DECEMBER 2026 · NEURIPS 2026

EGOSTREAM

A diagnostic benchmark for streaming episodic memory in egocentric vision

Rosario Forte · Giuseppe Lando · Antonino Furnari

Image Processing Laboratory · University of Catania · iplab.dmi.unict.it

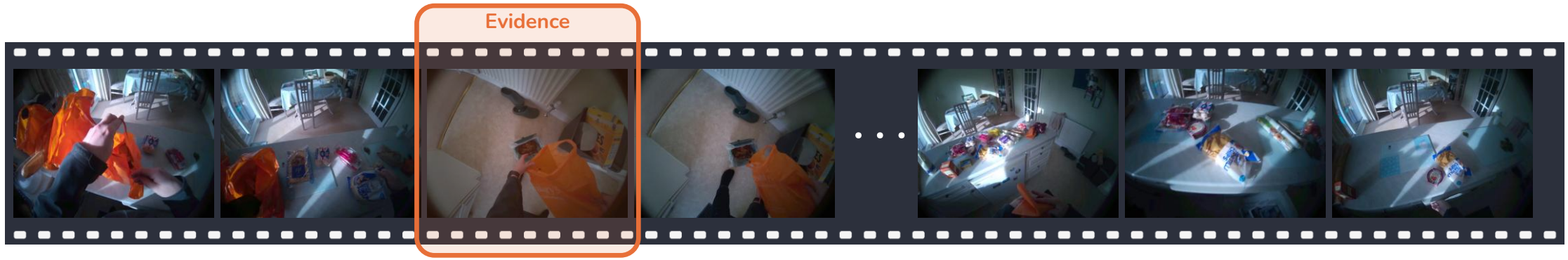
Collaborative AI Research Group · coai-iplab.github.io

rosario.forte@phd.unict.it · coai-iplab.github.io/EgoStream

paper · data · code

THE PROBLEM

One stream, two questions, one score



ASKED 5 S LATER · DETAIL MEMORY

What colour was the bag I shook over the bin?

model answer ✗ wrong

ASKED 1 H LATER · CAUSAL MEMORY

Why did I shake a bag over the food bin?

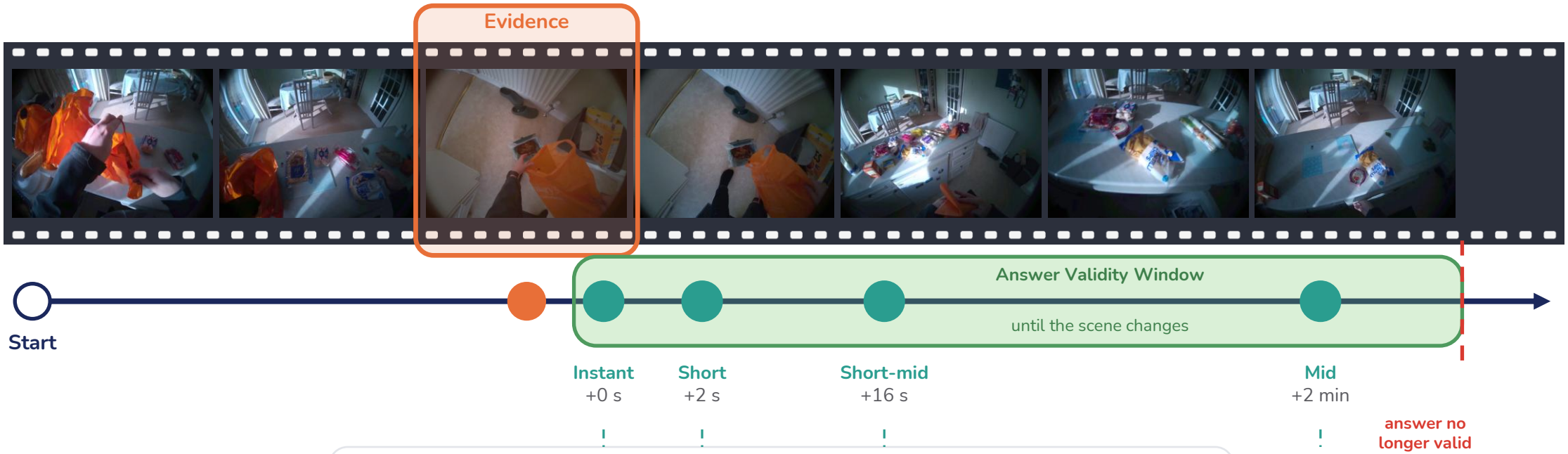
model answer ✗ wrong

PRIOR BENCHMARKS

✗ + ✗ → one accuracy, both questions

→ **Can't answer — or too far in time?**

The Answer Validity Window



SAME QUESTION, ASKED 4 TIMES INSIDE THE WINDOW · CAUSAL MEMORY

Why did I shake a bag over the food bin?

① Answer Validity Window
how long the answer to a question stays valid

② Same question, many recall times
all inside the window, so a wrong answer is forgetting — not a changed scene

③ What this unlocks
measuring for how long the information stays in memory

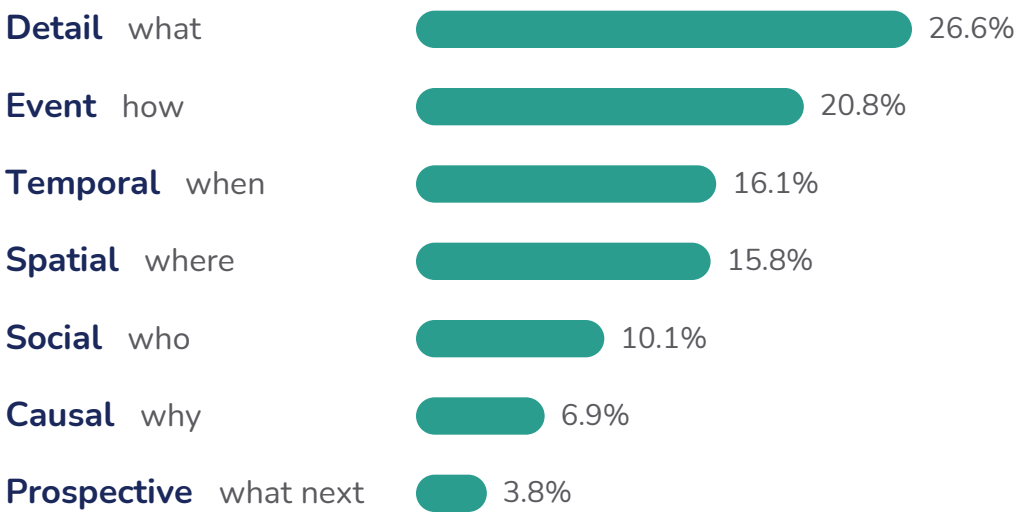
THE BENCHMARK

Seven categories, seven regimes

2,250 memory questions **8,528** recall-conditioned evaluations **5** source datasets

What is asked · 7 memory categories

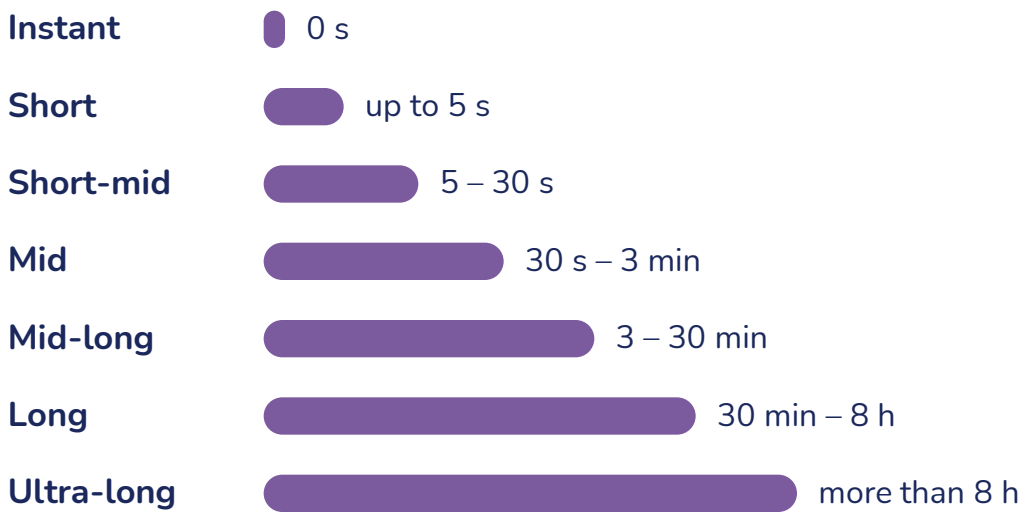
one primary category per question



Five sources, each skewed; together they cover all seven.

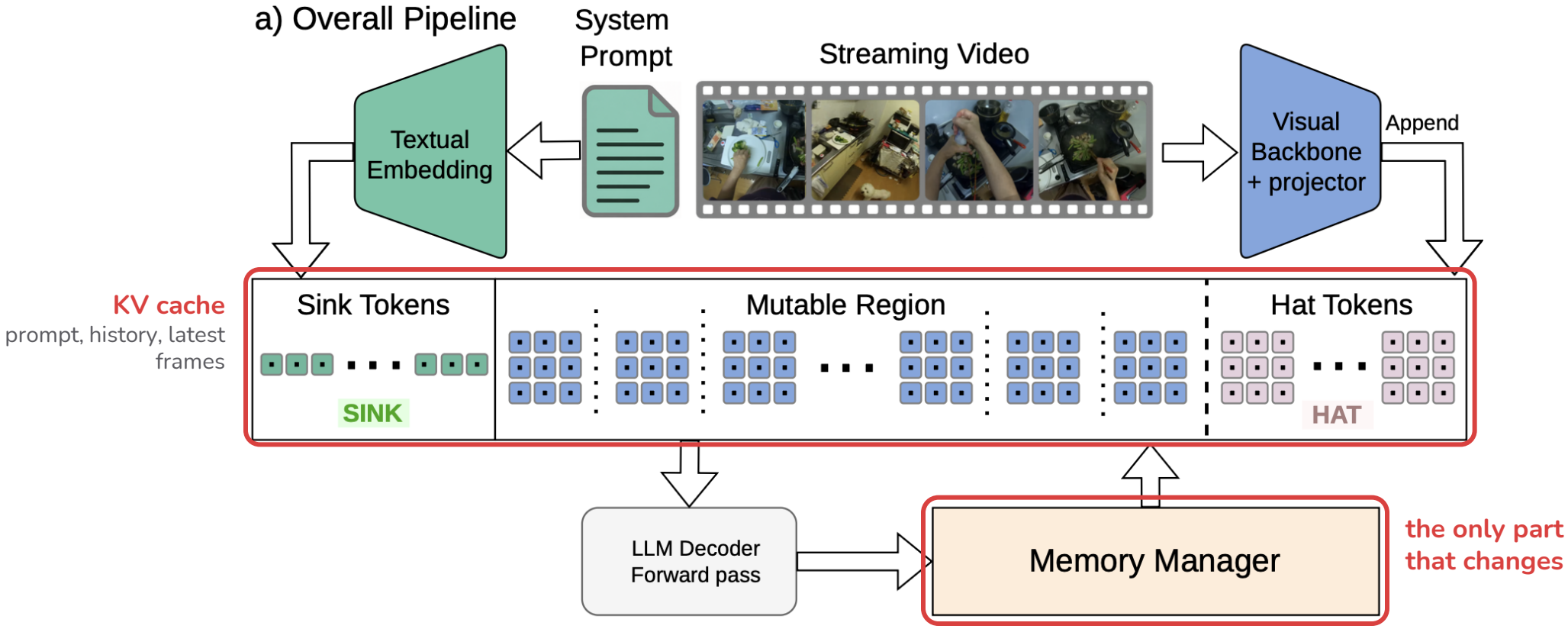
When it is asked · 7 recall regimes

time between the evidence and the question



Boundaries derived from the data: answer validity windows of the sources range from 0 s to 45 h.

A unified streaming testbed



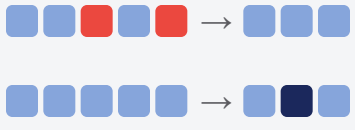
TESTBED

Inside the memory manager

1 Shrink the history

PRUNE drop redundant tokens

MERGE average similar tokens



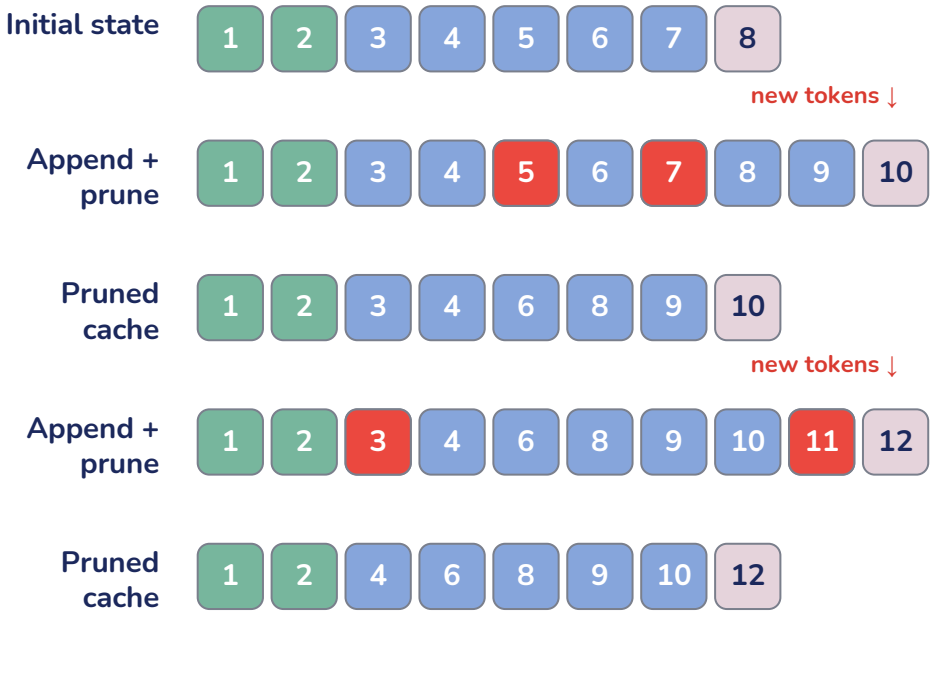
inside a frame or across frames

2 Handle what still overflows

DISCARD the oldest tokens are dropped

OFFLOAD moved to disk, retrieved at question time

optionally quantized to cut storage

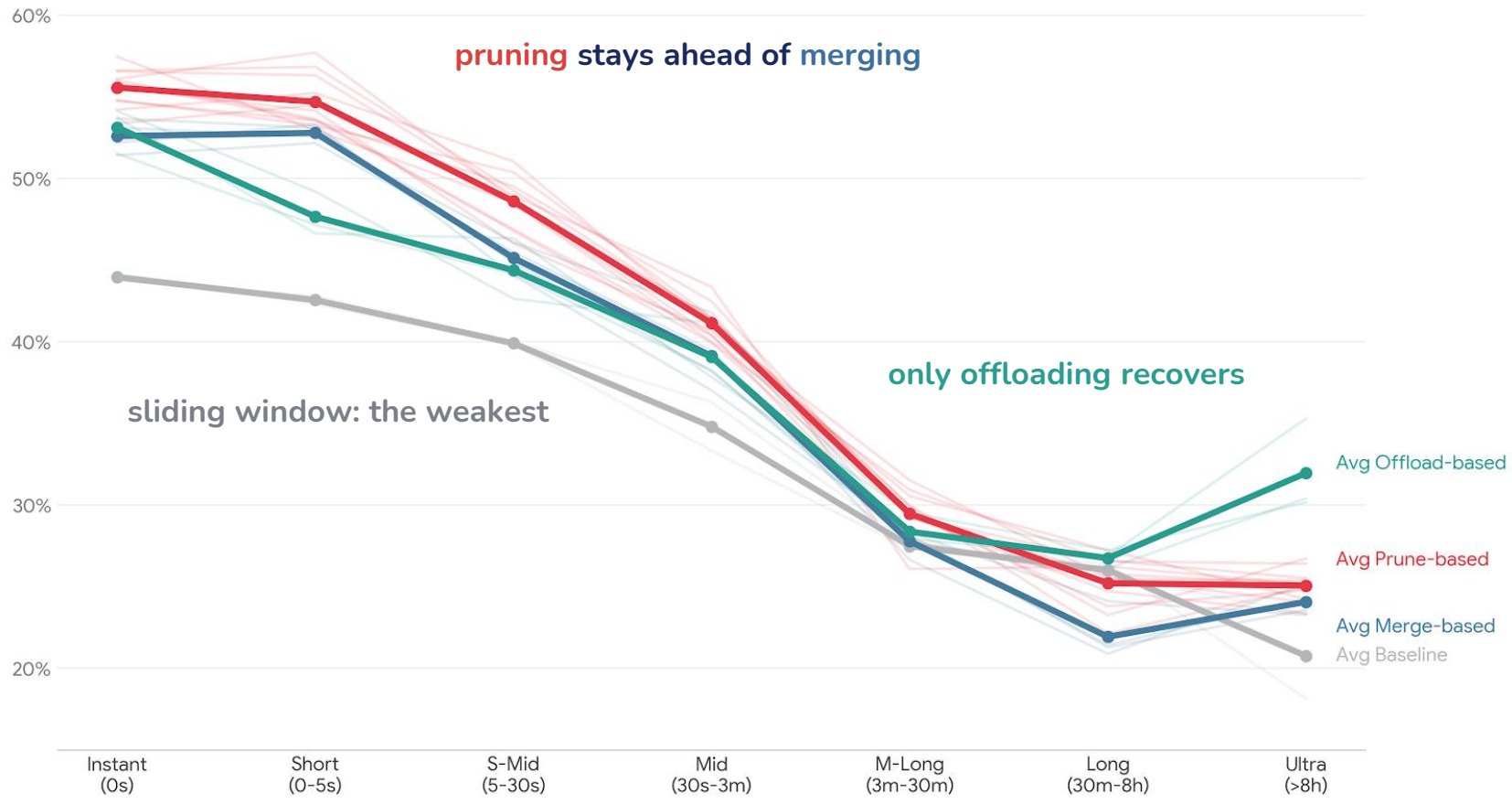


...

Cache update · SINK kept · mutable tokens · pruned on overflow · HAT = latest frames

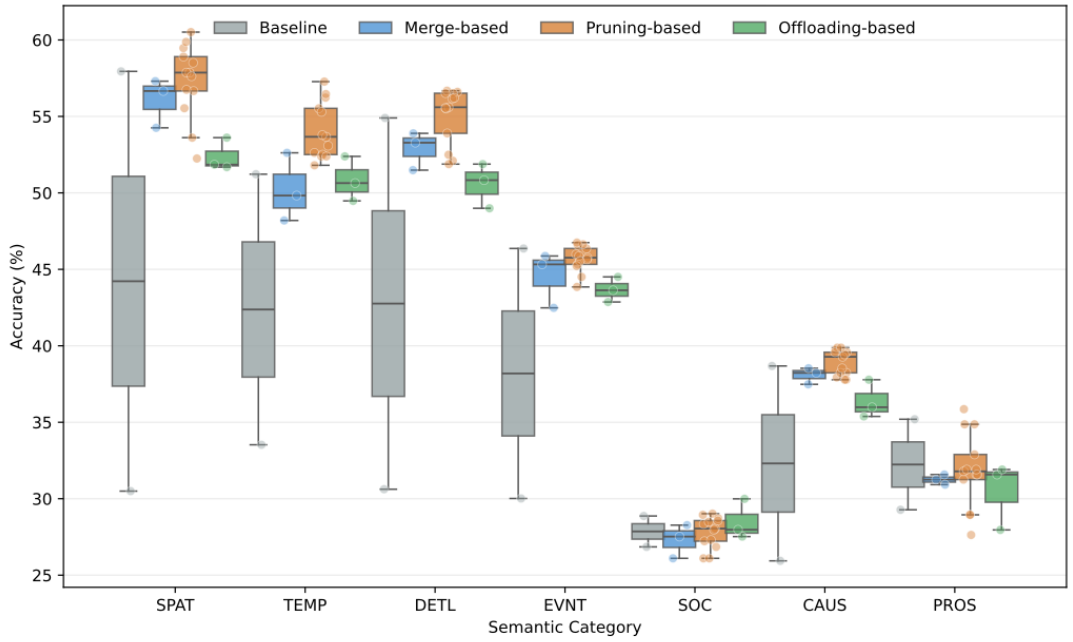
RESULTS

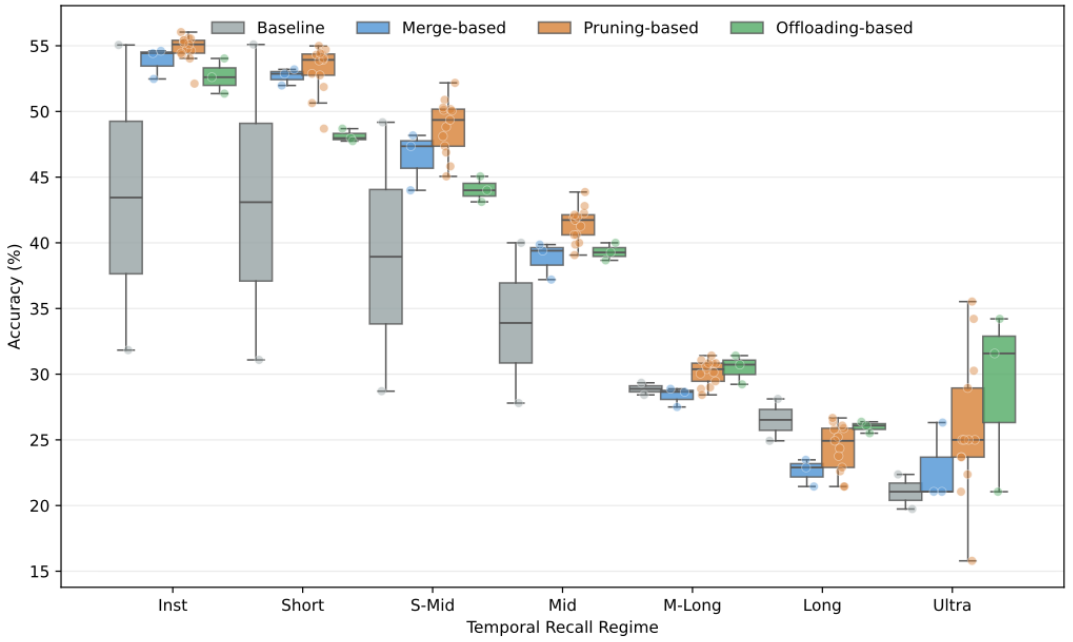
Accuracy decays with the recall horizon



RESULTS

Aggregate scores hide memory profiles





Pruning beats merge and offloading on most memory dimensions

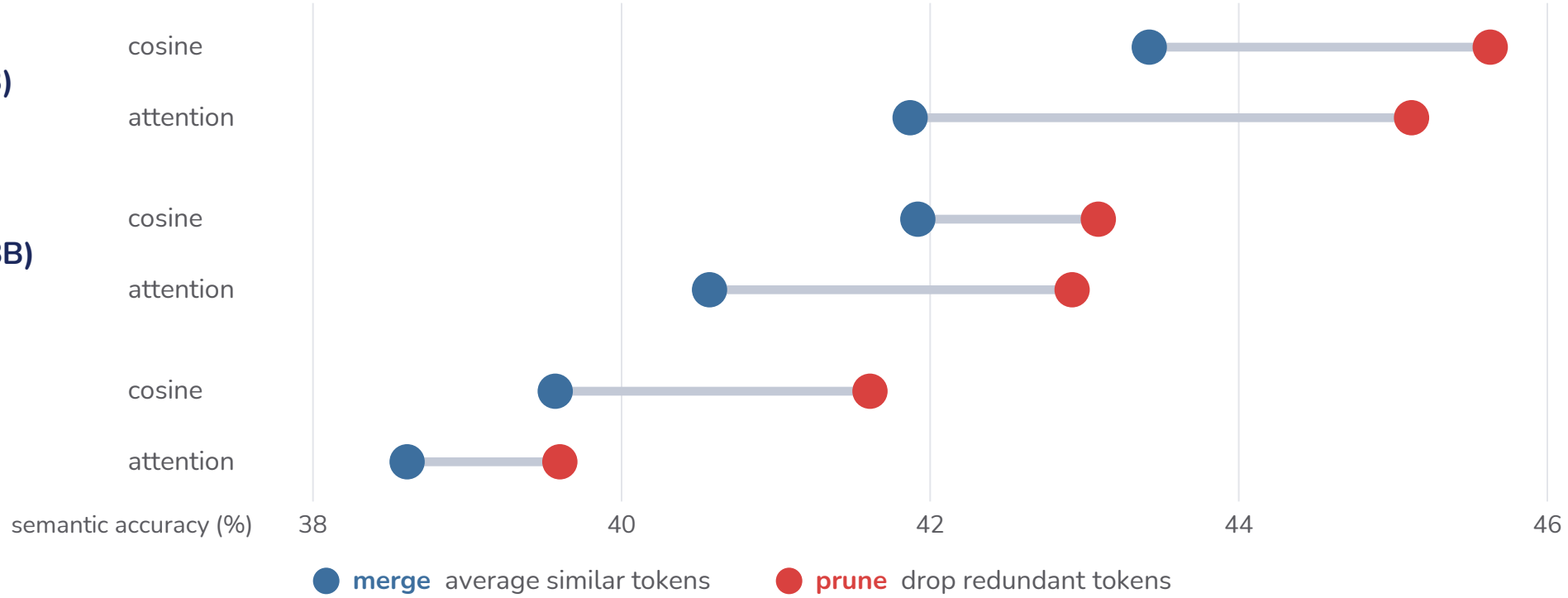
Gap narrows between pruning and merge on social and event memory

Offloading recovers the ultra-long regime and is competitive on social memory

RESULTS · GENERALIZATION

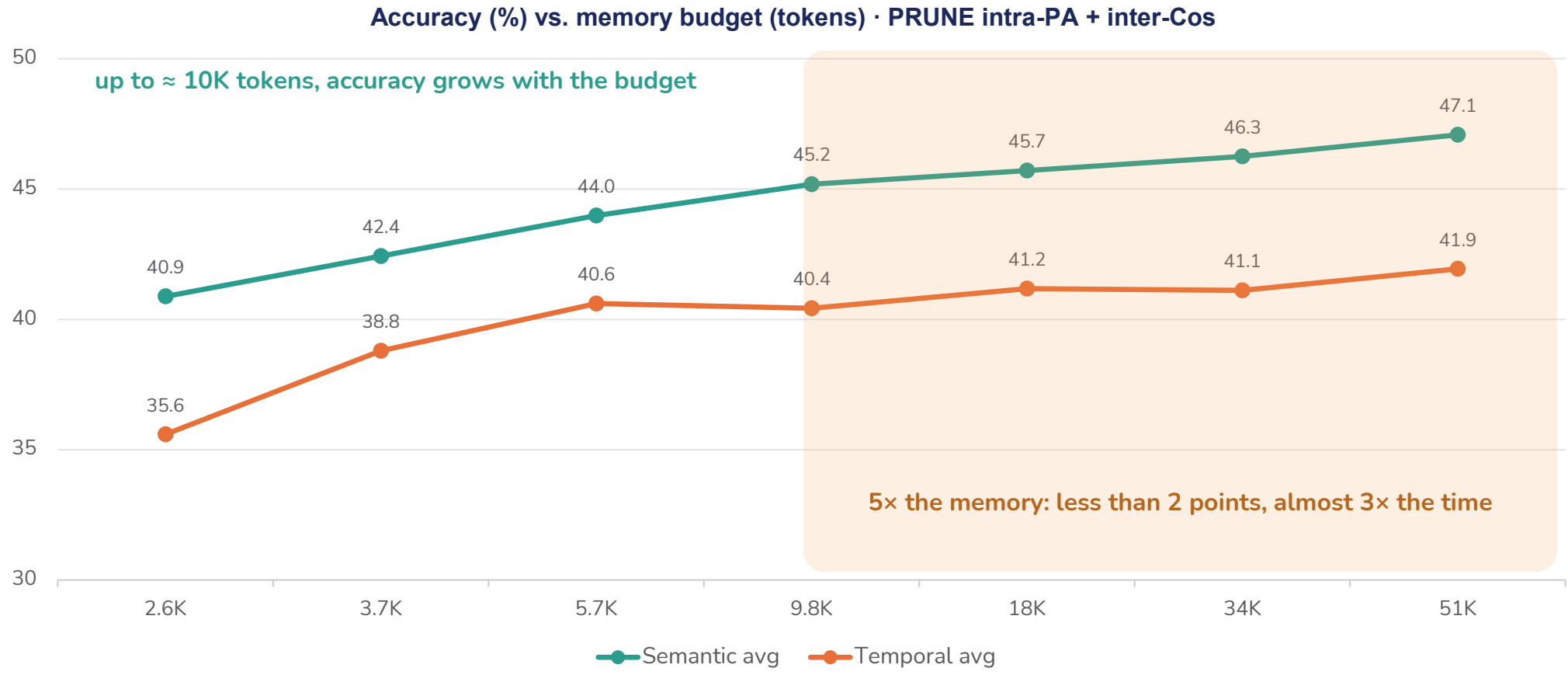
Pruning beats merging on every backbone

Setup we swap the vision encoder and the LLM decoder; budget, protocol, prompts and policies stay the same



Pruning ahead in 6 of 6 matched comparisons · merging averages tokens and blurs fine-grained evidence

More memory helps, with diminishing returns





Thank you!

Scan for paper, data and code!

Rosario Forte · rosario.forte@phd.unict.it
Collaborative AI Research Group · coai-iplab.github.io
coai-iplab.github.io/EgoStream · arXiv:2605.31557