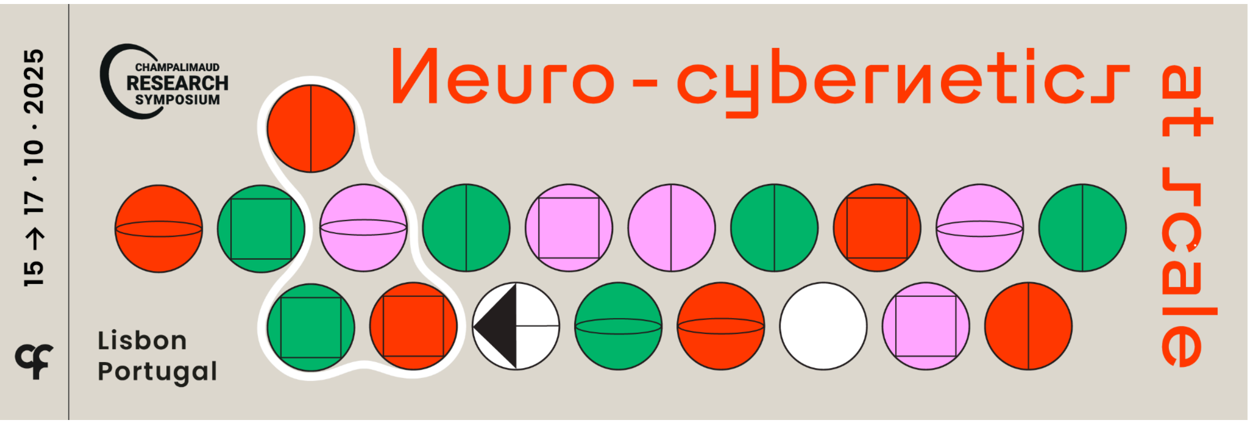

Learning What Matters: An Episodic Memory Perspective on Video-Language Models

Inês Baptista 1, Daniel MacNamee 1, 2
1. Champalimaud Foundation, Lisbon, PT
2. Champalimaud Foundation, Lisbon, PT

Human memory is remarkably efficient, actively compressing less important details while prioritizing and retaining what is novel, emotionally significant, or functionally useful. Episodic memory, in particular, orchestrates our continuous experience by dynamically segmenting it into discrete, meaningful events. This event segmentation supports crucial cognitive functions like abstraction, generalization from sparse data, and the understanding of narratives. Neurocognitive research suggests that this segmentation process is tightly linked to surprise — moments of high prediction error — triggered when expectations about the unfolding environment fail. At these points, the hippocampus initiates memory updates, marking event boundaries and selectively encoding contextually significant moments.

On the other hand, current AI models struggle to process extended, continuous and high-dimensional data like long videos and movies. Despite new developments in the field, Video-language models (VLMs) struggle in organizing asynchronously visual, auditory, and textual inputs; interpret nuanced emotional tone; and understand connections between temporally distant events in a narrative — for example, understanding how a character's early decision shapes the outcome of a later scene. Crucially, unlike humans who naturally structure their past experiences into coherent "mental videos" with a clear sense of how one event leads to another, current VLMs lack mechanisms to automatically break down a continuous stream of data into meaningful individual events or to encode the temporal coherence vital for narrative understanding.

In this work, we investigate whether the neurocognitive framework of surprise-driven event segmentation provides a viable model for enhancing memory in VLMs. We also study how latent internal representations in VLMs evolve over long videos and whether cues like prediction failure (surprise), contextual shifts, or affective salience correlate with changes in the model's latent space. Ultimately, our goal is to guide the development of more brain-aligned VLMs —

enhancing their ability to select and encode the most memorable events through mechanisms inspired by episodic recall, and improving performance on video question answering and long-range temporal reasoning tasks.

episodic memory, video-language models, event segmentation, computer vision, artificial intelligence, interpretability