CAME-Bench

Long, interleaved, interference-heavy trajectory generation

1) Construction Procedures

Closed-world environment → Storyboard Planning → Natural Language Trajectory → Pragmatic Refinement

Let's discuss day 1 planning...

I want to learn more about *the first hotel I raised*. I heard people discussing their stunning rating 4.6/5...

...
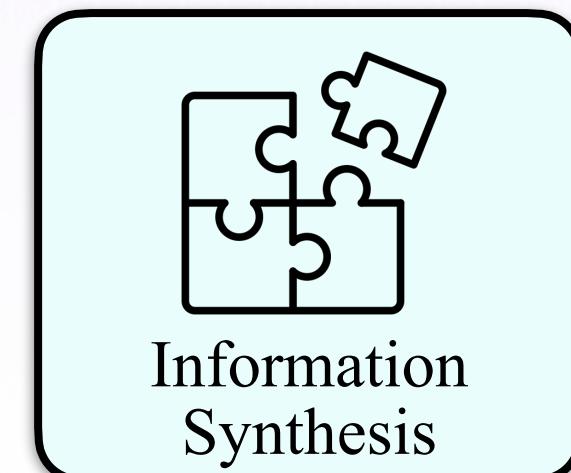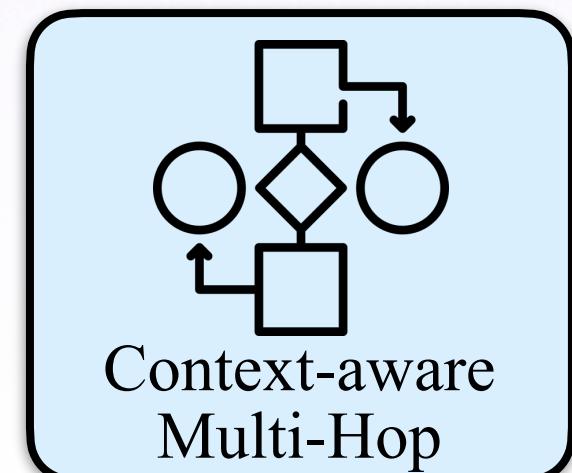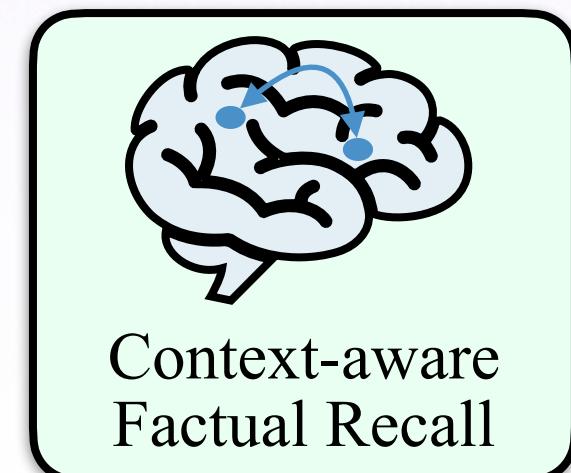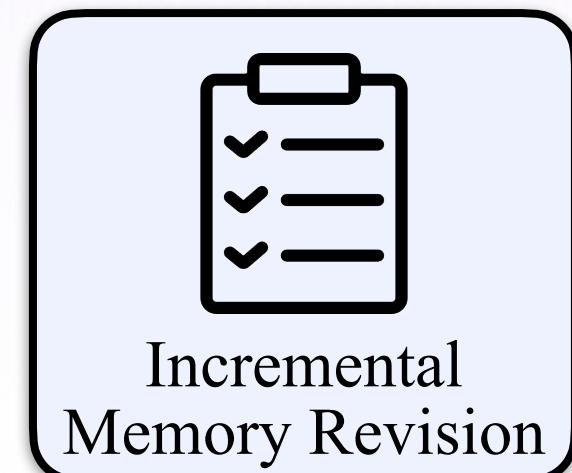
Let's discuss Day 2...

Sounds good! That's a great place for you to relax and enjoy international cuisines...

...

2) Trajectory Overview

A synthetic benchmark characterized by high contextual interference, recurring entities, and interleaved goals that require managing context over long horizons.
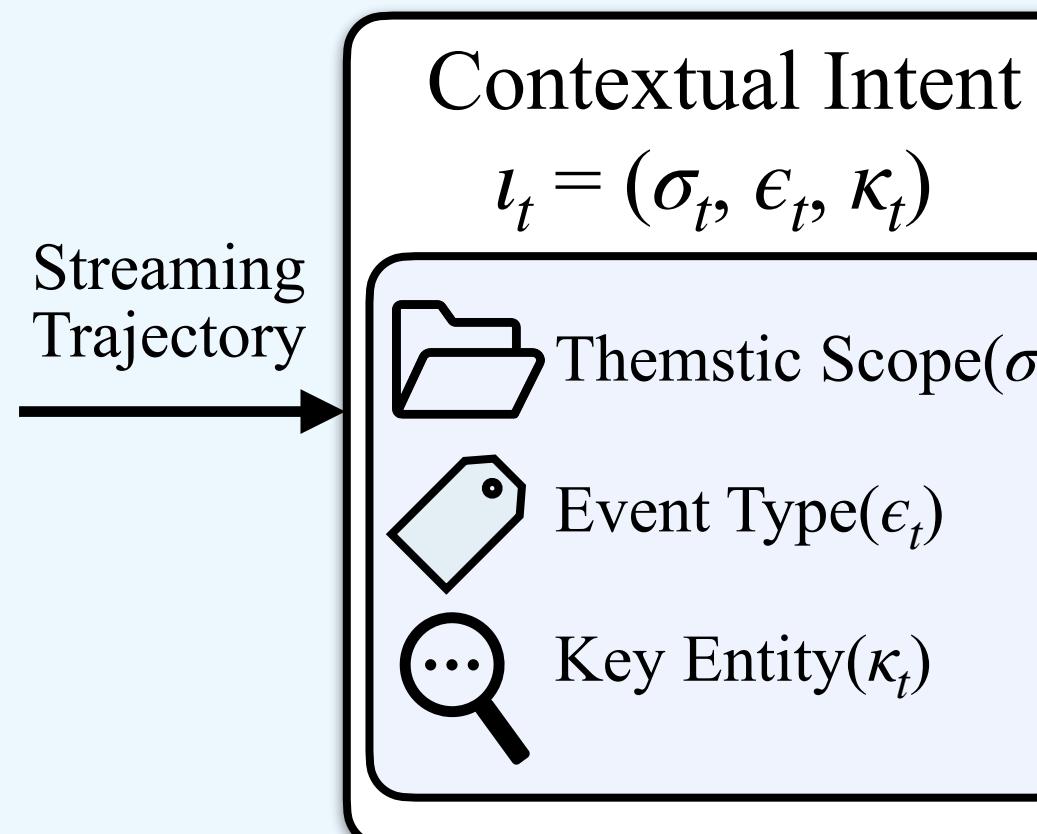
Causal Consistency   Controlled Semantic Interference   Referential Ambiguity

3) Evaluation Overview

Incremental Memory Revision   Context-aware Factual Recall   Context-aware Multi-Hop   Information Synthesis
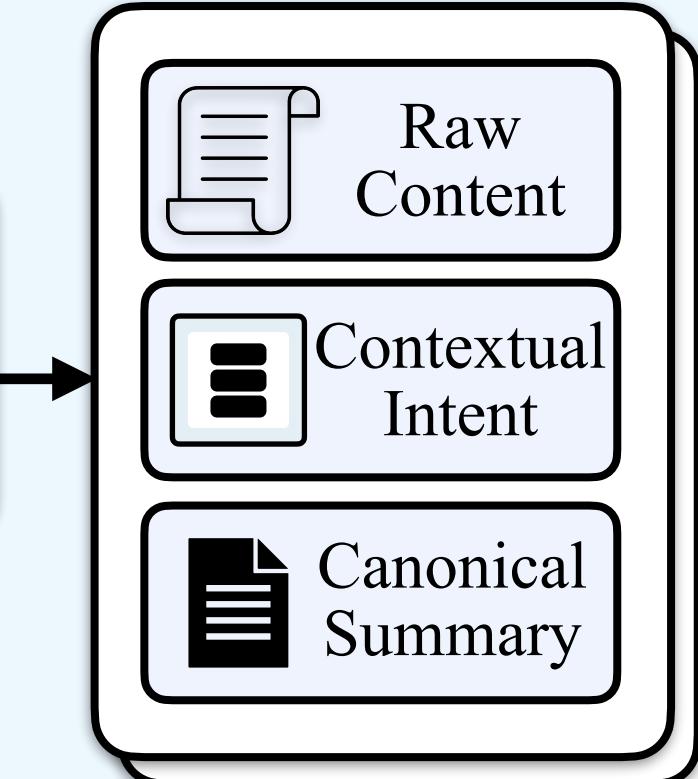
STITCH (Our Method)

Structured intent for context-aware retrieval

1) Intent Construction

Streaming Trajectory →

Contextual Intent
$$\iota_t = (\sigma_t, \epsilon_t, \kappa_t)$$

- Themstic Scope(σ_t)
- Event Type(ϵ_t)
- Key Entity(κ_t)

→ Coreference Resolution →

2) Structured Memory Snippet

- Raw Content
- Contextual Intent
- Canonical Summary

CAME-Bench trajectory + query ⟶ STITCH contextual-intent indexed memory ⟶ context-aware answer

3) Intent-aware Retrieval

Which hotel to book for day 1 →

Structured Filter Configuration(F_q)
- ☒ **Scope:** "Day-1 Itinerary"
- ☐ **Event Type:** N/A
- ☒ **Key Entity Type:** Hotel

→ Label Density Ranking → Final Answer