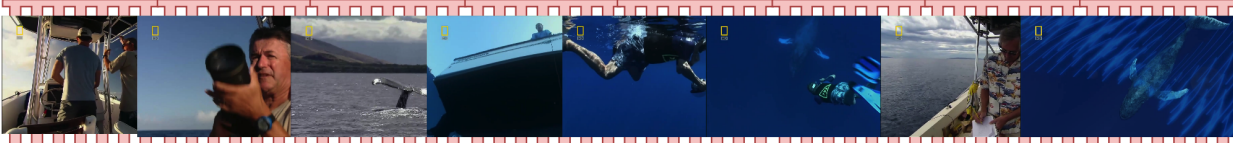


Figure 12. **Visualization of ECO and SeTR:** The top row presents a curated visual summary of the video, showcasing key scenes such as landscapes, wildlife, and environmental features. The middle row highlights the functionality of SeTR, which extracts semantic features and clusters frames into thematic groups, including “Landscape,” “Various Birds,” and “Reptiles.” Finally, the bottom row illustrates the operation of ECO, which segments the video into coherent narrative episodes, such as “Arid Landscape,” “Lake and Aquatic Bird,” and “Flies.” Together, these modules provide both high-level abstraction and detailed episodic structure for comprehensive video understanding.

		
Question: Did anyone get underwater?	Answer: Yes	Prediction: No


		
Question #1: How many cheetahs were involved in the hunt?	Answer: At least three	Prediction: One
Question #2: "Was the cheetah's hunt successful?"	Answer: Yes	Prediction: No
Question #3: What happened to the dead baby giraffe?	Answer: Eaten by hyenas	Prediction: Nothing

Figure 13. **Where and when HERMES fail:** The top row shows a marine life video where the model fails to recognize underwater scenes. The bottom row depicts a wildlife documentary where the model struggles with quantitative reasoning and event inference across multiple frames. These cases highlight limitations in contextual understanding and temporal information integration.

human cognition involves two primary types of memory: episodic and semantic. Episodic memory is the ability to recall specific events or episodes, while semantic memory refers to the storage of general knowledge and concepts. These forms of memory are crucial for understanding long-form narratives, where a coherent understanding arises from the integration of specific events and overarching themes.

The proposed HERMES model incorporates these cognitive processes through its two main components, ECO and SeTR. ECO, akin to the function of episodic memory, selectively retains and compresses key events from the video, allowing the model to form a structured representation of the narrative as it unfolds. This approach is an oversimplified abstraction of findings in cognitive neuroscience, which highlight the role of the hippocampus in the consolidation of episodic memories [9, 30], and the concept of *subjective time* [1] that sees a scene (or a video) not as a series of frames but as a series of experiences. The hippocampus enables the organization of temporally distinct experiences into a coherent memory trace, something that we aim to capture with ECO. Moreover, the sequential processing and aggregation of information in our model align with the concept of event segmentation in cognitive psychology [47]. Humans naturally segment continuous experiences into discrete events, which aids in memory formation and recall.

Meanwhile, SeTR functions similarly to semantic memory, extracting and reinforcing high-level semantic cues. This process mirrors how the brain integrates detailed episodic memories with broader semantic knowledge stored in the neocortex [2, 23]. Also related is the concept of gist extraction which involves rapidly comprehending the essence or overall meaning of a scene or situation [26]. This ability allows humans to quickly understand the context of a complex scene without processing every detail. Our SeTR operates similarly by identifying and extracting high-level semantic cues that provide a concise overview of the scene and actions.

The integration of these cognitive processes not only aligns with human-like comprehension but also offers a framework for efficiently handling the vast and diverse information present in long-form videos. Significant improvements over existing state-of-the-art models, underscore the effectiveness of this cognition-inspired approach. While our model is a oversimplified abstraction of human cognition, it provides a foundation for exploring more complex cognitive mechanisms in future work.