SELECTIVE ASSOCIATION IN CONTEXT MEMORY FOR TASK-SPECIFIC VIDEO UNDERSTANDING

Shaurjya Mandal
Harvard University
Massachusetts General Hospital
smandal2@mgh.harvard.edu

Nutan Sharma
Harvard University
Massachusetts General Hospital
nsharma@mgb.org

ABSTRACT

Multimodal foundation models, particularly those instruction fine-tuned for vision-language tasks, have recently gained prominence for their ability to parse and analyze complex video streams. Despite their effectiveness for broad, general-purpose queries, these models often struggle with domain-specific questions that demand deeper contextual understanding. The core limitation lies in their reliance on vision-language grounding extracted from raw video frames, which does not adequately capture nuanced context when the task is more specialized. In this paper, we introduce a method for "selective association in context memory" that addresses this shortcoming. Our approach leverages a targeted "association block" drawn from the extensive content in the model's context window, focusing attention on the most relevant sub-scenes. By selectively filtering and organizing the visual stream, we enable more precise alignment of textual and visual cues for task-specific understanding. This mirrors the human cognitive strategy of associating smaller, relevant incidents to effectively recall and interpret them. We state examples to demonstrate the utility of our approach using examples from the medical domain—specifically, in analyzing videos of neurological movement disorders, where identifying subtle clinical cues requires robust context awareness.

1 INTRODUCTION

Multimodal foundation models, such as large language models (LLMs) trained to parse extensive video data, are increasingly capable of addressing a range of vision-language tasks. Recent research has placed growing emphasis on associative memory—the capacity of systems to recall and leverage domain-specific information in context—both in understanding how humans learn from limited cues and in implementing more robust, adaptable AI architectures. For instance, the Neuro-Symbolic Concept Learner demonstrates how bridging object-based visual representations with executable symbolic programs can yield strong performance on visual question answering by learning a compositional structure of scenes, words, and questions from natural supervision Mao et al. (2019). Parallel lines of research draw from associative memory principles to tackle in-context learning (ICL), enabling Transformers to respond adaptively to new prompts that vary from the data observed during training Burns et al. (2024). Additionally, theories of entropic associative memory highlight the importance of handling partial or noisy cues in a structured, memory-like system that unifies sub-symbolic and symbolic representations Pineda et al. (2021). The state-of-the-art models still face limitations when they are confronted with specialized or domain-intensive tasks. Their general-purpose training often overlooks subtle cues that experts prioritize, especially in domains like neurology, where detecting disorders such as essential tremor or dystonia demands close attention to particular, nuanced features of patient movements. Simple vision-language grounding—where the model merely maps raw frames to text—falls short for queries like "Is essential tremor present?" or "Do you see signs of dystonia?" because it cannot readily associate local sub-scenes of clinical interest (for instance, short bursts of a patient's hand tremor) with the complex diagnostic criteria embedded in expert knowledge.

This paper addresses the necessity of a more **specialized context-association mechanism** for video-based understanding tasks that demand domain-specific reasoning. Building on neuro-symbolic methods, in-context associative memory architectures, and the concept of entropic representation,

we propose a system that isolates sub-scenes or "mini-episodes" of interest for each clinical query. In doing so, it not only improves interpretability—much like a human expert recalling the salient details of a past incident—but also substantially reduces the noise in video-based feature extraction. Although still a proposal, this selective association framework offers a pathway toward robust, context-rich video understanding in specialized areas, including medical diagnostics, where subtle visual cues carry critical diagnostic importance.

2 Related Works

The continued expansion of large-scale, multimodal datasets and high-capacity architectures has driven significant advances in video foundation models (ViFMs). As highlighted in a recent survey, these models aim to learn universal representations from massive video collections, often incorporating additional modalities such as text or audio to improve temporal grounding and semantic alignment Madan et al. (2024). In particular, the shift toward contrastive pretraining strategies has enabled ViFMs to map corresponding multimodal cues—like frames and textual captions—into a shared embedding space, facilitating tasks such as video retrieval, captioning, and question-answering. These developments lead to the state-of-the-art ViFMs increasingly tackling *long-form and multimodal reasoning*. For instance, InternVideo2 adopts a progressive training paradigm—combining masked video modeling, contrastive alignment, and next-token prediction—to yield flexible representations that can handle both short clips and prolonged video streams Wang et al. (2024). Such techniques parallel broader trends in computer vision and natural language processing, where unifying cross-modal learning objectives often leads to improved performance across diverse downstream applications.

3 Methodology

3.1 Overview of Selective Associative Memory

We propose a *memory-enabled* framework that progressively extracts and stores relevant sub-scenes from a video into a dedicated *associative memory block*, with the goal of performing more robust, context-aware tasks (e.g., detecting subtle clinical cues in neurological disorders). Let $\mathcal{V}$ be a video decomposed into frames or short clips $\{v_1, v_2, \ldots, v_T\}$, each of which is encoded by a backbone network F_θ, commonly a CNN+Transformer or a 3D CNN, yielding frame-wise embeddings $\{\mathbf{e}_t\} \in \mathbb{R}^d$. To incorporate domain knowledge such as "look for signs of essential tremor or dystonia," we introduce a *context vector* $\mathbf{c} \in \mathbb{R}^d$ that summarizes high-level task-specific instructions or prior knowledge. Our method maintains K memory slots $\{\mathbf{m}_k\}_{k=1}^K$, each also in $\mathbb{R}^d$, representing sub-scenes judged to be *relevant* for the task. Each video segment's *storage score* S_t governs whether it is added or whether an existing slot is replaced.

3.2 Memory Creation and Update

The decision to store a given segment embedding $\mathbf{e}_t$ is determined by balancing *domain relevance* (matching the context vector $\mathbf{c}$) and *novelty* (avoiding duplication). First, we compute a contextual relevance:

$$r(\mathbf{e}_t, \mathbf{c}) = \frac{\mathbf{e}_t^\top \mathbf{c}}{\|\mathbf{e}_t\| \|\mathbf{c}\|}, \tag{1}$$

which is a standard cosine similarity that measures how well the segment aligns with the domain or task cues. Second, we quantify *novelty* by comparing $\mathbf{e}_t$ to the *closest* memory slot in $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_K\}$:

$$\text{novelty}(\mathbf{e}_t, \mathcal{M}) = 1 - \max_{1 \le k \le K} \frac{\mathbf{e}_t^\top \mathbf{m}_k}{\|\mathbf{e}_t\| \|\mathbf{m}_k\|}. \tag{2}$$

By subtracting the maximum similarity from 1, high values indicate $\mathbf{e}_t$ does not closely resemble any stored items, fostering a more diverse set of memory entries. Combining these two terms yields:

$$S_t = \alpha\, r(\mathbf{e}_t, \mathbf{c}) \;+\; (1 - \alpha)\, \text{novelty}(\mathbf{e}_t, \mathcal{M}), \tag{3}$$

where $\alpha \in [0, 1]$ controls the trade-off. If $S_t > \tau_s$, the model either *adds* $\mathbf{e}_t$ as a new slot (if space remains) or *replaces* the least recently accessed slot. This can be done by overwriting the slot with

the lowest usage frequency or the smallest long-term relevance score. For efficient operation in longer videos, one may implement additional heuristics such as *soft updates*, where $\mathbf{m}_k$ is blended with $\mathbf{e}_t$:

$$\mathbf{m}_k \leftarrow \gamma\, \mathbf{m}_k + (1-\gamma)\, \mathbf{e}_t, \tag{4}$$

for some $\gamma \in [0,1]$. This avoids discarding similar embeddings altogether and allows the memory representation to evolve over time. In practice, we also track a recency weight or usage counter to avoid continually replacing the same slot and to better handle repetitive segments.

3.3 Context-Aware Retrieval

When processing queries at inference or encountering a *test* segment that must be analyzed in-depth, we retrieve from the associative memory to form a context-enriched representation. A query embedding $\mathbf{q}$ may come from: (1) the current video segment $\mathbf{e}_{t'}$, or (2) an external question embedding (in medical settings, e.g. "Is there a tremor in the patient's hand?"). We compute softmax-based similarities:

$$\alpha_k = \frac{\exp\big(\mathbf{q}^\top \mathbf{m}_k/\tau_r\big)}{\sum_{j=1}^{K} \exp\big(\mathbf{q}^\top \mathbf{m}_j/\tau_r\big)}, \qquad \mathbf{r} = \sum_{k=1}^{K} \alpha_k\, \mathbf{m}_k, \tag{5}$$

where τ_r is a softmax temperature that controls how concentrated or diffuse the retrieval distribution is. The retrieved vector $\mathbf{r}$ is then concatenated with $\mathbf{q}$ to form $[\mathbf{q};\mathbf{r}] \in \mathbb{R}^{2d}$, which is fed into a classification, detection, or regression head, depending on the task. This retrieval emphasizes *content actually stored in memory*—e.g., frames capturing visible tremors or relevant motions—yielding a context-aware feature that augments the original embedding. For interpretability, we can visualize which memory slots receive the highest α_k weights, linking final decisions back to critical sub-scenes.

3.4 Training Objective

In addition to the standard task loss, $\mathcal{L}_{\text{task}}\big([\mathbf{q};\mathbf{r}],\, y\big)$, we incorporate a *memory-utility* regularizer $\mathcal{L}_{\text{mem}}$ that encourages the model to *correctly prioritize* segments deemed crucial by a ground-truth label or heuristic. Let $\mathbb{I}[\text{important}(v_t)]$ indicate that v_t is labeled (in training) as essential for the downstream query. We define:

$$\mathcal{L}_{\text{mem}} = \sum_{t=1}^{T} \max\big(0, \tau_s - S_t\big)\, \mathbb{I}[\text{important}(v_t)], \tag{6}$$

which penalizes cases where the computed S_t is *below* τ_s despite v_t being an important segment. Thus, whenever an essential segment is overlooked by the memory gating, the penalty grows. The overall objective is then:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \lambda\, \mathcal{L}_{\text{mem}}, \tag{7}$$

where λ moderates the emphasis on correct memory decisions. Training can be end-to-end, back-propagating through F_θ, the memory gating rules (implemented via differentiable approximations or carefully managed straight-through updates), and the final classifier head. Alternatively, parts of F_θ may be frozen if the domain embeddings are already robustly pre-trained, focusing learning capacity on refining the gating rules and task head. By jointly optimizing for both *accuracy on the domain task* and *the selective memory process*, the model learns to highlight precisely those sub-scenes that matter most, thereby improving both performance and interoperability.

4 Conclusion and Future Works

We have introduced a *selective associative memory* framework that stores task-relevant sub-scenes from a video stream into a dedicated memory block, dynamically guided by contextual relevance and novelty. Our proposed memory-utility regularizer further drives the model to store important frames and enhances interpretability, creating a direct link between task outputs and associated sub-scenes. We aim to extend the memory representation to capture temporal linkages among stored frames, enabling more robust reasoning over multi-step processes. Incorporating additional modalities (e.g., audio or sensor data) could also improve retrieval precision in tasks demanding multimodal evidence, such as neurological assessments involving motion and speech cues.

REFERENCES

Thomas F Burns, Tomoki Fukai, and Christopher J Earls. Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture. *arXiv preprint arXiv:2412.15113*, 2024.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

Luis A Pineda, Gibrán Fuentes, and Rafael Morales. An entropic associative memory. *Scientific reports*, 11(1):6948, 2021.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416. Springer, 2024.