∞-VIDEO: A Training-Free Approach to Long Video Understanding via Continuous-Time Memory Consolidation

Saul Santos[1] António Farinhas[1] Daniel McNamee[2] André F. T. Martins[1,3]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon ELLIS Unit, Portugal

[2]Neuroscience Programme, Champalimaud Research, Lisbon, Portugal [3]Unbabel, Lisbon, Portugal

Outline

- **Problem:** Long videos exceed model context; subsampling loses critical details.
- **Inspiration:** Human memory consolidates key events over time.
- **Goal:** Understanding full videos in one pass without missing important content.
- **Approach:** Adapt ∞-former [1] to video via continuous-time visual memory—no retraining needed.
- **Our method shows improved performance on video QA tasks with Video-LLaMA and VideoChat2.**

From Discrete to Continuous Attention

Discrete attention uses tokens; we instead model input as a continuous signal $x(t)$ on $[0, 1]$, expressed via basis functions $\psi(t)$:

$$x(t) = B^\top \psi(t)$$

This allows using a probability *density* $p(t)$ instead of softmax. B is computed via Ridge regression.

Continuous Attention and Long-Term Memory (LTM)

Project the continuous input $x(t) = B^\top \psi(t)$ to get keys and values:

$$k^h(t) = (W_K^h)^\top x(t) = (W_K^h)^\top B^\top \psi(t),$$
$$v^h(t) = (W_V^h)^\top x(t) = (W_V^h)^\top B^\top \psi(t)$$

Queries: $Q^h = [q_1^\top, \ldots, q_R^\top] = YW_Q^h$. Similarity and Gibbs PDF:

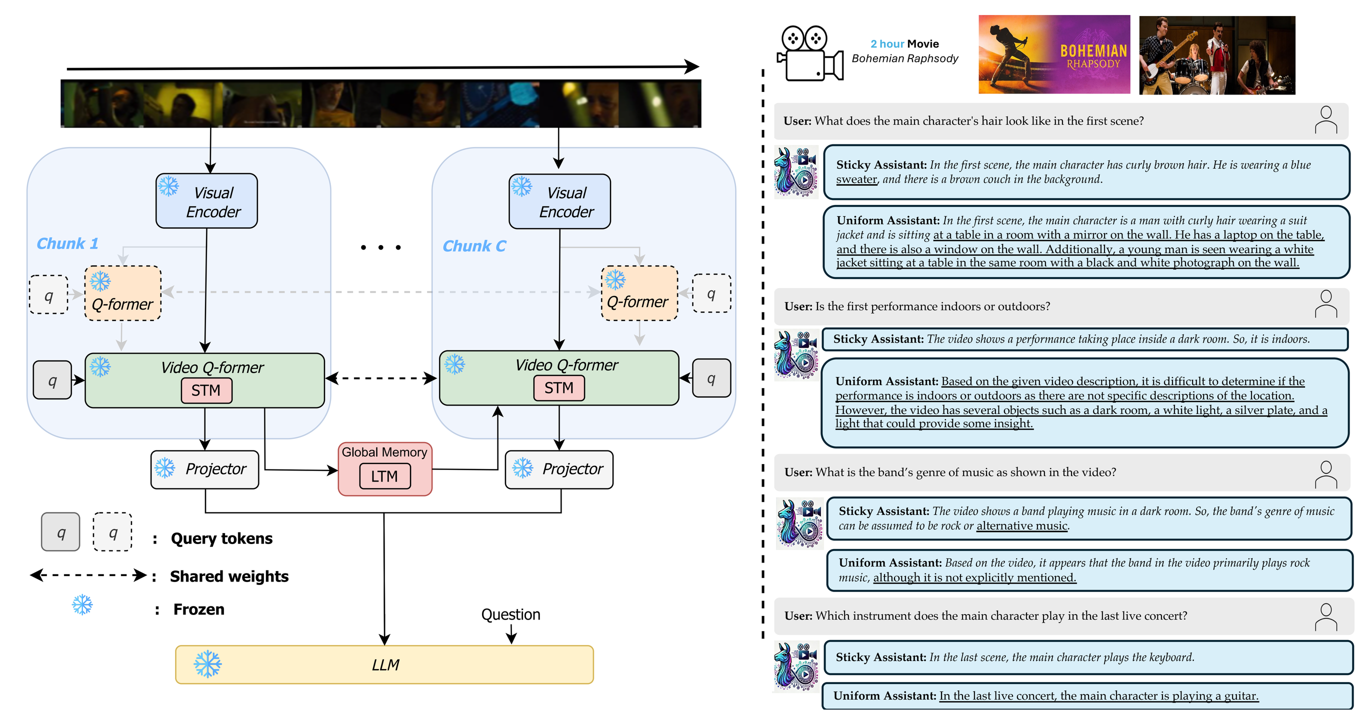
$$s_i^h(t) = q_i^\top k^h(t), \qquad p_i^h(t) = \frac{e^{s_i^h(t)}}{\int e^{s_i^h(t')} dt'}$$

Attention via expectation:

$$Z_i^h = \mathbb{E}_{p_i^h}[v^h(t)] = (W_V^h)^\top B^\top \int p_i^h(t) \psi(t) dt$$

Final LTM: concatenate heads and apply output projection.

Overall Architecture



The final output of our modified video Q-former layer is a weighted average of the standard Short-Term Memory (STM) attention and our new LTM context:

$$Z = \alpha Z_{\text{STM}} + (1 - \alpha) Z_{\text{LTM}}$$

To feed a fixed-size representation to the LLM, we compute a running average of the projected video token embeddings E_c from each of the C chunks:

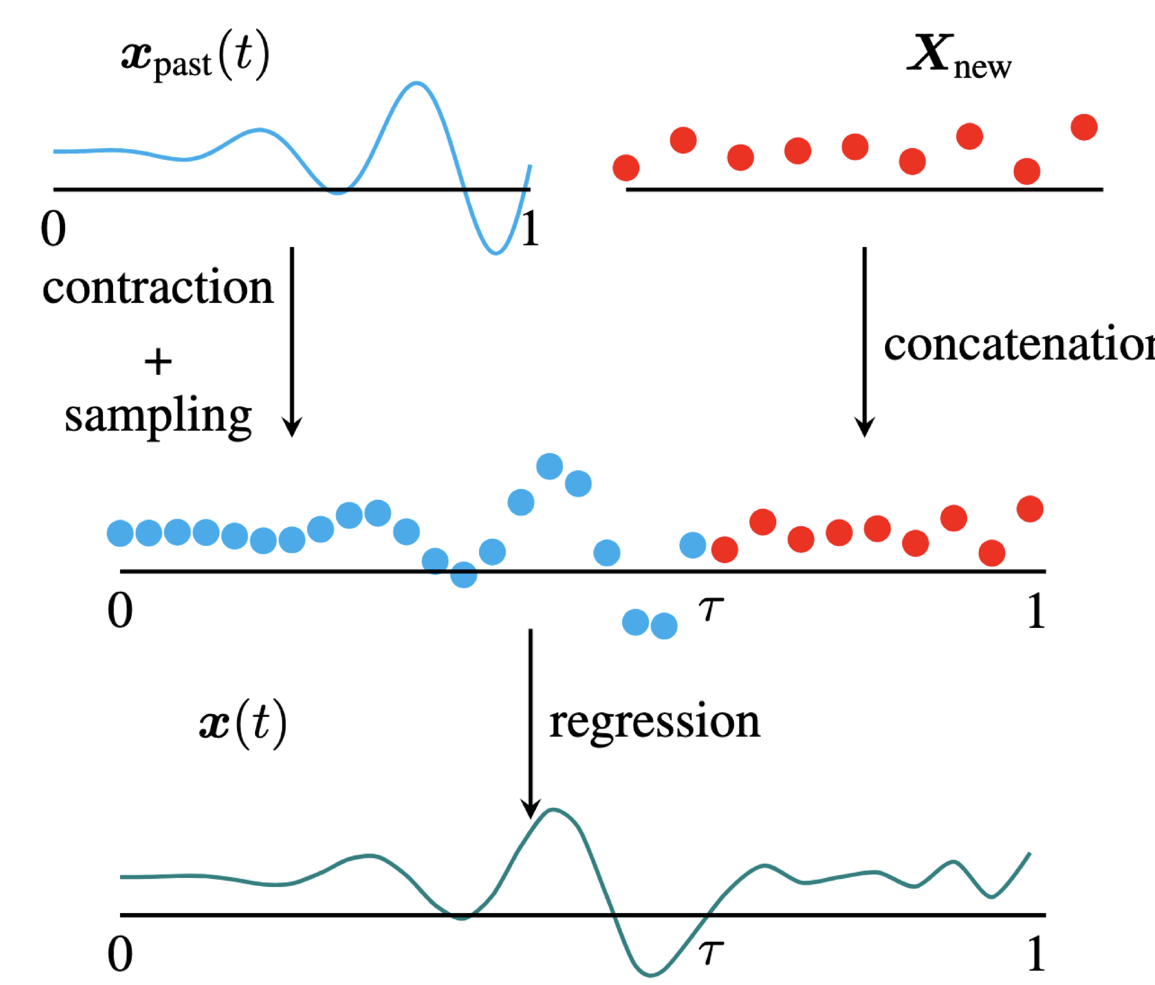
$$\bar{E}_c = \frac{C-1}{C}\bar{E}_{c-1} + \frac{1}{C}E_c$$

Continuous-Time Memory Consolidation

The LTM is updated for each new chunk in FOUR steps: As new video chunks arrive, the LTM is updated:

1. **Sample:** Evaluate the current LTM signal $x(t)$ at T locations.
2. **Contract:** The past context is mapped to a smaller interval $[0, \tau]$, inducing a "forgetting" factor.
3. **Concatenate:** The contracted past context is combined with the new chunk's context.
4. **Regress:** A new continuous signal $x_{\text{new}}(t)$ is fit to the combined context over $[0, 1]$.
5. **Continuous Attention:** Apply continuous attention over $x_{\text{new}}(t)$.

Continuous-Time Memory Consolidation



Sticky Memories

Motivation: Uniform memory allocation is inefficient.
Idea: Prioritize relevant regions by allocating memory proportionally, inspired by resource-based models [2, 3, 4].
Analogy: Mimics *non-local replay* in the brain, where past events are selectively reactivated [5, 6].

Long-Term Open-Ended Question Answering

Method	LLM	#Frames	Medium	Long	Avg
Video-LLaVA	Vicuna-7B	8	38.0	36.2	39.9
ShareGPT4Video 8B	-	16	36.3	35.0	39.9
Chat-UniVi-v1.5	Vicuna-7B	64	**40.3**	35.8	40.6
Qwen-VL-Chat	Qwen-7B	4	38.7	37.8	41.1
VideoChat2	Mistral-7B	32	37.9	38.0	42.1
∞-VideoChat2 (no LTM)	Mistral-7B	128	39.6	38.8	42.3
∞-VideoChat2 (uniform)	Mistral-7B	128	40.0	38.8	42.4
∞-VideoChat2 (sticky)	Mistral-7B	128	40.2	**38.9**	**42.4**

Our sticky memory method yields gains over baselines on Video MME!

Long-Term Open-Ended Question Answering

Method	LLM	Number of Frames	Accuracy	Score	CI	DO	CU
Video Chat	Vicuna-7B	32	61.0	3.34	3.26	3.20	3.38
Video-ChatGPT	Vicuna-7B	100	44.2	2.71	2.48	2.78	3.03
Video LLaMA-Based Models							
Video LLaMA	Vicuna-7B	32	51.4	3.10	3.30	2.53	3.28
MovieChat	Vicuna-7B	2048	67.8	3.81	3.32	3.28	3.44
MovieChat+	Vicuna-7B	2048	66.4	3.67	3.70	3.30	3.62
∞-Video LLaMA (no LTM)	Vicuna-7B	2048	68.0	3.76	3.72	3.33	3.71
∞-Video LLaMA (uniform)	Vicuna-7B	2048	66.5	3.69	3.60	3.31	3.58
∞-Video LLaMA (sticky)	Vicuna-7B	2048	**72.2**	**3.88**	**3.89**	3.47	3.79
∞-Video LLaMA (no STM uniform)	Vicuna-7B	2048	62.4	3.75	3.36	3.38	3.52
∞-Video LLaMA (no STM sticky)	Vicuna-7B	2048	59.2	3.68	3.30	3.30	3.44
VideoChat2-Based Models							
VideoChat2	Mistral-7B	16	62.2	3.72	3.46	3.60	3.69
∞-VideoChat2 (no LTM)	Mistral-7B	128	63.9	3.74	3.54	3.60	3.73
∞-VideoChat2 (uniform)	Mistral-7B	128	64.1	3.73	3.54	3.60	3.75
∞-VideoChat2 (sticky)	Mistral-7B	128	63.9	3.74	3.55	3.63	3.74
∞-VideoChat2 (no STM uniform)	Mistral-7B	128	65.7	3.78	3.65	3.60	3.84
∞-VideoChat2 (no STM sticky)	Mistral-7B	128	66.5	3.85	3.71	**3.68**	**3.96**

∞-Video with sticky memories achieves the best results on MovieChat-1K, outperforming all baselines!

Qualitative Analysis

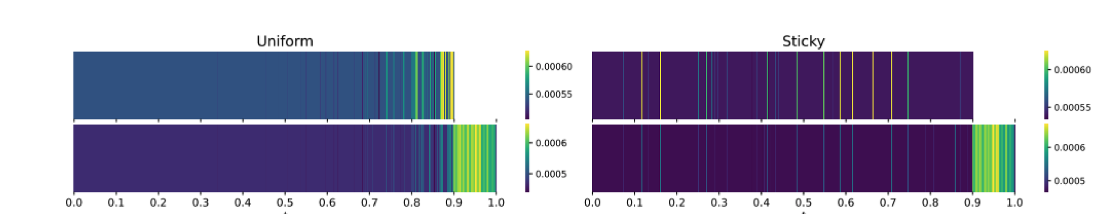


Figure 3. **(Top)** LTM attention density on the $[0, \tau]$ interval for the *Interstellar* trailer, using sticky memories in the final chunk of the ∞-Video LLaMA video Q-former's last layer. **(Bottom)** The same attention density map, extended over the full t interval.

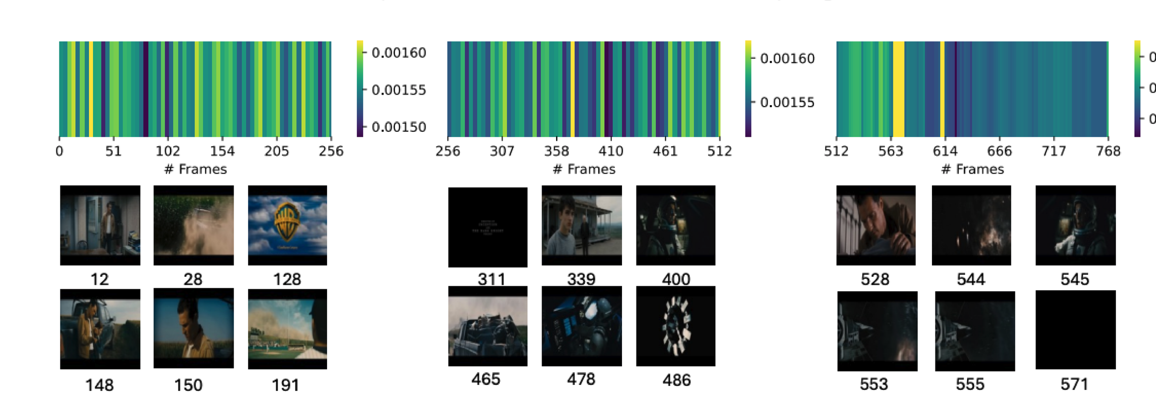


Figure 4. Highest continuous attention density frames selected using sticky memories in the *Interstellar* trailer for ∞-Video LLaMA across 3 chunks. **(Left)** Interval: $[0, \tau^2]$. **(Middle)** Interval: $(\tau^2, \tau]$. **(Right)** Interval: $(\tau, 1]$.

References

[1] Pedro Henrique Martins, Zita Marinho, and André FT Martins. ∞-former: Infinite memory transformer. In *Proc. ACL*, 2022.

[2] W. Ma, M. Husain, and P. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17: 347–356, 2014. doi: 10.1038/nn.3655. URL https://doi.org/10.1038/nn.3655.

[3] Oliver Hardt, Einar O Einarsson, and Karim Nader. A bridge over troubled water: reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual Review of Psychology*, 61:141–167, 2010. doi: 10.1146/annurev.psych.093008.100455.

[4] Yadin Dudai, Avi Karni, and Jan Born. The consolidation and transformation of memory. *Neuron*, 88:20–32, 2015.

[5] Margaret F Carr, Shantanu P Jadhav, and Loren M Frank. Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, 14:147–153, 2011.

[6] Daniel C McNamee. The generative neural microdynamics of cognitive processing. *Current Opinion in Neurobiology*, 85:102855, 2024. doi: 10.1016/j.conb.2024.102855. Epub 2024 Feb 29.

Acknowledgments: ERC (DECOLLAGE, ERC-2022-CoG 101088763), EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), Portuguese PRR (Center for Responsible AI, C645008882- 00000055), FCT (contract UIDB/50008/2020).