Learning Safe Robot Planning from Unsafe Experiences: An Episodic Memory Approach for LLM-based Agents

ICLR 2026 Workshop — MemAgents

Hang Zhao · Jing Du · Shengwei An

Northeastern University · Virginia Tech

Problem

Challenge:

- LLM-based robotic agents generate unsafe commands
- Risk harming humans, objects, or environment

Existing approaches (formal verification):

- Require manual specification of every constraint
- Cannot learn from experience
- Do not adapt to emergent hazards

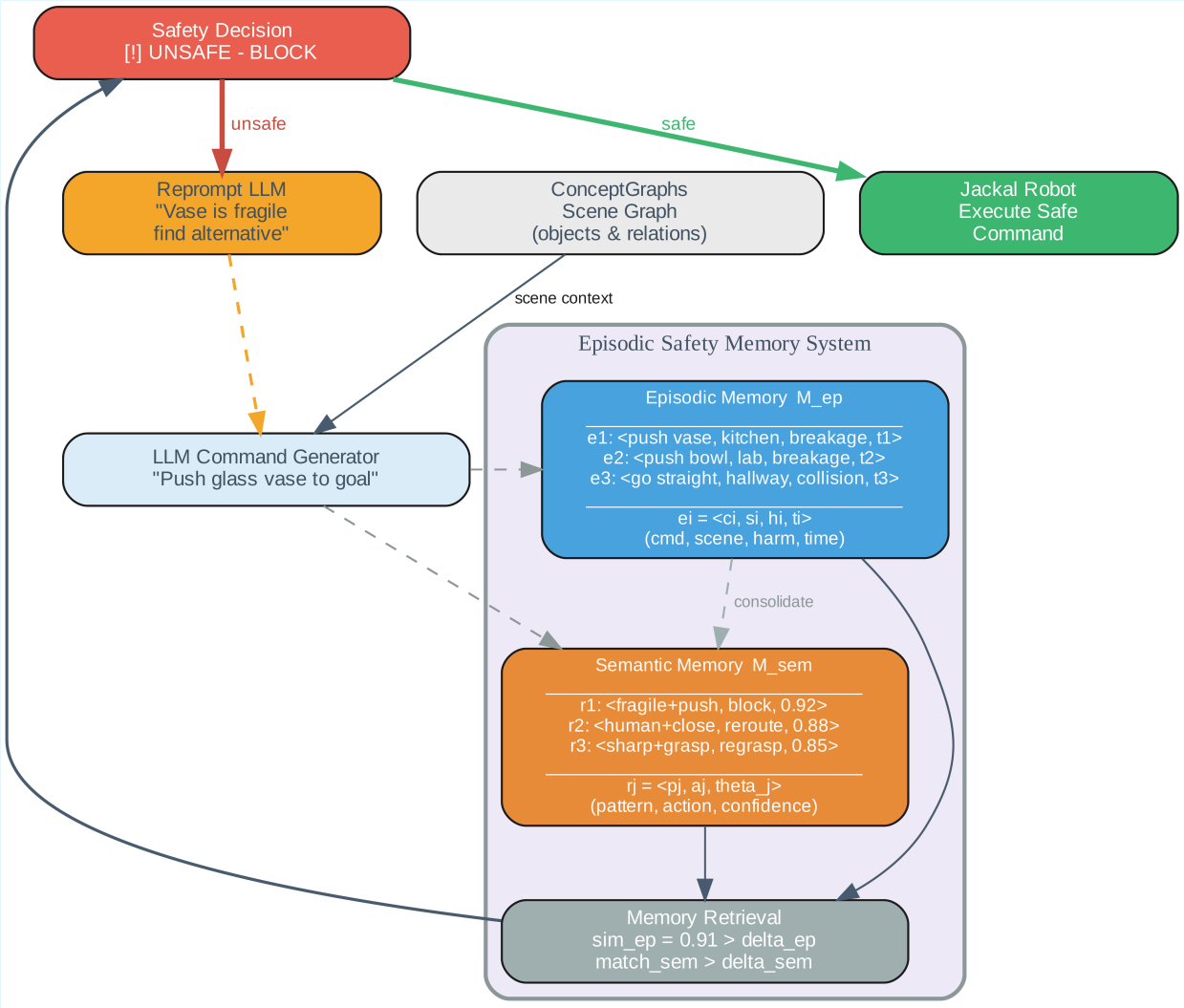
Our Approach

Reframe safety as a memory problem

Episodic Safety Memory:

- Store unsafe instances with consequences
- Retrieve similar cases in real-time
- Consolidate patterns into semantic rules
- Integrate with ConceptGraphs planning

System Architecture



Memory Components

Episodic Memory (M_{ep})	**Semantic Memory (M_{sem})**
Stores specific violation instances	Stores generalized safety rules
e_i = ■*cmd, scene, harm, time*■	r_j = ■*pattern, action, confidence*■
Example: "push vase" → breakage	**Example:** "fragile + push" → block (0.92)
Function: Single-shot learning	**Function:** Pattern generalization

Experimental Results

Setup: 50 scenarios · Jackal robot · ConceptGraphs · 10 seed violations

Method	Safety Rate	Task Success	False Positive
No Safety	52%	98%	0%
Static Rules	78%	89%	11%
Ours	**94%**	**97%**	**3%**

+42% safety improvement over baseline

Key Findings

Memory Dynamics:

- 78 episodes → 12 semantic rules (200 interactions)
- Single-shot learning from one violation
- Generalizes to novel objects

Example Case:

- "push glass vase" → matched (0.91) → blocked
- Reprompted: "navigate around" → safe ✓

Contributions

1. First episodic memory approach for robot safety
2. Dual-memory architecture (episodic + semantic)
3. Real-time filtering via CLIP + graph matching
4. 94% safety with 97% task success
5. Adapts to emergent hazards without manual rules

vs. Formal Verification:

- Learns from experience
- Handles novel hazards

Thank You!

Questions?

zhao.hang1@northeastern.edu