EVALUATING TEMPORAL STATE REASONING IN EMBODIED AGENTS UNDER NOISY OBSERVATION STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must reason over observation streams produced by perception modules, action logs, trackers, task monitors, and simulators. These streams are often noisy, delayed, out of order, incomplete, and contradictory, making it challenging to maintain temporally consistent beliefs about the world: whether an object was in a given state at a specific time, what information was available before late evidence arrived, which observations support or contradict a belief, and whether task goals hold. Existing embodied benchmarks primarily evaluate task execution, navigation, question answering, or memory recall, but do not isolate temporal state reasoning from imperfect evidence streams. We present EviStateBench, a benchmark for evaluating temporal task-state reasoning in embodied agents. EviStateBench constructs hidden state timelines from simulator truth and task specifications, releases sanitized observation streams as public inputs, and evaluates standardized predictions against hidden answer sets. The benchmark covers bitemporal reasoning, evidence attribution, state changes, uncertainty, precondition checking, failure localization, and goal-state queries. The released artifact contains 570 simulator-grounded episodes from 36 activities, 5,848 hard queries, 2.86M clean observations, and 2.80M hard mixed observations. We also provide EviStateDB, a reference temporal-state maintenance system that tracks valid time, transaction time, and supporting or contradicting evidence. Experiments with retrieval-memory, temporal-log, and EviStateDB baselines show that EviStateBench distinguishes retrieval, temporal lookup, and explicit state-maintenance strategies, with exact accuracy ranging from 25.2% to 75.6%.

1 INTRODUCTION

Embodied agents increasingly operate over observation streams rather than clean symbolic states. A robot may receive detector outputs, action logs, task monitors, and simulator feedback that describe the world at different times and with different reliability. Observations can be delayed, missing, out of order, or contradictory: a later visual observation may revise an earlier relation estimate, while a task executor still needs to determine what state held, what was known at a given time, which evidence supports an answer, and whether task goals are satisfied. This creates a fundamental challenge for embodied intelligence: maintaining temporally consistent task-state beliefs from imperfect evidence streams.

Existing embodied benchmarks evaluate important capabilities including household task execution Srivastava et al. (2022); Li et al. (2024), navigation and rearrangement Savva et al. (2019); Szot et al. (2021), visual interaction Kolve et al. (2017), program-based activities Puig et al. (2018), dialogue-based interaction Padmakumar et al. (2022), episodic question answering Majumdar et al. (2024), and instruction following Shridhar et al. (2020). However, these benchmarks primarily measure task success, perception, planning, or memory recall. They do not isolate whether an agent can maintain a temporally grounded state representation from noisy and incomplete observations.

We introduce EviStateBench, a benchmark for evaluating temporal state reasoning in embodied agents. The benchmark formulates this challenge as maintaining queryable task-state views over imperfect observation streams. Unlike simple retrieval from episodic memory, the evaluated sys-

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.

tems must answer temporal and evidence-aware questions: whether a state held at a valid time, what information was available before delayed evidence arrived, which observations support or contradict a belief, what changed during an episode, and whether task-level conditions are satisfied. These requirements connect to temporal data management through valid-time and transaction-time semantics Snodgrass (1999); Jensen & Snodgrass (1999), stream processing under continuous updates Babcock et al. (2002), incremental view maintenance Gupta & Mumick (1995); Zhuge et al. (1995), and provenance-aware reasoning under uncertainty Aggarwal & Yu (2009); Green et al. (2007); Cheney et al. (2009).

As shown in Figure 1, each EviStateBench instance provides public task specifications, observation streams, queries, and metadata while withholding hidden state timelines and answer sets derived from simulator truth. Systems may use arbitrary internal mechanisms, including retrieval memory, temporal logs, learned memory modules, rules, database systems, or hybrid approaches. They are compared only through a shared QueryAnswer interface and hidden evaluation protocol, preventing access to oracle states while enabling reproducible comparison across different reasoning strategies. EviStateBench contains 570 simulator-grounded episodes from 36 activities, 5,848 hard queries, 2.86M clean observations, and 2.80M hard mixed observations. It evaluates bitemporal reasoning, evidence attribution, state changes, uncertainty handling, precondition reasoning, failure localization, and goal-state queries. Experiments with retrieval-memory, temporal-log, and EviStateDB baselines show that EviStateBench separates retrieval, temporal lookup, and explicit state-maintenance strategies, with exact accuracy ranging from 25.2% to 75.6%. The remaining errors concentrate on structured, set-valued, and task-derived queries, revealing open challenges for reliable temporal reasoning in embodied agents.

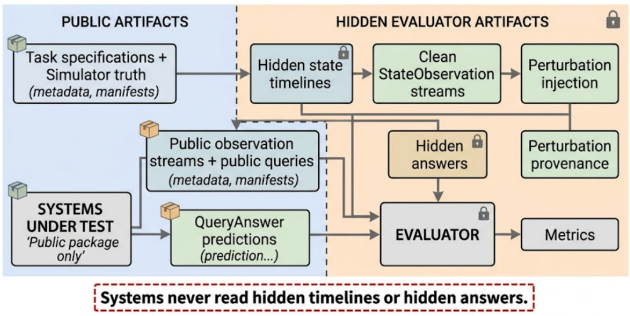

Figure 1: Overview of EviStateBench and pipeline.

We additionally provide EviStateDB, a reference temporal-state maintenance system for EviStateBench. EviStateDB maintains valid-time intervals, transaction-time revisions, supporting and contradicting evidence, confidence, and task-derived views. It serves as an interpretable baseline rather than an oracle: hidden answers are generated independently from simulator truth and task specifications, allowing EviStateBench to evaluate EviStateDB alongside future learned or hybrid systems. This paper makes four contributions.

- We introduce EviStateBench, a benchmark for evaluating temporal state reasoning in embodied agents under noisy, delayed, incomplete, and contradictory observation streams.
- We define a public-input/hidden-oracle evaluation protocol with StateObservation streams, QueryAnswer outputs, temporal semantics, evidence tracking, and diagnostic query workloads.
- We provide a simulator-grounded generation pipeline that creates reproducible episodes, observation streams, perturbations, hidden answer sets, and validation artifacts.
- We establish reference baselines and analyses that characterize the capabilities and limitations of retrieval-based, temporal-log, and explicit state-maintenance approaches.

2 Motivation and Benchmark Scope

Why Temporal State Reasoning Matters. Embodied agents rarely receive a clean symbolic description of the world. Instead, they observe streams produced by perception models, action logs, task monitors, and simulators. These observations may disagree, arrive late, or provide incomplete evidence. Consider a household task where an agent places a cup into a cabinet and closes the door. A detector may report that the cup is inside the cabinet, an action log may indicate that the placement succeeded, while a later observation may revise the earlier belief. A reliable agent must distinguish

Table 1: Existing embodied benchmarks evaluate task execution, simulation, instruction following, question answering, or memory recall. EviStateBench focuses on temporal state reasoning from imperfect observation streams.

Benchmark family	Primary target	Temporal state queries	Event/ arrival time	Perturbed observations	Task-derived views	Hidden oracle
BEHAVIOR / BEHAVIOR-1K	Household task execution and simulation	Partial	–	–	Partial	Simulator truth
ALFRED	Instruction following	–	–	–	Task success	Demo labels
OpenEQA	Embodied question answering	Episodic context	–	–	NL answers	Human/LLM scoring
Retrieval memory evals	Recall from stored observations	Limited	–	Usually no	–	Dataset labels
EviStateBench	Temporal state reasoning from observation streams	✓	✓	✓	✓	Simulator truth + specs

Note: ✓ indicates that the property is explicitly evaluated as a first-class scoring target; "Partial" indicates that the property may appear in task definitions or simulator state but is not the primary evaluation target.

what actually happened, what evidence was available at a given time, and what state can be inferred from the accumulated observations. This distinction separates three concepts: the *world state*, the *observation evidence*, and the *maintained task-state belief*. The world state describes what holds at a valid time; observations provide partial and potentially unreliable evidence; the maintained state is the agent's answerable representation after processing evidence. EviStateBench evaluates this final capability by comparing system predictions against hidden answers derived from simulator truth while exposing only observation streams as input.

Benchmark Formulation. EviStateBench evaluates systems that infer temporally grounded task states from imperfect observation streams. Each instance provides task specifications, observations, and queries. A system outputs standardized QueryAnswer predictions, without requiring any specific internal representation. Systems may use retrieval memory, temporal logs, learned memory modules, rules, database representations, or hybrid approaches. The benchmark covers eight query families. `CHECK_STATE` and `CHECK_GOAL` evaluate state and goal reasoning. `AS_OF_STATE` tests whether systems respect information availability by separating valid time from arrival time. `WHY_STATE` evaluates evidence attribution. `STATE_DIFF`, `FIND_UNCERTAIN_STATES`, `CHECK_PRECONDITION`, and `FAILURE_LOCALIZATION` evaluate structured reasoning over changes, uncertainty, task constraints, and failures.

Scope Boundary. EviStateBench is designed to isolate temporal state reasoning rather than full embodied control. It does not directly evaluate manipulation policies, exploration strategies, perception accuracy, or language-grounded action execution. Including these factors in the primary score would make failures difficult to attribute because performance differences could originate from perception, control, simulation, or planning rather than state reasoning. Therefore, EviStateBench uses simulator-grounded hidden timelines and controlled observation streams. All systems receive the same public evidence and query workload, and differences reflect how they interpret, update, and reason over observations.

Benchmark Design Principles. The benchmark is designed around five principles. (i) *Temporal semantics.* The benchmark evaluates current and historical states rather than only final task success. (ii) *Information availability.* The benchmark separates event time from arrival time so that systems cannot use future evidence when answering historical queries. (iii) *Evidence awareness.* The benchmark represents supporting and contradicting observations to evaluate reasoning under uncertainty and conflict. (iv) *Task-level reasoning.* The benchmark evaluates goals, preconditions, changes, and failure explanations rather than isolated object attributes. (v) *Reproducibility.* The benchmark provides public inputs, hidden answers, validation procedures, and a shared evaluator to ensure comparable measurements. Table 1 compares EviStateBench with existing embodied benchmarks. Rather than replacing task execution or perception benchmarks, EviStateBench targets an intermediate capability: transforming imperfect observations into temporally consistent and evidence-grounded task-state beliefs.

3 Benchmark Construction and Evaluation Framework

3.1 Benchmark Overview and Boundary

EviStateBench evaluates whether embodied agents can maintain temporally grounded and evidence-aware task states from imperfect observation streams. Figure 1 illustrates the overall benchmark

workflow. Simulator-grounded task executions are converted into hidden state timelines, public observation streams, query workloads, and hidden evaluation targets.

Each benchmark instance provides task specifications, observation streams, queries, and metadata as public inputs. Hidden state timelines and answer sets are generated from simulator truth and task specifications and are used only by the evaluator. The boundary in Figure 2 separates information available to systems from information reserved for evaluation. This design enables representation-independent evaluation. A system may use retrieval memory, learned memory, symbolic reasoning, temporal logs, database representations, or hybrid architectures. The benchmark evaluates only standardized QueryAnswer predictions, allowing different reasoning mechanisms to be compared under the same observation stream and evaluation protocol.

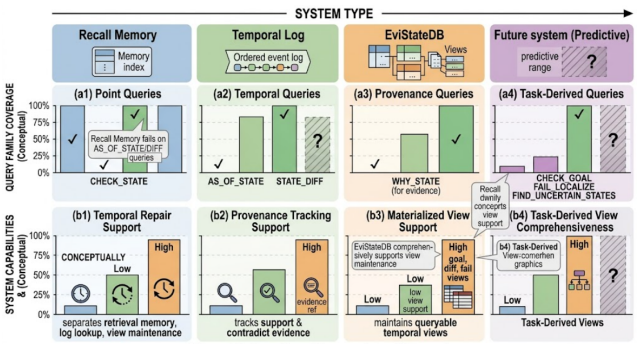

Figure 2: Boundary in EviStateBench.

3.2 Benchmark Generation

The benchmark generation pipeline creates controlled observation streams from simulator-grounded executions. For each task episode, EviStateBench first extracts semantic state transitions and constructs a hidden temporal state timeline. The timeline serves as the source for both public observations and hidden answers. Clean observation streams are generated from the hidden timeline with minimal noise. Perturbed streams introduce missing evidence, delayed observations, confidence degradation, and contradictory observations. These perturbations are designed to evaluate reasoning under imperfect evidence rather than to model a specific sensor failure mode. The resulting artifact is summarized in Table 2.

Table 2: EviStateBench profile.

Item	Count or status
Episodes	570
Activities	36
Task specifications	570
Target predicates	11
Clean observations	2,856,815
Hard mixed observations	2,797,127
Queries	5,848
Validation errors	0

3.3 Workloads and Perturbation Settings

EviStateBench evaluates multiple dimensions of temporal reasoning rather than a single task-success score. Table 4 summarizes the query families, which cover state reasoning, historical reasoning, evidence attribution, uncertainty handling, task constraints, state transitions, and failure analysis. The hard mixed setting combines multiple evidence challenges, including missing observations, delayed arrivals, confidence degradation, and contradictions. Unlike end-to-end embodied evaluation, EviStateBench fixes the observation stream and evaluates how systems interpret, update, and reason over available evidence.

Table 3: Query workloads in EviStateBench.

Query	Input	Capability
`CHECK_STATE`	Predicate and time	State reasoning
`AS_OF_STATE`	Valid and arrival time	Temporal availability
`WHY_STATE`	State and evidence	Evidence attribution
`STATE_DIFF`	Time interval	State transition reasoning
`CHECK_GOAL`	Goal specification	Goal reasoning
`FIND_UNCERTAIN_STATES`	State scope	Uncertainty handling
`CHECK_PRECONDITION`	Task constraints	Task reasoning
`FAILURE_LOCALIZATION`	Goal subset	Failure analysis

3.4 Reference System and Evaluation Protocol

Figure 3 summarizes the capability dimensions evaluated by EviStateBench. Instead of measuring only final task completion, the benchmark analyzes whether systems can maintain temporal consistency, attribute evidence, handle uncertainty, and answer structured task-level queries.

We provide EviStateDB as an interpretable reference system. EviStateDB maintains temporal state representations by tracking valid-time intervals, transaction-time revisions, confidence, and supporting or contradicting evidence. It also derives task-level states required for goal checking, uncertainty analysis, state changes, and failure localization. EviStateDB is not used to generate benchmark labels. Hidden answers are independently generated from simulator truth and task specifications, allowing EviStateDB to be compared fairly with retrieval-based, learned, and hybrid approaches. All systems output standardized QueryAnswer predictions. The evaluator reports exact correctness, value accuracy, status accuracy, structured-answer metrics, and confidence calibration when applicable. This protocol enables controlled comparison across different architectures while keeping the benchmark focused on temporal reasoning in embodied agents.

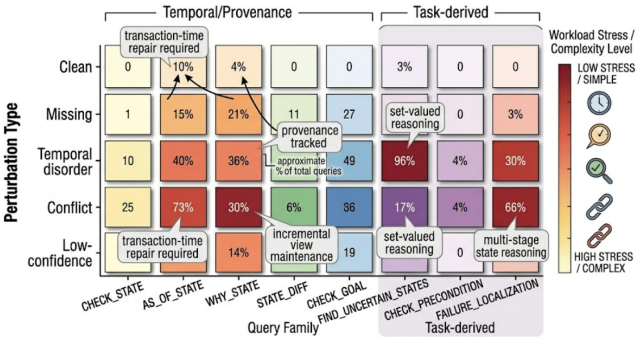

Figure 3: Capability dimensions evaluated by EviStateBench. The benchmark measures temporal reasoning, evidence attribution, uncertainty handling, state changes, and task-level reasoning.

4 Evaluation and Analysis

4.1 Implementation and Characterization

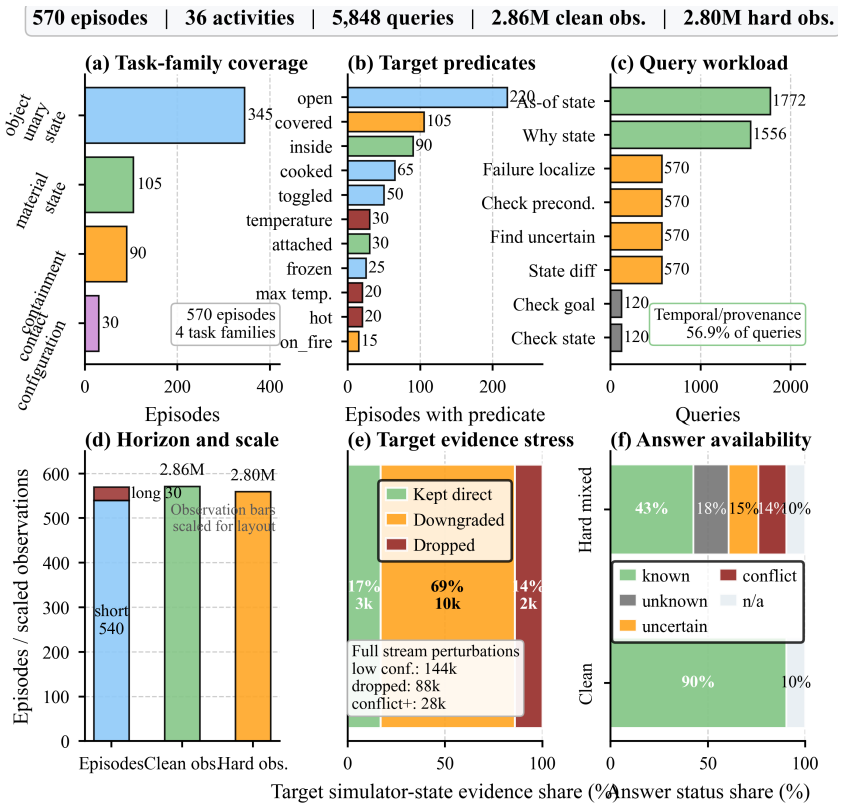

Figure 4: EviStateBench covers 570 simulator-grounded episodes over 36 activities, four task-state families, and 11 target predicates.

The first evaluation question is whether EviStateBench is a concrete, validated benchmark package rather than only a proposal. Figure 4 summarizes the released workload. The reported artifact contains 570 simulator-grounded episodes over 36 activities, four task-state families, 11 target predicates, 5,848 hard queries, 2.86M clean observations, and 2.80M hard mixed observations. Public-artifact validation, query-semantics audit, and answer-schema validation all pass, so the reported experiments are run on a fixed and reproducible release. The artifact covers object unary state, material state, containment, and contact configuration, with target predicates ranging from `open`, `covered`, and `inside` to numeric and material predicates such as `temperature`, `max_temperature`, `cooked`, and `frozen`. Its query mix is dominated by temporal and evidence-centric workloads such as `AS_OF_STATE` and `WHY_STATE`, while also including structured and task-derived queries such as `STATE_DIFF`, `FIND_UNCERTAIN_STATES`, `CHECK_PRECONDITION`, and `FAILURE_LOCALIZATION`. This diversity is important: a benchmark limited to one predicate type or one easy query shape would be vulnerable to specialized rules rather than testing general temporal task-state view maintenance.

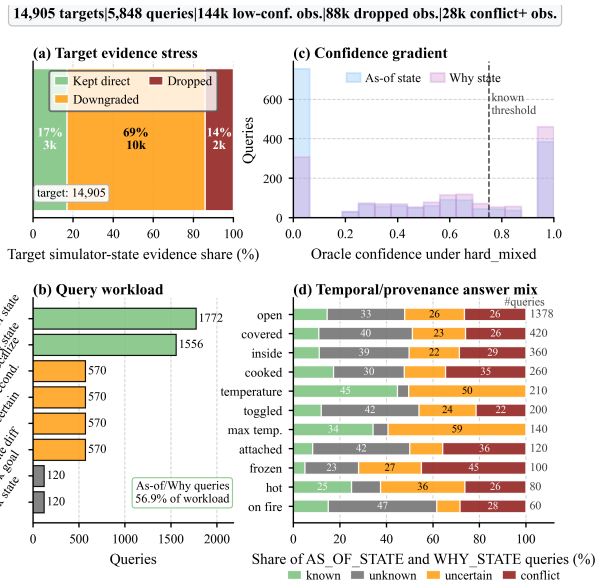

Figure 5: Hard-split mechanism and query difficulty profile. The hard stream keeps only a minority of target simulator-state observations as direct evidence, while downgrading or dropping the rest.

4.2 Hard-Split Difficulty

Figure 5 explains why the hard split is more than a larger file. Among 14,905 target simulator-state observations, only 2,512 remain as direct target evidence, 10,282 are downgraded to lower-confidence evidence, and 2,111 are dropped. The overall stream also contains 87,571 dropped observations, 144,011 low-confidence observations, and 27,883 injected conflicts. This perturbation is target-aware: it stresses evidence near the states that queries ask about, while leaving most background observations intact. The hard split is therefore a controlled stress test for data-management behavior. It is not an empirical model of every real sensor failure. Its purpose is to create missing, low-confidence, delayed, and conflicting evidence regimes in which systems must decide whether a state is known, unknown, uncertain, or conflicting. The confidence and predicate panels in Figure 5 show that the resulting workload is not binary. Different predicates exhibit different answer



Figure 6: Aggregated long state-chain templates.

mixtures, and temporal/provenance queries include high-confidence, low-confidence, missing, and conflicting cases.

4.3 Long State Chains

Most episodes in the reported artifact are short controlled transitions, but the release also includes a 30-episode long state-chain subset over cooking and freezing activities. The cooking chains move from an actuator state to temperature and maximum-temperature changes, and then to `hot` and `cooked` states. The freezing chain moves from opening the refrigerator to temperature change and then to `frozen` state. These chains create cases in which later task predicates depend on earlier state changes, so a system must maintain more than a single isolated predicate flip. Figure 6 makes this diagnostic role explicit. The subset contains three long-chain activity templates and 180 truth-level state changes, producing 360 hard as-of queries under the hard mixed stream. The answer mix includes `known`, `unknown`, `uncertain`, and `conflict` cases, showing that chained state consequences are converted into temporal availability tests rather than only final-state checks. This subset is not large enough to claim broad long-horizon BEHAVIOR coverage; instead, it demonstrates that EviStateBench can express multi-stage temporal state maintenance beyond a single target transition.

4.4 Baseline Results

Figure 7 compares three reference baselines on the final hard split. Recall memory answers from retrieved or recent evidence without maintaining a temporal state view. Temporal log keeps an explicit event log and evaluates temporal queries over it. EviStateDB maintains a bitemporal task-state view with support and contradict evidence.

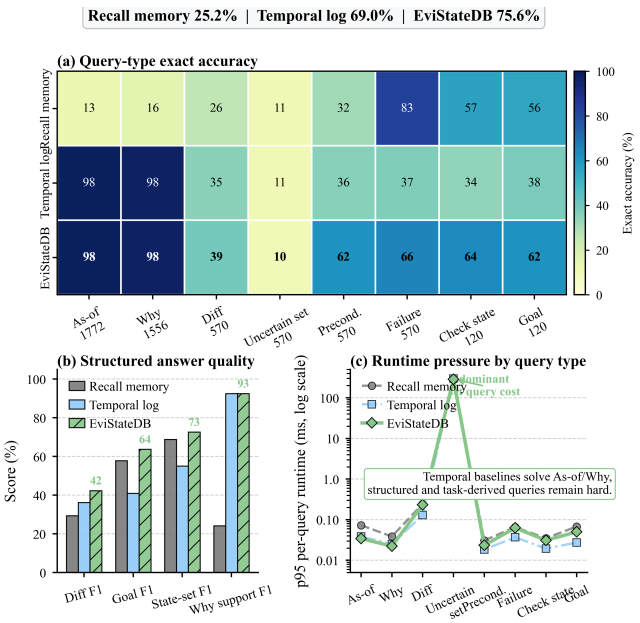

Figure 7: Baseline results on the final hard split. The upper heatmap reports query-type exact accuracy, the lower-left panel reports structured answer metrics, and the lower-right panel reports p95 per-query runtime from prediction outputs. All methods cover the full 5,848-query workload.

The overall exact accuracies are 25.2% for Recall memory, 69.0% for Temporal log, and 75.6% for EviStateDB. These results support the benchmark's diagnostic goal: the workload separates retrieval-style memory, append-only temporal lookup, and materialized temporal-state views. The largest gap is between retrieval-style memory and the two temporal baselines; the additional gain from Temporal log to EviStateDB is smaller and concentrates in structured and task-derived workloads. Temporal log and EviStateDB reach 98% exact accuracy on `AS_OF_STATE` and `WHY_STATE`, but they still struggle on structured workloads. EviStateDB reaches only 39% exact accuracy on `STATE_DIFF` and 10% on `FIND_UNCERTAIN_STATES`, although its partial structured metrics are higher: 42.4% Diff F1, 63.8% Goal F1, 72.6% State-set F1, and 92.5% Why-support F1. The runtime panel reports per-query prediction time, not full end-to-end indexing or stream-loading time. It shows a different stress point: most point-state queries are inexpensive after indexing, while `FIND_UNCERTAIN_STATES` dominates p95 query time for all methods. This is useful benchmark information because it separates semantic difficulty from query-time pressure.

4.5 Failure and Difficulty Diagnostics

The baseline diagnostics separate two different messages. First, answer status matters. Recall memory fails especially on unknown, uncertain, and conflict answers, while Temporal log and EviStateDB recover most unknown and uncertain statuses. Second, the high predicate-level accuracy for Temporal log and EviStateDB on `AS_OF_STATE` and `WHY_STATE` should not be overstated. It means that explicit event-time and transaction-time maintenance is effective for these two query families. It does not mean that all of EviStateBench is easy, because the residual errors in Figure 7 concentrate in set-valued uncertainty, state diffs, preconditions, failure localization, and goal views. Figure 8 explains where residual difficulty comes from. Retrieval fails broadly, while Temporal log and EviStateDB largely recover temporal and provenance answers. The remaining difficulty is therefore concentrated in answer shapes that require sets, diffs, preconditions, failure localization, and goal aggregation.

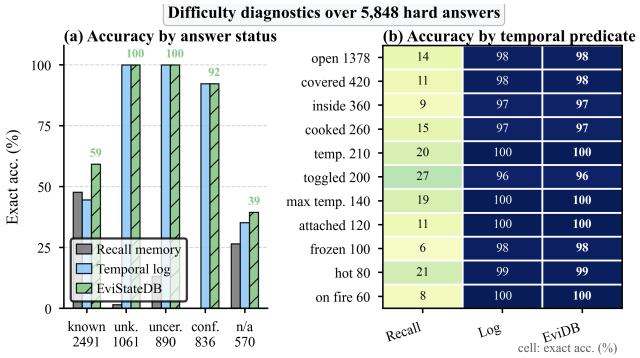

Figure 8: Difficulty analysis by answer status and predicate. Recall memory degrades across answer statuses and target predicates, while explicit temporal-state baselines largely solve `AS_OF_STATE` and `WHY_STATE`.

The interpretation should avoid two common mistakes. It should not treat EviStateDB as the oracle merely because it is the reference engine. It should also not claim that a lower score means a method is bad for all embodied AI tasks. A retrieval memory that performs poorly on state diffs may still be useful for open-ended question answering. The EAB claim is narrower and more useful: EviStateBench identifies which systems are good at temporal task-state view maintenance under a shared, reproducible workload.

5 Discussion

Insights from Benchmarking. EviStateBench reveals several challenges for temporal reasoning in embodied agents. First, maintaining only the latest state is insufficient: delayed evidence can change what an agent should believe at a previous time, while contradictory observations require explicit uncertainty and conflict handling. Second, memory retrieval and state reasoning are fundamentally different capabilities. Retrieving relevant observations may help answer recall questions, but embodied tasks often require maintaining consistent states, changes, preconditions, and goals over time. Third, evidence tracking is essential for reliable reasoning. Supporting and contradicting observations improve not only explanation queries, but also uncertainty estimation, failure analysis, and trust calibration. Fourth, task-level reasoning remains challenging even when individual predicates can be inferred correctly, suggesting that future embodied agents need structured mechanisms for composing temporal evidence into higher-level beliefs. Finally, reliable evaluation requires careful benchmark construction: hidden-state generation, query semantics, and validation procedures are necessary to distinguish reasoning failures from evaluation artifacts.

Limitations and Future Directions. EviStateBench intentionally focuses on a controlled temporal reasoning setting. The current artifact uses simulator-grounded and structured observation streams rather than raw perception inputs, and it does not directly evaluate perception, low-level control, exploration, or language-grounded planning. This design improves reproducibility and isolates temporal reasoning capability, but limits conclusions about complete real-world embodied systems. Future extensions can incorporate raw visual observations, multimodal evidence, and longer-horizon tasks while preserving the same evaluation principles. In addition, the current benchmark provides reference systems based on retrieval, temporal logging, and explicit state maintenance. Learned memory systems, LLM-based agents, and full embodied pipelines represent important future directions, provided that they operate under the same public-input and hidden-evaluation protocol.

Broader Implications for Embodied Intelligence. Reliable embodied agents require more than collecting observations or retrieving past experiences; they must maintain coherent beliefs as evidence accumulates, changes, and sometimes conflicts. EviStateBench provides a controlled framework for studying this intermediate capability between perception and action. By separating observation collection from temporal reasoning, the benchmark enables future research on memory architectures, uncertainty-aware agents, and multimodal reasoning systems. More broadly, the benchmark highlights the importance of evaluating not only whether an agent succeeds, but also whether it can explain, revise, and maintain the beliefs that support its decisions.

6 Related Work

Embodied Task and Simulation Benchmarks. BEHAVIOR introduced a benchmark of everyday household activities with predicate-logic activity specifications and simulation-based evaluation Srivastava et al. (2022). BEHAVIOR-1K scales this line of work to 1,000 activities and introduces OmniGibson as a simulation environment for realistic physics and rendering Li et al. (2024). Habitat and Habitat 2.0 support embodied navigation and rearrangement evaluation Savva et al. (2019); Szot et al. (2021); AI2-THOR provides an interactive 3D environment for visual AI Kolve et al. (2017); VirtualHome represents household activities as executable programs Puig et al. (2018); and TEACh studies embodied agents that interact through dialogue Padmakumar et al. (2022). EviStateBench uses the same general idea that task specifications and simulator truth can ground evaluation, but it changes the target from embodied task success to temporal task-state view maintenance. ALFRED evaluates instruction following Shridhar et al. (2020), and OpenEQA evaluates embodied question answering over episodic memory and active exploration Majumdar et al. (2024).

Temporal and Provenance Data Management. Temporal databases separate the time when a fact is valid in the modeled world from the time when a database records or knows the fact Snodgrass (1999); Jensen & Snodgrass (1999). Stream systems study continuous processing over arriving data Babcock et al. (2002); Garofalakis et al. (2016). Incremental view maintenance studies how derived database state can be updated after base changes Gupta & Mumick (1995); Zhuge et al. (1995). Provenance and uncertain-data systems address evidence and noisy facts Green et al. (2007); Cheney et al. (2009); Aggarwal & Yu (2009); Dalvi & Suciu (2007). EviStateBench brings these ideas into embodied task-state maintenance, where event-time and arrival-time mismatch, missing observations, conflicting evidence, and task-derived goal views appear together.

Experiment Analysis and Benchmark. Database benchmark and EAB papers often contribute workload models, artifact rules, and diagnostic analyses rather than only a new algorithm. LDBC standardizes graph workloads Erling et al. (2015); Iosup et al. (2016); TSM-Bench benchmarks time-series systems Khelifati et al. (2023); and FinBench defines a financial graph transaction workload Qi et al. (2025). Recent EAB papers similarly contribute evaluation frameworks and fine-grained diagnostics, including SQLMorph Malekpour et al. (2026), BEACON Najafi et al. (2025), and plan-based AQP analysis Mu et al. (2025).

7 Conclusion

EviStateBench is a data-management benchmark for temporal task-state view maintenance over embodied observation streams. It separates public observation streams and queries from hidden timelines and answers, evaluates systems through standardized QueryAnswers, and makes event time, arrival time, evidence, perturbations, and task-derived state views first-class benchmark concerns. EviStateDB provides a reference baseline engine, but the benchmark oracle comes from simulator truth and task specifications. The reported artifact and experiments show that temporal maintenance strategies outperform retrieval-style memory on this workload, while structured, set-valued, and task-derived queries remain open stress points for submitted systems. By fixing the public artifact, hidden oracle, prediction schema, and evaluator, EviStateBench provides a reproducible basis for comparing future retrieval, database, and learned-memory systems on the same temporal state-maintenance problem.

AI USE STATEMENT

AI tools assisted with literature search, manuscript organization and writing edits, discussion of experimental design, and code for figures and analysis. The authors are responsible for the experiments, numerical results, references, analyses, and final manuscript.

ETHICS STATEMENT

This study uses simulator-grounded observation streams and task specifications. It does not involve human participants or physical robot deployment. The reported results do not establish reliability in real-world settings.

REPRODUCIBILITY STATEMENT

The main text describes benchmark construction, perturbations, baselines, and evaluation metrics. Appendix A specifies hidden state timelines, public evidence streams, query workloads, answer normalization, and scoring semantics.

REFERENCES

Charu C. Aggarwal and Philip S. Yu. A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623, 2009. doi: 10.1109/TKDE.2008.190.

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pp. 1–16, 2002.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. doi: 10.1561/1900000006.

Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *Proceedings of the International Conference on Very Large Data Bases*, pp. 864–875, 2007.

Orri Erling, Alex Averbuch, Josep Larriba-Pey, Hassan Chafi, Andrey Gubichev, Arnau Prat, Minh-Duc Pham, and Peter Boncz. The LDBC social network benchmark: Interactive workload. *Proceedings of the VLDB Endowment*, 8(12):2022–2033, 2015. doi: 10.14778/2824032.2824070.

Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi. *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2016.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pp. 31–40, 2007.

Ashish Gupta and Inderpal Singh Mumick. Maintenance of materialized views: Problems, techniques, and applications. *IEEE Data Engineering Bulletin*, 18(2):3–18, 1995.

Alexandru Iosup, Tim Hegeman, Wing Lung Ngai, Stijn Heldens, Arnau Prat-Pérez, Thomas Manhardt, Hassan Chafi, Mihai Capotă, Narayanan Sundaram, Michael Anderson, Ilie Gabriel Tănase, Yinglong Xia, Lifeng Nai, Peter Boncz, and Georgos Siganos. LDBC graphalytics: A benchmark for large-scale graph analysis on parallel and distributed platforms. *Proceedings of the VLDB Endowment*, 9(13):1317–1328, 2016. doi: 10.14778/3007263.3007270.

Christian S. Jensen and Richard T. Snodgrass. Temporal data management. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):36–44, 1999. doi: 10.1109/69.755613.

Abdelouahab Khelifati, Mourad Khayati, Anton Dignös, Djellel Difallah, and Philippe Cudré-Mauroux. Tsm-bench: benchmarking time series database systems for monitoring applications. *Proceedings of the VLDB Endowment*, 16(11):3363–3376, 2023.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

Chengshu Li et al. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.

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. doi: 10.1109/CVPR52733.2024.01560.

Mohammadhossein Malekpour, Mohamed Riahi, Maxime Lamothe, and Amine Mhedhbi. SQL-Morph: Query mutation and fine-grained metrics for text-to-SQL evaluation. In *Proceedings of the IEEE International Conference on Data Engineering*, 2026. Experiment, Analysis, and Benchmark paper.

Pei Mu, Anderson Chaves Carniel, Antonio Barbalace, and Amir Shaikhha. [experiment, analysis, and benchmark] systematic evaluation of plan-based adaptive query processing. *arXiv preprint arXiv:2511.16455*, 2025.

Mohammad Matin Najafi, Xianju Zhu, Chrysanthi Kosyfaki, Laks V. S. Lakshmanan, and Reynold Cheng. BEACON: A benchmark for efficient and accurate counting of subgraphs. *arXiv preprint arXiv:2504.10948*, 2025.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018.

Shipeng Qi, Bing Tong, Jiatao Hu, Heng Lin, Yue Pang, Wei Yuan, Songlin Lyu, Zhihui Guo, Ke Huang, Xujin Ba, et al. The ldbc financial benchmark: Transaction workload. *Proceedings of the VLDB Endowment*, 18(9):3007–3020, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Richard T. Snodgrass. *Developing Time-Oriented Database Applications in SQL*. Morgan Kaufmann, 1999.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martin-Martin, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C. Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. BEHAVIOR: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Proceedings of the Conference on Robot Learning*, 2022.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vibhav Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X. Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, 2021.

Yue Zhuge, Hector Garcia-Molina, Joachim Hammer, and Jennifer Widom. View maintenance in a warehousing environment. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 316–327, 1995. doi: 10.1145/223784.223848.