THE COST OF REASONING OVER FRAMES INSTEAD OF WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing embodied question answering benchmarks evaluate selecting and reading a single frame, analogous to needle-in-a-haystack (NIAH) retrieval. Frontier models excel at this, reflecting their saturation on OpenEQA. We identify an axis of complexity in the scene representation a question requires, defined by how much persistent state must be maintained across frames and how many deductions are needed to integrate observations; existing evaluations occupy its degenerate single-frame case and require neither. We formalize these demands in a compact deductive language whose rules integrate observations across frames into a consistent scene representation, and introduce a novel generative campus environment that poses questions far beyond existing benchmarks along this axis, each certified by a derivation in the language. Frontier Models answer NIAH controls with over 90% accuracy, but fall to 12–41% when evidence must be carried across frames, at derivation depths of at most seven. Results are consistent with models assembling a scene representation ad-hoc within the chain of thought at query time, from the frames themselves, rather than reasoning over one constructed beforehand. Existing benchmarks cannot observe these costs; our axis, language, and environment make them measurable, and motivate evaluations and architectures that treat the representation, rather than the frames, as the object of reasoning.

1 INTRODUCTION

A student approaches an unfamiliar campus building and notices a nameplate: "Hall of Computer Science." Inside, the student carries the name with them, despite it disappearing from view: *I am in the Hall of Computer Science*. By the elevator, a floor indicator reads "3." Climbing two flights of stairs, they update their persistent floor state even in the absence of floor signs: four, then five. Down a hallway, they spot a printer alcove. Although the nameplate, floor indicator, stairwells, and printer were never co-visible and were separated by several minutes of experience, the student binds them: a printer on the fifth floor of the Hall of Computer Science. Later, the student crosses a bridge from a neighbouring building onto a floor with no sign. Recognizing the printer alcove, they know they are on the fifth floor of the Hall of Computer Science again, and bind a second printer in a back office there. When asked "How many printers are on the fifth floor of the Hall of Computer Science?", the student queries their cognitive map, and responds: "two." No single view alone contained the evidence for the answer; only by maintaining a persistent state, independent of any view and updated by each, could the question be answered.

Suppose instead that each printer sat beneath a sign reading "Floor 5" and carried an asset label: "Property of the Hall of Computer Science." Then each printer's floor and building would have been established by a single view, and the question would not have required maintained state to answer. The question, answer, and route remain unchanged, but the distribution of evidence changes how much state (building name, floor number) must persist across observations and how many deductions must integrate it. We study these two demands as an axis of representational complexity (Figure 1) and formalize them in Section 3.3.

We argue that prominent Episodic Memory Embodied QA (EM-EQA) benchmarks evaluate the degenerate case of this axis. They are built on short-horizon environments — single apartments, homes, or individual buildings traversed over minute-scale episodes — characterized by visually unique semantic zones and minimal aliasing. Evidence is locally recoverable. 'What color is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

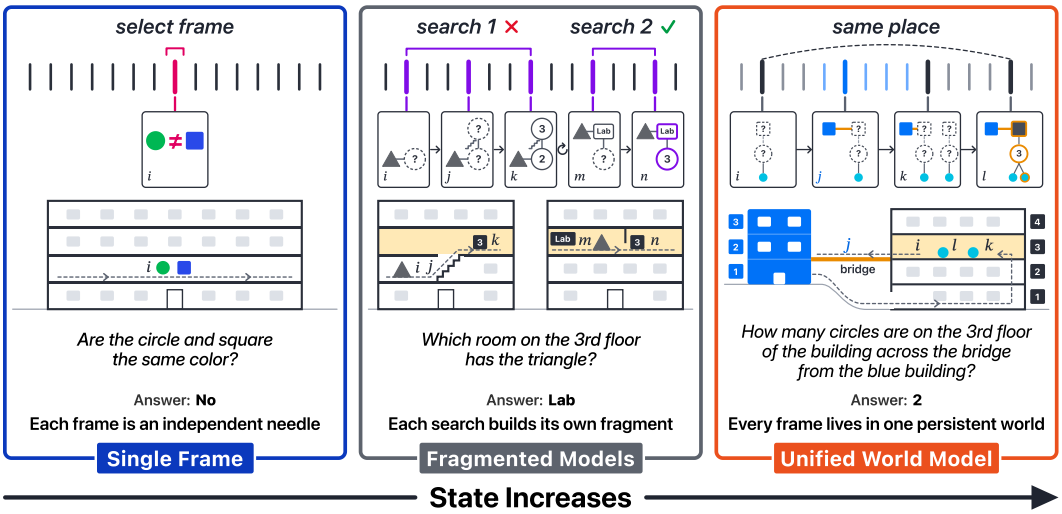

Figure 1: **From frames to worlds.** Three points along an axis of increasing scene representation capability. *Single frame* (left): the agent matches a query to one frame and answers from it, as in needle-in-a-haystack retrieval. *Fragmented models* (middle): the agent indexes into a frame the question hints at, then searches around it, keeping only local state; when a search fails, as the first does here, it starts over from another frame. *Unified world model* (right): the agent builds one coherent representation integrating all its observations before answering, avoiding exhaustive search. Existing embodied QA benchmarks target the first; our questions require the others.

the chair in the bedroom?' requires only retrieving frames containing both a bedroom and chair, then reading the attribute. This resembles Needle-In-A-Haystack (NIAH) retrieval, a task at which frontier models excel. On the HM3D split of OpenEQA, we find that frontier models exceed the published human baseline, and that a single oracle-selected frame per question essentially matches full-episode performance (Section 3.4), suggesting that the benchmark largely rewards selecting and reading evidence available within a single frame.

To fill this gap, we introduce **CampusGen**, which programmatically generates campuses of arbitrarily many buildings and hours-long trajectories through them. Building counts, floor counts, exterior and interior facades, trajectory length and routes, and evidence placement are all configurable. By modulating evidence distribution while preserving questions, answers, and trajectories, CampusGen controls what facts must be carried across observations and how they must be updated or combined. Each answer is then certified by a derivation in **CampusLog**, a compact deductive language, making the required state and integration steps, and thus the question's difficulty, explicit.

On CampusGen, both models nearly saturate needle controls (98% and 91%) but fall to 66% and 37% under *Full* and 41% and 12% under *Reduced*, as mean derivation depth rises from 4.1 to 6.5. Thinking rises with depth, and Gemini's incorrect answers cluster at its output limit. On two real campuses, Gemini 3.1 Pro falls from 62.7% on NIAH questions to 20.5% on questions requiring carried state.

These results are consistent with models assembling the required representation at question time, from the frames themselves (Figure 1, middle): selecting a keyframe the question points to, then searching outward from it while carrying only local state, such as the current building and floor, rather than maintaining one unified world model (Figure 1, right). This strategy's limits are obscured by benchmarks where evidence is locally recoverable, since there one keyframe suffices.

2 Related Work

From retrieval to state maintenance. Frontier models excel at NIAH retrieval (Kamradt, 2023; Gemini Team et al., 2024; Hsieh et al., 2024), with GPT-6-Astra retrieving the right needle near-perfectly at up to a million tokens among seven close distractors (OpenAI, 2026b). They remain

strong but less reliable on long-context multihop reasoning (Hsieh et al., 2024; Kuratov et al., 2024; Modarressi et al., 2025; Li et al., 2025b; OpenAI, 2025), with Claude Mythos 5 scoring 97.5% on one-hop parent lookups and falling to 79.4% on breadth-first search over million-token graphs (Anthropic, 2026). In text, models demonstrate some mechanistic state tracking (Li et al., 2021; 2023; 2025a), but their implicit world models can be incoherent (Vafa et al., 2024) and their accuracy falls with cumulative updates and complexity (Hou et al., 2026; Shojaee et al., 2025), though the collapse at high complexity is partly attributed to output token limits (Lawsen, 2025). In video, VSTAT and SpaMEM find that models succeed at state tracking once the video is translated to text, but fail at grounding that state in the frames (Yu et al., 2026; Liao et al., 2026). By varying trajectory length and congestion over short rendered sequences, Thompson et al. (2025) find that models lose track of which objects are distinct as complexity rises. Accuracy thus falls as questions move from retrieval to state, but video studies of this axis use short clips or a few dozen states of one house.

Embodied and long-video question answering. Embodied question answering (EQA) (Das et al., 2018) is the task of understanding an environment to answer natural language queries. OpenEQA (Majumdar et al., 2024) introduced the first open-vocabulary benchmark, with subsequent work extending it to grounded answers, multi-object reasoning, query ambiguity, metric queries, scene edits and agent actions over space and time (Jiang et al., 2025; Zhai et al., 2025; Cherniavskii et al., 2024; Yang et al., 2025a; Liao et al., 2026; Yadav et al., 2025; Lin et al., 2025). All stay inside one home, apartment or scan, and on OpenEQA one oracle frame matches the full episode (Section 3.4), placing it at the retrieval end of this axis. NaVQA (Anwar et al., 2024) and CityEQA (Zhao et al., 2025) extend to longer campus trajectories and city scale, but evidence remains locally recoverable: NaVQA's metric answers are read from sensor metadata stored with retrieved frames, and CityEQA's agents handed the frame containing the answer outscore its embodied agents. Long-video benchmarks reach hours with one wearer's recordings or spliced clips (Chandrasegaran et al., 2024; Yang et al., 2025b;c; Chen et al., 2026), and EgoSchema and EgoMemReason measure how much video an answer needs (Mangalam et al., 2023; Wang et al., 2026). CampusGen extends procedural generators (Deitke et al., 2022; Raistrick et al., 2024) to multi-story campuses, certifies each answer and its difficulty with a derivation, and modulates evidence placement at a fixed question and camera path, moving the question along this axis (Section 3.3). Note we focus on episodic memory (EM-EQA (Majumdar et al., 2024)) rather than active exploration.

Harnesses, world models, and progress. A *world model* is a coherent, unified representation of an environment's state and rules (Ying et al., 2025). Recent games intuitively demand more capable world models than older benchmarks: ARC-AGI-3 asks agents to infer each game's rules by acting, while Pokémon involves pursuing a goal over thousands of decisions (Ying et al., 2025; ARC Prize Foundation, 2026a; Zanichelli et al., 2026). Harnesses facilitate construction and reuse of an external world model: Codex agents keep a world model as code or text and plan through it (OpenAI, 2026a; Rodionov, 2026a;b), and at the same effort, GPT-6-Astra rises on ARC-AGI-3 from 59.3% under ARC Prize's standard harness to 98.4% under OpenAI's Responses API harness, which preserves its reasoning state between requests (ARC Prize Foundation, 2026b; Kamradt, 2026; OpenAI, 2026b). Progress is fast, with Astra completing Pokémon FireRed in 18 hours where GPT-5.6 Sol needed 96, and, in Codex, both Fallout 3 and Portal (Schreiner, 2026; cozyblaze, 2026). Newer models and harnesses alone outscore existing scene representation methods (Gorlo et al., 2025; Long et al., 2026; Yeo et al., 2026; Wang et al., 2025; Google DeepMind, 2026). This recent jump in performance, and the fact that embodied QA benchmarks test older models, motivate our setup: we evaluate GPT-6-Astra in the Codex CLI and Gemini 3.8 Flash with the episode in context, neither given a prebuilt scene representation (Section 5.1).

3 Representation Complexity in Embodied Question Answering

3.1 The Structure of Embodied Experience

In EM-EQA (Majumdar et al., 2024), a model answers a question Q from an episode history $\mathcal{H} = \{I_t\}_{t=1}^{N}$, the ordered egocentric RGB frames I_t of an embodied agent's trajectory, each showing only part of the environment. Consider going up the stairs: at step t, an agent reads a sign s in a stair landing u: I_t grounds the evidence that s is a floor sign displaying 2 and that u contains it, so $\mathrm{Level}(u) = 2$. It then climbs a flight of stairs. The images that follow show only steps and walls,

until at step t' it reaches the next landing u'. Nothing in $I_{t'}$ says which floor u' is on; its level follows only by carrying $\mathrm{Level}(u) = 2$ from I_t and applying the climb observed in between:

$$\underbrace{s : \mathsf{FloorSign} \ \wedge \ \mathrm{Displays}(s, 2) \ \wedge \ \mathrm{Contains}(u, s)}_{\text{grounded in } I_t} \ \Rightarrow \ \mathrm{Level}(u) = 2,$$

$$\underbrace{\mathrm{Level}(u) = 2}_{\text{carried from } I_t} \ \wedge \ \underbrace{\mathrm{Up}(u, u')}_{\substack{\text{grounded in} \\ I_{t+1}, \ldots, I_{t'}}} \ \Rightarrow \ \mathrm{Level}(u') = 3.$$

These inferences are not complicated; people constantly construct and chain them as they move through the world. Suppose the agent had entered this stairwell without seeing the building's name. Stairwells exist in buildings, so the landing's building has some identity, not yet known:

$$\mathrm{Building}(u) = \beta.$$

Leaving later through the lobby, the agent reads the name on the door. Since they never left the building in between, the name binds β retroactively:

$$\underbrace{\mathrm{Building}(u'') = \text{Hall of CS}}_{\text{grounded in } I_{t''}} \ \wedge \ \underbrace{\mathrm{Building}(u) = \mathrm{Building}(u'')}_{\text{carried through } I_{t'}, \ldots, I_{t''}} \ \Rightarrow \ \beta = \text{Hall of CS}.$$

Exiting the building, crossing an interior bridge, or even passing a sign renaming a section of the building complex in-between would intuitively change this deduction.

3.2 A World Model of That Structure

Component	Purpose	Example
Type signatures	Posit structure not yet observed	$\mathrm{Building} : \mathsf{Room} \to \mathsf{Building}$
Typed grounded facts	Ground the model in frames	$s : \mathsf{FloorSign}$, $\mathrm{Displays}(s, 2)$ in I_t
Inference rules	Derive facts in no single view	$\mathrm{Lvl}(u) = n \wedge \mathrm{Up}(u, u') \Rightarrow \mathrm{Lvl}(u') = n{+}1$
Persistent state	Permit derived facts to be reused and bound elsewhere	Carry $\mathrm{Level}(u) = 2$ from I_t to $I_{t'}$; keep $\mathrm{Building}(u)$ unbound until $I_{t''}$

Table 1: The components of a world model, with examples from Section 3.1. Types and rules are open-ended.

Both examples in Section 3.1 draw on the same four components listed in Table 1.

Type signatures posit structure before it is seen: every room, for example, lies on the floor of some building, which has a name and entrances. A room's parent building exists as an unknown the moment it is seen, the β of Section 3.1, so the name read in the lobby can bind it retroactively.

Typed grounded facts determine which inference rules apply. For example, the same numeral means different things in different contexts: a 2 beside a classroom door probably denotes a room number, while a 2 in a stairwell is likely a floor number, and only the latter sets a floor level.

Inference rules derive facts that **persistent state** then retains between views: a floor's level, a building's name, and whether two visits reached the same place.

Together these form a *world model* (Section 2; Figure 1, right): its rules Θ are the types and inference rules, and its state W is the facts grounded and derived so far, written $\mathcal{M} = (\Theta, W)$. CampusLog supplies Θ, the same for every campus; CampusGen supplies the campus and the episode, whose frames ground the facts in W. Rules are applied to those facts until nothing new follows, and each derived fact keeps the rule and premises that produced it, so every answer comes with a checkable CampusLog derivation, which Section 3.3 measures.

Consider an item that recurs across buildings or rooms whose interiors look alike. With such a world model, asking which one is on floor F of building B becomes a single lookup:

$$\begin{aligned} &\textbf{select} \ \ \mathrm{Room}(x) \\ &\textbf{where} \ \ x : \mathsf{Backpack} \ \wedge \ \mathrm{Level}(\mathrm{Floor}(\mathrm{Room}(x))) = F \\ &\qquad\qquad \wedge \ \mathrm{Name}(\mathrm{Building}(\mathrm{Room}(x))) = B \end{aligned}$$

Without a world model, each candidate must be traced back to its floor and building, rejected, and re-tested per question, each trace a separate fragment (Figure 1, middle). Difficulty compounds when evidence is nonlocal, such as with loop closure, when a place seen before must be recognized.

3.3 Question Difficulty as a Minimal World Model

We define the difficulty of a question Q with respect to an episode $\mathcal{H}$ as the complexity of the minimal world model needed to answer Q from $\mathcal{H}$.

A world model $\mathcal{M} = (\Theta, W)$ answers Q through a derivation D that is sound for the pair $(Q, \mathcal{H})$, that is, one that gives the correct answer to Q from $\mathcal{H}$. We measure the world model's complexity on that derivation, along three axes:

- its **theory**, the number of rules $|\Theta_D|$ it uses;
- its **depth** $\delta(D)$, the length of its longest chain of rule applications, and their total **size** $\sigma(D)$;
- its **span** $\pi(D)$, the number of frames between the earliest and latest frames whose grounded facts the derivation uses: how far the evidence for the answer is spread across the episode.

Collecting these as $c(D) = (|\Theta_D|, \delta, \sigma, \pi)$, compared in some order of priority,

$$d_{\mathcal{H}}(Q) \;=\; \min\big\{\, c(D) \;:\; D \text{ sound for } (Q, \mathcal{H}) \,\big\}.$$

Consider the first printer in Section 1. If the floor and name signs hang in the printer's room, one view grounds everything, and the minimal world model needs only rules that apply each sign to its room (or frame), none relating one view to another. Otherwise, the level and building must be carried throughout the building, adding rules that carry facts across views: a doorway joins two spaces and keeps the floor, a staircase joins two landings and adds one level, and all of these things inherit the exterior building sign.

Episode	Theory	Depth	Size	Span
All evidence with printer	2 (4)	1 (2)	1 (7)	0
Evidence carried to printer	5 (9)	4 (5)	10 (30)	3

Table 2: Difficulty of locating the first printer in two episodes under CampusLog. Absolute counts depend on how a theory factors its rules; what matters is relative magnitude. Parentheses include cost of perception; span is in frames.

In this paper we fix Θ to CampusLog's rules and focus on D, minimizing its depth and then its size, with depth δ as the primary measure.

3.4 Prominent Benchmarks Occupy the Degenerate Case

The first episode of Table 2 has low depth, zero span, and needs no persistent state. Prominent EQA benchmarks reflect this degenerate condition (Table 3). They are built on compact indoor scans — homes, apartments, or single buildings — where relevant facts are often locally recoverable: there is typically one kitchen, one living room, or one relevant functional region, with target objects that are themselves unique in the scene. On OpenEQA's HM3D split, one oracle-selected frame per question matches Gemini's full-episode score (Table 4), so gains there measure selecting and reading a frame, not maintaining a world model.

Larger-scale mapping work moves beyond compact indoor scans, but still does not target the environment–query pairing needed for state maintenance. CODa and NaVQA provide longer robot trajectories, but only through a small number of adjacent buildings, no multi-floor traversal, and little in-situ text for setting semantic state (Anwar et al., 2024). Moreover, the associated NaVQA and OC-NaVQA questions are primarily object-centric, asking where, when, or whether objects and events were observed, rather than making answers depend on memory updated outside the target observation or local cluster (Anwar et al., 2024; Gorlo et al., 2025).

Dataset	Multi-Bldg	Multi-Floor	In-situ Text	Max Path
ScanNet	✗	✗	✗	∼10m
MP3D	✗	✓ (1–3)	✗	∼50m
HM3D	✗	✓ (1–3)	✗	∼20m
Gibson	✗	✓ (1–4)	✗	∼40m
CODa	✓ (1–2)	✗	✗	∼1.5 km
Ours	✓ **(18+)**	✓ **(45+)**	✓	**∼10 km**

Table 3: **Environment comparison** (Dai et al., 2017; Chang et al., 2017; Ramakrishnan et al., 2021; Xia et al., 2018; Anwar et al., 2024). Buildings and floors are what rules for carrying facts act on; in-situ text supplies the facts that anchor a derivation; path length bounds span.

Input	Astra	Gem-3.8-F
Full episode	—	87.84±0.78
50 frames	84.35±0.44	87.82±0.30
Oracle frame	80.67±0.29	87.78±0.45
Blind	45.44±0.39	43.61±1.04
Human†	85.1	

Table 4: **OpenEQA** (HM3D, 557 questions). LLM-judge accuracy of GPT-6 Astra (xhigh) and Gemini 3.8 Flash (High), mean ± sample SD over three runs. †Reported by Majumdar et al. (2024).

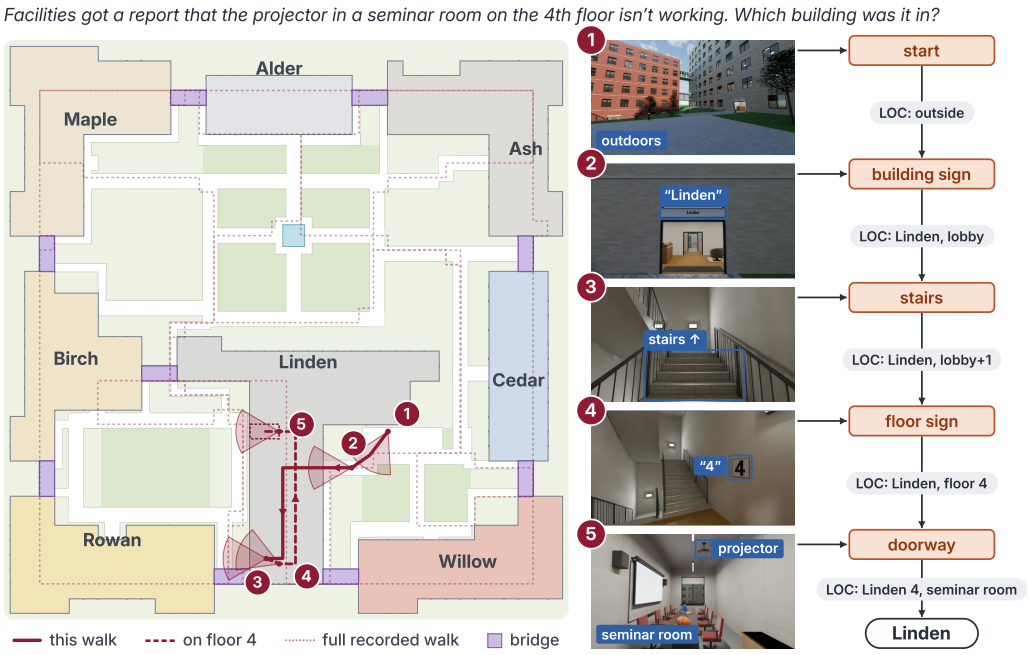

Figure 2: **An example CampusGen episode and derivation.** A campus layout (left), five selected observations (middle), and the corresponding state updates (right). This question is made more difficult by deleting the floor sign and including many projectors in visited seminar rooms.

4 CampusGen

Existing environments lack many buildings and floors, signs, long paths, and distractors (Table 3), allowing one keyframe to answer their questions (Figure 1, left). CampusGen, a Blender-based procedural environment, instantiates the properties existing environment–query pairs lack in Table 3. Difficulty is controlled through how the campus is built (Section 4.1), where evidence is placed (Section 4.2), and how many distractors survive each clue (Section 4.3).

4.1 Building a Campus

As shown in Figure 2, a campus physically consists of multi-floor buildings containing hallways connecting rooms of 28 types. Buildings can be directly connected by interior bridges, and are

Sign	Where	States	Reduced
Entrance nameplate	entrances	building name	kept
Building manifest	lobby	(sub)department;	kept
Floor sign	stair landings	floor number	ground floor only
Bridge panel	bridge ends	far building's name	off
Room plaque	hallways	type, name	kept
		department, subdepartment	off

Table 5: **The evidence CampusGen toggles.**

indirectly connected by exterior paths. Each building is assigned a department, such as "Physics"; each floor a subdepartment, such as "Geometry".

Buildings stand apart, on hills or on flat ground, so their ground floors can sit at different heights. Appearance is configurable: each building can have its own interior and exterior facades or share them, and room types permit randomized furnishings.

CampusGen searches over campus layouts, room contents, and routes for a campus that permits questions with the requested derivation dificulty and distractor properties. One continuous walk through hallways, stairs, bridges, paths, and rooms, rendered in Blender once per condition. It can be arbitrarily long; ours are 10,000–16,000 frames.

4.2 Evidence

Table 5 lists the signs CampusGen places; their placement determines how far each fact must be carried, and so each question's depth and span. A sign beside the target makes a question degenerate, answerable from one frame (Figure 1, left). A condition only turns signs off, keeping the campus, route, questions, and answers fixed; under *Reduced* (Table 5, last column), floors, buildings, and subdepartments must be carried further.

4.3 Questions and Distractors

Each question is a query in the language of Table 6, such as the one in Section 3.2; it asks for a property of the one item that satisfies all its clues. For each clue, the episode includes distractors that satisfy every other clue but have a different answer, such that a model that loses one fact proceeds down a "rabbit hole" search, as in Figure 1 (middle). We define chance as a uniform guess among a question's plausible answers: the distinct values of the requested type that the episode shows within the question's scope. We require at least four, so chance is at most 25%. CampusLog certifies every question under each condition (Section 3.3).

select	colour / room / building / department / floor / subdepartment
where	*item* $\wedge$ *clue* $\wedge \cdots$
item	backpack / bottle / mug / ...
clue	colour = green / building = Linden / department = Mathematics / floor = 4 / subdepartment = Geometry / room type = classroom

Table 6: **CampusGen's formal question vocabulary.**

5 Experiments

5.1 Setup

Campuses and episodes. We generate three campuses with CampusGen, each with 8–12 buildings, rendered under *Full* and *Reduced* (Table 5). Episodes run at one frame per second: 10,000–16,000 frames, or 2.8–4.4 hours. Questions are sampled from each campus's question bank and are identical across conditions: NIAH controls, answerable from one frame, and primary questions, which must be derived across frames. We run only three episodes, since more would be cost-prohibitive given

the model class, long-context evaluation, and high thinking levels; caching each episode's context lets us ask many questions of it.

Questions are sampled from each campus's question bank and are identical across conditions: in total, 40 needle-in-a-haystack (NIAH) controls, answerable from one frame, and 94 hard questions, which must be derived across frames.

Models. GPT-6-Astra in the Codex harness with xhigh thinking, and Gemini 3.8 Flash with maximum thinking and an explicit 65.5k output limit. An LLM judge, GPT-6-Astra with low thinking, scores answers following OpenEQA (Majumdar et al., 2024).

Chance. A uniform guess among a question's plausible answers, at most 25% (Section 4.3).

5.2 Removing Evidence Raises Difficulty and Lowers Accuracy

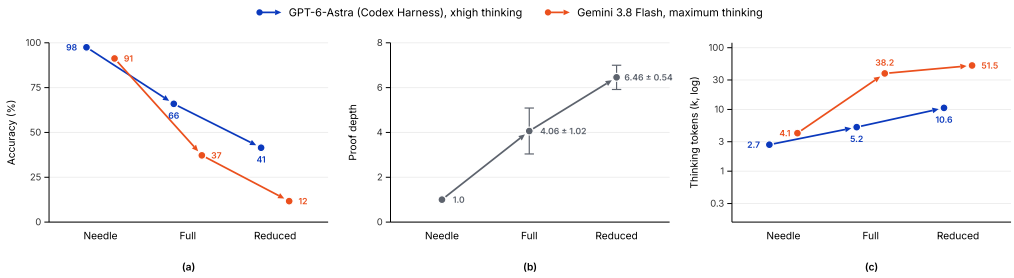

Figure 3: **Accuracy falls as depth rises.** (a) Accuracy, (b) CampusLog derivation depth and (c) mean thinking tokens (log scale) on needle-in-a-haystack controls and on the hard questions under *Full* and *Reduced*, pooled over three campuses (94 hard and 40 needle questions per condition). Depth is mean ± SD; needle questions are plotted at depth 1.

Both models nearly saturate the needle controls (Figure 3a): 98% for GPT-6-Astra and 91% for Gemini 3.8 Flash. On the same campuses and routes, accuracy falls to 66% and 37% under *Full* and to 41% and 12% under *Reduced*, as mean depth rises from 4.06 to 6.46 (Figure 3b). Thinking tokens rise with depth for both models (Figure 3c).

5.3 Gemini Fails Closer to Its Thinking Limit

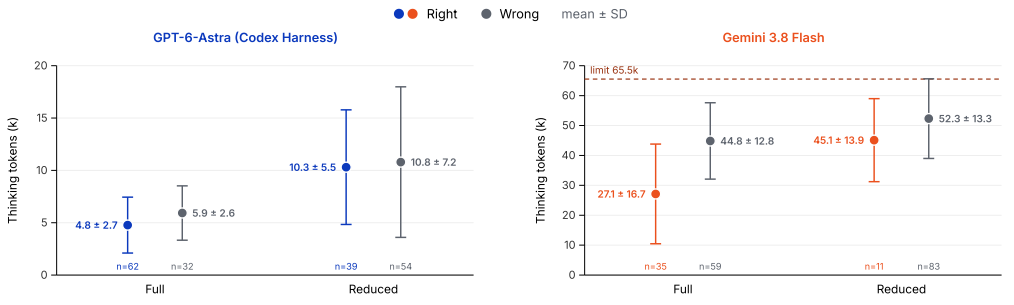

Figure 4: **Thinking tokens on correct and incorrect answers** (mean ± SD), for GPT-6-Astra (left) and Gemini 3.8 Flash (right), under *Full* and *Reduced*. The dashed line marks Gemini's 65.5k output limit.

For GPT-6-Astra, thinking is similar on correct and incorrect answers (Figure 4). For Gemini 3.8 Flash, incorrect answers average 44.8k tokens under *Full* and 52.3k under *Reduced*, against a 65.5k limit. Under *Full*, it is right on 93% of answers that think for under 20k tokens and on 15% of those at the limit; under *Reduced*, on 25% and 5%.

5.4 Real Campuses

We additionally evaluate Gemini-3.1-Pro-Preview on two real, anonymized campuses (Table 7), with frames sampled at 0.4 FPS and the whole episode, about 900k tokens, in context; results are averaged over five runs. Accuracy falls from 62.7% on NIAH questions to 20.5% on State Required questions, which must carry state across frames.

	Duration	**Buildings**	**Frames**	**NIAH**	**State Required**
Anon A	2:03:36	18	2,966	96	38
Anon B	2:12:00	21	3,168	103	37
Total	4:15:36	39	6,134	199	75
Accuracy (%)				62.7	20.5
Tokens, correct				470	2,702
Tokens, incorrect				961	2,114

Table 7: **Real campuses.** Questions per campus, and Gemini-3.1-Pro-Preview's accuracy and mean tokens with the whole episode in context, averaged over both campuses and five runs.

6 Conclusion

Prominent embodied QA benchmarks occupy the degenerate end of an axis set by how much state must persist across frames and how many deductions must integrate it. There, one keyframe supplies the evidence for the question, where answering reduces to needle-in-a-haystack retrieval, which frontier models nearly saturate: 98% and 91% on our needle controls.

Moving along that axis, retrieval evolves into state maintenance. **CampusLog** formalizes these demands, and **CampusGen** moves questions along the axis by removing evidence while keeping campus, route, questions, and answers fixed. The same models fall to 66% and 37% when facts must be carried across frames, and to 41% and 12% when signs are removed, at derivation depths of at most seven; on two real campuses, Gemini-3.1-Pro-Preview falls from 62.7% on NIAH questions to 20.5% on State Required questions.

These results are consistent with models assembling a representation from the frames at question time rather than reasoning over one built beforehand, a cost that single-frame benchmarks cannot observe. Because CampusGen can add buildings, lengthen routes, and remove evidence at will, it can scale with models, and we hope it motivates evaluations and architectures that treat the representation, rather than the frames, as the object of reasoning.

Ethics statement

Real-campus episodes were recorded in public spaces and anonymized; no identifiable individuals appear. Generated campuses contain no real people.

AI use statement

We used AI coding assistants (Claude Code with Claude Fable 5, Claude Fable 5.1, and Claude Opus 5.5; and Codex with GPT-6-Astra) to produce code for this work, including implementing our methods, experimental pipelines, and evaluation scripts. The authors manually verified and reviewed all AI-generated code and thoroughly checked outputs for correctness. We also used AI tools (Claude Code and Claude Design) to aid with designing figures and refining the wording of the paper; all figures and text were reviewed by the authors.

References

Anthropic. System Card: Claude Fable 5 & Claude Mythos 5. System card, 2026. URL https://anthropic.com/claude-fable-5-mythos-5-system-card. June 9, 2026. Accessed 2026-09-25.

Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. *arXiv preprint arXiv:2409.13682*, 2024. URL https://arxiv.org/abs/2409.13682.

ARC Prize Foundation. ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence. *arXiv preprint arXiv:2603.24621*, 2026a. URL https://arxiv.org/abs/2603.24621.

ARC Prize Foundation. OpenAI GPT-6 Astra: ARC-AGI-3 Results. Evaluation results, 2026b. URL https://arcprize.org/results/openai-gpt-6-astra. Accessed 2026-09-25.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-Hour Video-Language Understanding. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2411.04998.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017.

Guo Chen, Lidong Lu, Yicheng Liu, Liangrui Dong, Lidong Zou, Jixin Lv, Zhenquan Li, Xinyi Mao, Baoqi Pei, Shihao Wang, Zhiqi Li, Karan Sapra, Fuxiao Liu, Yin-Dong Zheng, Yifei Huang, Limin Wang, Zhiding Yu, Andrew Tao, Guilin Liu, and Tong Lu. Towards Multimodal Lifelong Understanding: A Dataset and Agentic Baseline. *arXiv preprint arXiv:2603.05484*, 2026. URL https://arxiv.org/abs/2603.05484.

Daniil Cherniavskii, Phillip Lippe, Andrii Zadaianchuk, and Efstratios Gavves. Benchmark for temporal, ambiguous, and grounded embodied question-answering, 2024. URL https://openreview.net/forum?id=toqQYz2N2X. Preprint; withdrawn submission to ICLR 2025.

cozyblaze. Portal agent. GitHub repository, 2026. URL https://github.com/cozyblaze/portal-agent. Portal run completed September 5, 2026. Accessed 2026-09-25.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, pp. 1–10, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Das_Embodied_Question_Answering_CVPR_2018_paper.html.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. *arXiv preprint arXiv:2206.06994*, 2022. URL https://arxiv.org/abs/2206.06994.

Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Google DeepMind. Gemini 3.8 Flash Model Card. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. September 2026. Accessed 2026-09-25.

Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Describe anything anywhere at any moment, 2025. URL https://arxiv.org/abs/2512.00565.

Dengzhe Hou, Lingyu Jiang, Deng Li, Zirui Li, Fangzhou Lin, and Kazunori D Yamada. WMF-AM: Probing LLM Working Memory via Depth-Parameterized Cumulative State Tracking. *arXiv preprint arXiv:2603.27343*, 2026. URL https://arxiv.org/abs/2603.27343.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.06654.

Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering. *arXiv preprint arXiv:2503.11117*, 2025.

Greg Kamradt. Needle In A Haystack. GitHub repository, 2023. URL https://github.com/gkamradt/needle-in-a-haystack. Accessed 2026-09-25.

Greg Kamradt. OpenAI's GPT-6 Astra on ARC-AGI-3. ARC Prize, 2026. URL https://arcprize.org/blog/astra. September 3, 2026. Accessed 2026-09-25.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.10149.

A. Lawsen. Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv preprint arXiv:2506.09250*, 2025. URL https://arxiv.org/abs/2506.09250.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit Representations of Meaning in Neural Language Models. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2106.00737.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (How) Do Language Models Track State? *arXiv preprint arXiv:2503.02854*, 2025a. URL https://arxiv.org/abs/2503.02854.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Miaoyu Li, Qin Chao, and Boyang Li. Two Causally Related Needles in a Video Haystack. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.19853.

Chih-Ting Liao, Xi Xiao, Chunlei Meng, Zhangquan Chen, Yitong Qiao, Weilin Zhou, Tianyang Wang, Xu Zheng, and Xin Cao. SpaMEM: Benchmarking Dynamic Spatial Reasoning via Perception-Memory Integration in Embodied Environments. *arXiv preprint arXiv:2604.22409*, 2026. URL https://arxiv.org/abs/2604.22409.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-Bench: Evaluating the Capabilities of MLLMs in Online Spatio-temporal Scene Understanding. *arXiv preprint arXiv:2507.07984*, 2025. URL https://arxiv.org/abs/2507.07984.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PMz29A7Muq.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Majumdar_OpenEQA_Embodied_Question_Answering_in_the_Era_of_Foundation_Models_CVPR_2024_paper.html.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. *arXiv preprint arXiv:2308.09126*, 2023. URL https://arxiv.org/abs/2308.09126.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.05167.

OpenAI. GraphWalks. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/openai/graphwalks. Accessed 2026-09-25.

OpenAI. Codex. Software, GitHub repository, 2026a. URL https://github.com/openai/codex. Accessed 2026-09-25.

OpenAI. GPT-6 Astra: A new generation of intelligence. OpenAI, 2026b. URL https://openai.com/index/gpt-6-astra/. Accessed 2026-09-25.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2406.11824.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv:2109.08238*, 2021.

Sergey Rodionov. Executable World Models for ARC-AGI-3 in the Era of Coding Agents. *arXiv preprint arXiv:2605.05138*, 2026a. doi: 10.1007/978-3-032-33195-3_15. URL https://arxiv.org/abs/2605.05138.

Sergey Rodionov. Do Coding Agents Need Executable World Models, Simplification, and Verification to Solve ARC-AGI-3? *arXiv preprint arXiv:2607.15439*, 2026b. URL https://arxiv.org/abs/2607.15439.

Maximilian Schreiner. GPT-6 Astra crushes Pokemon, Factorio, and Fallout 3 then spirals into Minecraft potato farming after one bad Creeper. The Decoder, 2026. URL https://the-decoder.com/gpt-6-astra-pokemon-champion-in-18-hours-potato-farmer-after-one-creeper-mishap/. September 17, 2026. Accessed 2026-09-25.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.06941.

Jacob Thompson, Emiliano Garcia-Lopez, and Yonatan Bisk. REM: Evaluating LLM Embodied Spatial Reasoning through Multi-Frame Trajectories. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2512.00736.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the World Model Implicit in a Generative Model. *arXiv preprint arXiv:2406.03689*, 2024. URL https://arxiv.org/abs/2406.03689.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_LVBench_An_Extreme_Long_Video_Understanding_Benchmark_ICCV_2025_paper.html.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. EgoMemReason: A Memory-Driven Reasoning Benchmark for Long-Horizon Egocentric Video Understanding. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.09874.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. FindingDory: A Benchmark to Evaluate Memory in Embodied Agents. *arXiv preprint arXiv:2506.15635*, 2025. URL https://arxiv.org/abs/2506.15635.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. *arXiv preprint arXiv:2503.03803*, 2025b. URL https://arxiv.org/abs/2503.03803.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. *arXiv preprint arXiv:2511.04670*, 2025c. URL https://arxiv.org/abs/2511.04670.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic Multimodal Memory Agent for Long Video Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yeo_WorldMM_Dynamic_Multimodal_Memory_Agent_for_Long_Video_Reasoning_CVPR_2026_paper.html.

Lance Ying, Katherine M. Collins, Prafull Sharma, Cedric Colas, Kaiya Ivy Zhao, Adrian Weller, Zenna Tavares, Phillip Isola, Samuel J. Gershman, Jacob D. Andreas, Thomas L. Griffiths, Francois Chollet, Kelsey R. Allen, and Joshua B. Tenenbaum. Assessing Adaptive World Models in Machines with Novel Games. *arXiv preprint arXiv:2507.12821*, 2025. URL https://arxiv.org/abs/2507.12821.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, and Saining Xie. Benchmarking Visual State Tracking in Multimodal Video Understanding. *arXiv preprint arXiv:2606.03920*, 2026. URL https://arxiv.org/abs/2606.03920.

Niccolò Zanichelli, Gavin Leech, and Peli Grietzer. I want to be the very best. Paradigm, 2026. URL https://www.paradigm3.org/research/pokemon. September 23, 2026. Accessed 2026-09-25.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-Centric Embodied Question Answering. *arXiv preprint arXiv:2505.13948*, 2025.

Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. CityEQA: A Hierarchical LLM Agent on Embodied Question Answering Benchmark in City Space. *arXiv preprint arXiv:2502.12532*, 2025. URL https://arxiv.org/abs/2502.12532.

A CAMPUSGEN DETAILS

Table 8 gives CampusGen's defaults, all configurable, and Table 9 its 28 room types. Table 10 shows where each fact comes from in each condition.

Setting	Default
Buildings	6–10 floors; a main hallway of 16, 24 or 32 m lined with rooms, and side wings
Stairs	one core, two in the longest buildings, with a landing and a hallway door on every floor
Entrances	three on the ground floor: one through the lobby and two at hallway ends
Campus	outdoor paths join all entrances; bridges join upper floors at equal height, giving each bridged building 1–3 neighbours
Floor numbers	each building counts up from a ground-floor label of 1–4
Labels	unique building, café and room names (25% of eligible rooms); departments may repeat, and residences have none

Table 8: **Campus defaults.**

Size (m)	Room types
4×4	single office, shared office, kitchenette, clinic†
4×4 or 4×8	restroom, storage
4×4 or 8×4	dorm room
4×4, 8×4 or 8×8	residence laundry
8×4 or 4×8	seminar room†, study room†
8×8	lobby, open computer lab†, woodshop†, library reading room†, dorm suite, server room, mail room, mechanical room, locker room
8×8 or 12×12	classroom†, lab†, lounge†
12×12	café, lecture hall†, computer lab†, machine shop†, library stacks†, gym

Table 9: **Room types.** Sizes are width along the hallway by depth. †Eligible for a name.

Fact	Full	Reduced
Floor	its landing's floor sign	the ground floor's sign, counted up the flights
Building	nameplate or bridge panel	nameplate
Department	room plaque	manifest header
Subdepartment	room plaque	manifest, once the building and floor are known

Table 10: **Where each fact comes from**, before doors and flights carry it to the target. Both conditions also recognize rooms by unique names.

Routes and observations. Routes are sampled to meet quotas on buildings and floors visited, arrivals by stairs and by bridge, and returns to a floor by another path. Frames are 1 m apart with a 120° field of view; in each room the camera looks left, centre and right. An item counts as seen in a frame if its mesh is visible above a minimum size, and a sign as read if its text is in view and unoccluded; these measurements, not a recognizer, give CampusLog its facts.

Questions. Queries (Table 6) are rendered in English, e.g. "At one point you saw a backpack in a classroom. Which building and which floor was that classroom on?" A question is kept only if:

- exactly one item on the campus satisfies every clue;
- every clue matters: dropping it admits an item with a different answer, which the route shows if the question names a place;
- CampusLog derives its clues and answer from the route in every condition;
- a room number comes with its building, since numbers repeat across buildings.

The bank caps questions per target and spreads them across answer types, clue types and buildings. Needle-in-a-haystack questions (Section 5.1) have depth at most 1.

Chance. Section 4.3 defines chance over a question's plausible answers. If the question names a place, these are the distinct answers of the target and of every other item of its kind the route shows, of its colour if one is given; otherwise, they are every value of the requested type the route shows. The campus must also hold at least four values of the requested type, overall and within the places the clues name (e.g. four subdepartments in the named department), so the layout alone does not give the answer away.