

The Gap

Why prior work falls short

PRIOR WORK

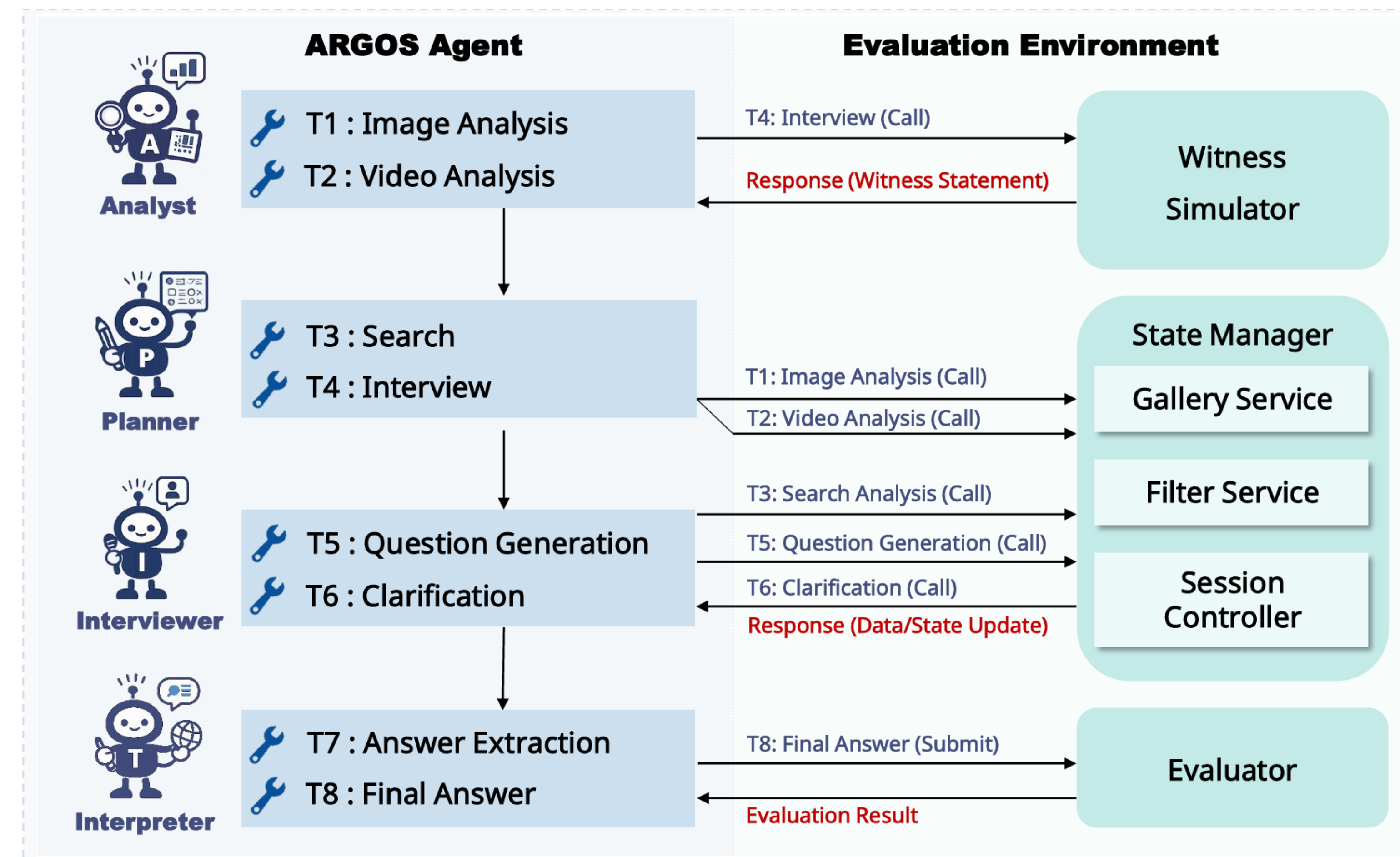
- Perfect Query**
A visual crop or a full text description, given upfront and assumed sufficient.
- One-shot Matching**
A single similarity search over a static gallery. No second chance to ask.
- Appearance Only**
Even multi-modal ReID reasons purely about how the person looked.

ARGOS

- Partial Clues**
Fragments collected over dialogue turns, hedged and sometimes contradictory.
- Interactive Reasoning**
The agent plans, asks, and filters under uncertainty within a turn budget.
- + Space + Time**
Camera topology and empirically validated transit times rule candidates out.

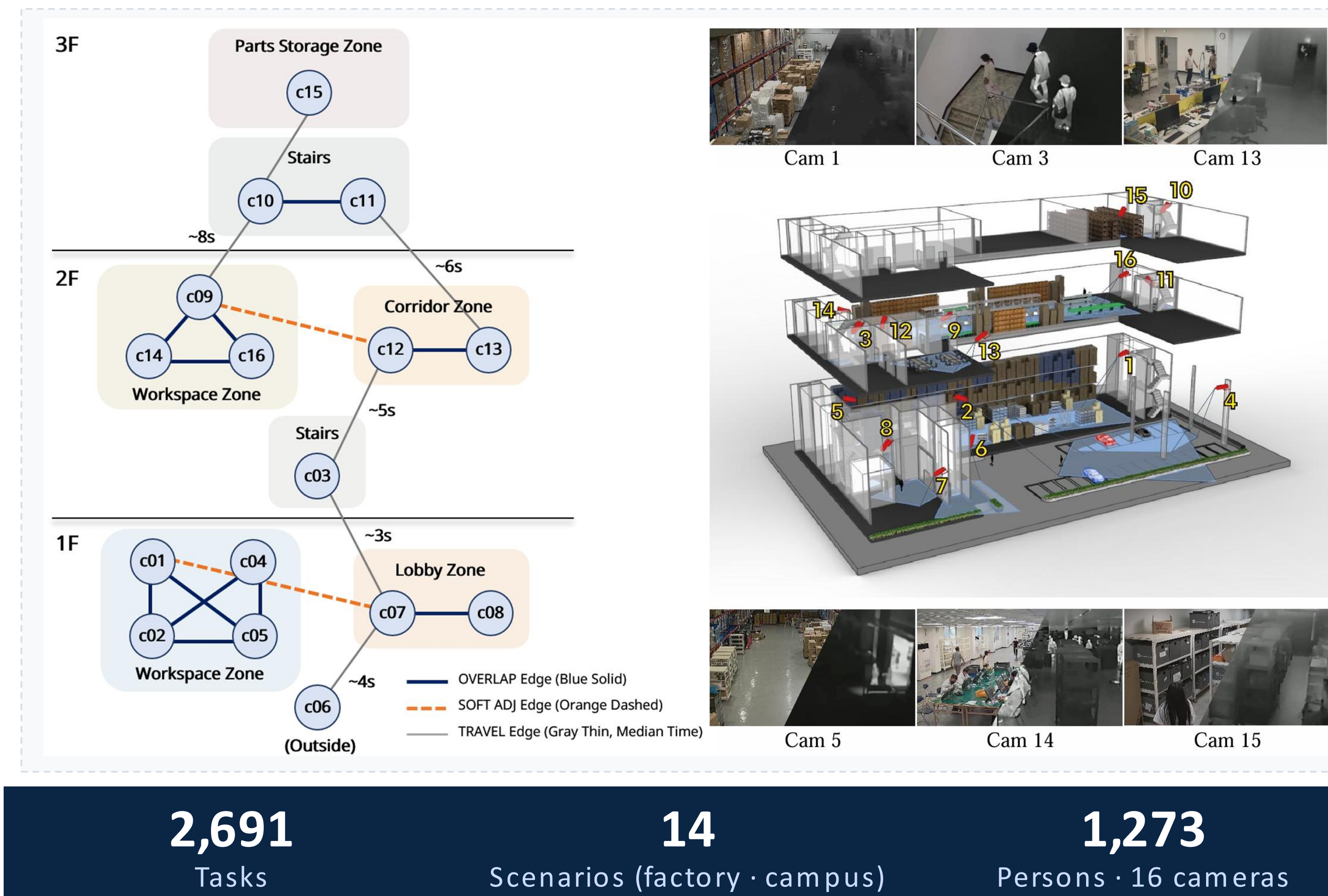
ARGOS Agent

Tool-augmented LLM agent · 4 modules · 8 tools



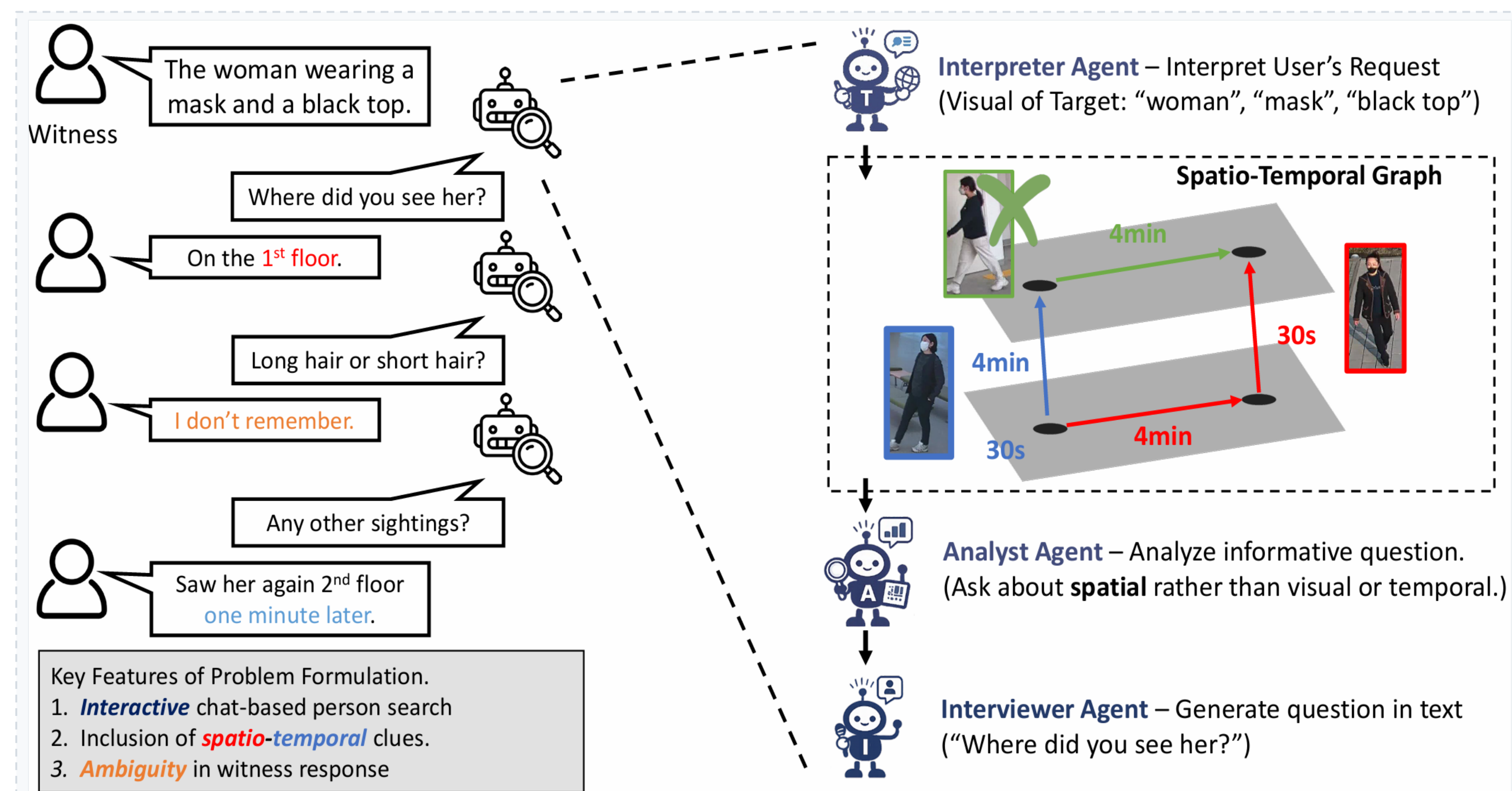
Spatio-Temporal Topology Graph

Camera connectivity × empirically validated transition times



Problem Formulation

An agent that plans, asks, and eliminates — over a vague witness



- Interactive chat-based search
- Spatio-temporal clues integrated
- Ambiguity in witness response

First interactive benchmark coupling witness dialogue with camera-network topology.

Benchmark at a Glance

	TRACK 1 Who	TRACK 2 Where	TRACK 3 When
Task count	989	550	1,152
Interaction	Single-turn	Multi-turn	Multi-turn
Core capability	Semantic parsing	Spatial reasoning	Temporal reasoning
Primary metric	Top-1 Acc	TWS	TWS

Key Findings

Four contributions and the evidence behind them

- New task: interactive person search**
Multi-turn witness dialogue × camera-network topology
- Benchmark + TWS metric**
2,691 tasks · 14 scenarios · 3 tracks; TWS scores correctness × turn efficiency
- STTG grounding**
One temporal check: 19→3 candidates; removing it drops Top-1 by **-49.6 pp**
- Agentic interaction >> direct LLM**
Single-pass 11% → agent 80.6%; w/o Strategy 0.567→0.373, w/o Spatial **-33.8 pp**

Direct LLM is insufficient.

Only tool-augmented agentic interaction grounded in the STTG achieves meaningful performance — the benchmark remains far from solved.

CONTACT gritycda@{kaist.ac.kr · gmail.com}
 LinkedIn in/myungchul-jaden-kim

Experiments — “Far from Solved”

4 LLM backbones · 20-turn budget · Turn-Weighted Success (TWS)

Table 1. Main results — Tracks 2 & 3 (4 backbones).

Backbone	Track 2 (Spatial)				Track 3 (Temporal)			
	TWS	Top-1	SR _t @5	AvgT _s	TWS	Top-1	SR _t @5	AvgT _s
Oracle	1.000	100.0	100.0	2.05	1.000	100.0	100.0	1.88
GPT-5.2	0.338	73.1	33.6	7.04	0.590	88.2	65.8	3.91
GPT-4o	0.323	74.5	31.6	7.49	0.567	80.6	60.8	3.91
GPT-5-mini	0.319	74.9	32.2	7.48	0.556	88.0	60.5	4.40
Claude Sonnet 4	0.383	76.0	38.4	6.71	0.548	83.6	59.5	4.25

SR_t@T: success rate (budget T); AvgT_s: avg. turns (success cases)

Table 3. Ablation (GPT-4o) — removing the temporal tool: **-49.6 pp Top-1**.

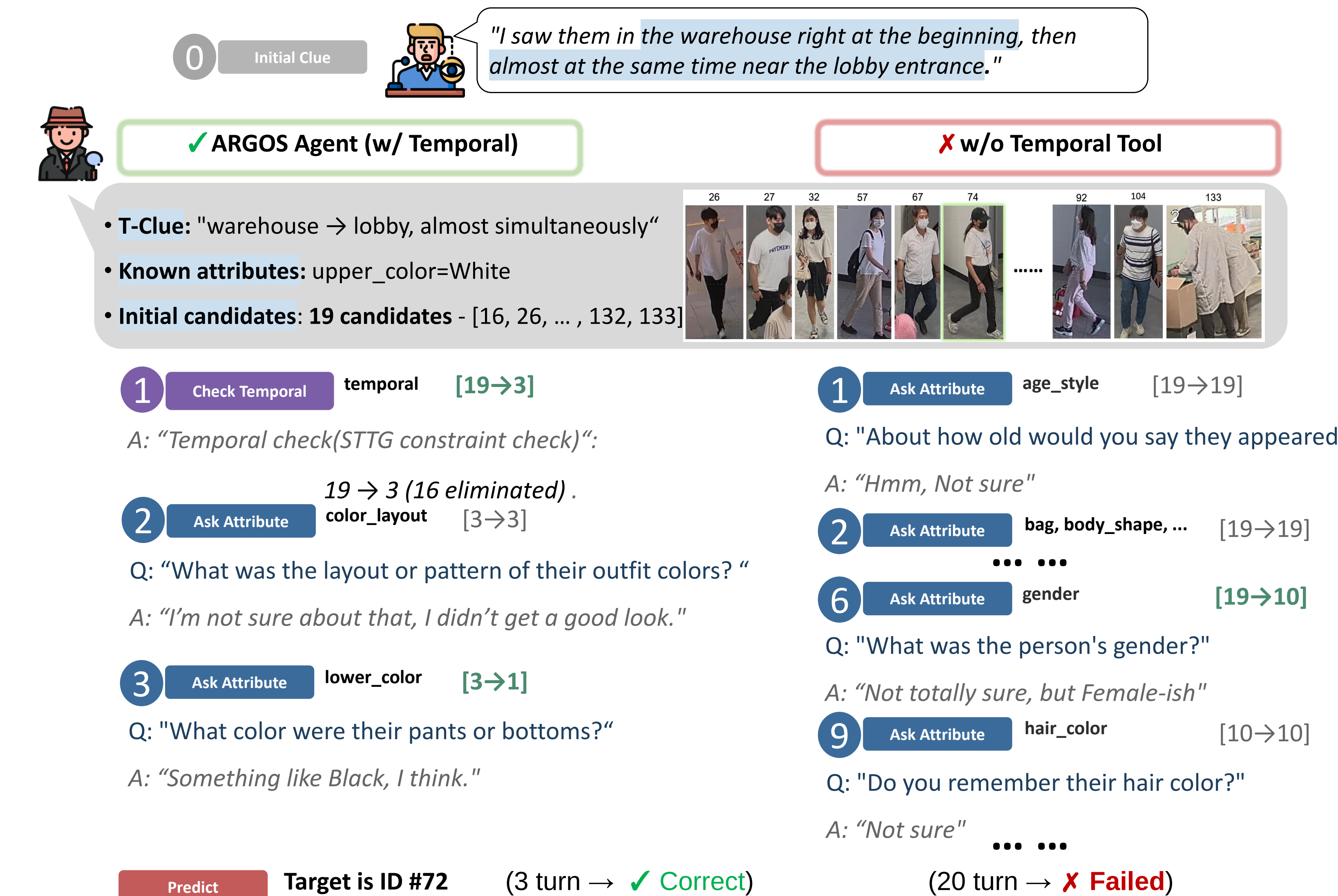
Agent	Removed	Track 2			Track 3		
		TWS	Top-1	AvgT _s	TWS	Top-1	AvgT _s
Oracle	(upper bound)	1.000	100.0	2.05	1.000	100.0	1.88
ARGOS Agent	(none)	0.323	74.5	7.49	0.567	80.6	3.91
w/o Strategy	LLM reasoning	0.136	47.6	10.78	0.373	76.9	7.73
w/o Spatial Tool	Spatial queries	0.063	40.7	14.20	—	—	—
w/o Temporal Tool	Temporal queries	—	—	—	0.054	31.0	13.44
Single-Pass (Full)	Interaction	—	56.2	—	—	29.6	—
Single-Pass (Initial)	Interaction + dialogue	—	64.7	—	—	11.3	—
w/o All Tools	All tools	N/A	—	—	—	11.4	—

ROBUSTNESS & EXTENSIONS

- Autonomous tool use 80.6 → 83.9 Top-1 (GPT-4o)
- Seed stability ± 0.006 TWS across 3 seeds
- Open-weight LLM Qwen2.5-72B: T2 0.26 · T3 0.45
- Witness paraphrasing Pool A ≈ Default

Qualitative Case Study

How a single temporal check solves what 20 turns cannot



Candidate Trajectory

