CEM-AGENT: COUPLING TOOL SCHEDULING AND COMPOSITIONAL EVIDENCE FOR NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional navigation requires binding observations across objects, encounters, and time before committing progress. We introduce CEM-Agent, which couples a frozen vision–language planner with Compositional Evidence Memory (CEM). CEM constructs source-linked witnesses for bounded task claims and exposes unresolved roles and relations to guide subsequent tool use. A shared gate checks semantic evidence and execution facts before accepting progress. We test whether evidence composition benefits more from autonomous than fixed tool scheduling through matched 2×2 comparisons on 100 compositional navigation episodes. With oracle pose, CEM raises strict episode success from 43% to 51% under autonomous scheduling, versus 34% to 40% under fixed scheduling. The resulting interaction is $+2$ percentage points; the estimated-pose comparison yields the same point estimate. On fixed histories, witness recovery improves from 69% to 79%. Exposing gap feedback raises autonomous CEM success from 47% to 51% and improves short-horizon gap resolution. These results support the usefulness of evidence diagnosis and its feedback to planning, while the small positive interaction remains descriptive because paired uncertainty is not established.

1 INTRODUCTION

Vision-and-language navigation (VLN) studies how an embodied agent follows language instructions using visual observations. The field has progressed from graph-based route following to continuous control and increasingly diverse navigation tasks (Anderson et al., 2018; Ku et al., 2020; Krantz et al., 2020; Qi et al., 2020; Sun et al., 2026). Foundation models further enable language-conditioned planning, video-based action prediction, and hierarchical navigation (Zhou et al., 2024; Zhang et al., 2024b; Cheng et al., 2025; Wei et al., 2025). Across these settings, an agent must determine not only where to move, but also which parts of an instruction its observations and actions actually establish.

Consider the instruction, "Pass the red doorway, turn left, find the chair, and return to the same doorway." Detecting a doorway, recording a left turn, and later observing another doorway are individually relevant events. They do not establish that the turn followed the intended passage or that the final doorway is the original instance. Successful execution requires binding evidence to the correct object, encounter, temporal order, and spatial context. A relevant memory can therefore be an insufficient or misleading witness of progress.

This problem interacts with how the agent obtains information. Tool-using navigation agents can request depth measurements, recall earlier views, or issue pixel-grounded movement commands (Li et al., 2026; Zhou et al., 2026). Flexible tool selection does not itself ensure that the resulting history justifies progress. Conversely, detecting missing evidence has limited behavioral value if the planner cannot act on that diagnosis. We therefore ask: *does explicit evidence composition improve navigation more when tool scheduling is autonomous, and what connects the two?*

We propose CEM-Agent, which couples a frozen vision–language planner with Compositional Evidence Memory (CEM). Given a bounded claim about the active subgoal, CEM constructs candidate witnesses from source-linked observations and motion records. It evaluates both the suitability of each record and the relations that make those records jointly meaningful. When a witness is incomplete, CEM returns the unresolved roles or bindings to the planner, which can retrieve history,

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.

request a measurement, or move for another view. A separate progress gate combines checked semantic evidence with the applicable execution facts before updating the task ledger.

Our hypothesis is that the two components interact on the episode-success scale. Evidence composition can expose a missing binding, and autonomous scheduling can select a read or motion that addresses it. A fixed schedule may instead repeat an uninformative tool or stall behind a stricter progress gate. This hypothesis is stronger than expecting the full system to perform best: it predicts that CEM's improvement over generic memory is larger under autonomous than fixed scheduling. Tool calling, mapping, and low-level control are shared components; only the evidence heads are trained. Our contributions are:

- An evidence-grounded agent architecture in which unresolved task requirements inform autonomous tool selection and source-linked checks govern progress commitment.
- A compositional memory that scores task roles and their instance, encounter, temporal, and spatial bindings under a bounded context, while distinguishing missing evidence from refutation.
- An evaluation organized around interaction, mechanism, and localization dependence: matched factorial comparisons, fixed-history evidence tests, and an intervention on feedback to the planner.

The experiments follow one evidence chain. Closed-loop factorial comparisons yield a small positive interaction of $+2$ percentage points under both oracle and estimated pose. Fixed-history tests show higher witness recovery and less false support; exposing gap feedback then improves both gap resolution and episode success. The evidence is strongest for these component and feedback gains. Without paired uncertainty estimates, the factorial results suggest complementarity but do not establish a statistically reliable interaction or its invariance to localization error.

2 Related Work

Vision-and-language navigation and foundation models. R2R and RxR ground route instructions in indoor trajectories (Anderson et al., 2018; Ku et al., 2020); REVERIE adds referred-object identification (Qi et al., 2020), and VLN-CE requires continuous execution beyond a navigability graph (Krantz et al., 2020). Foundation-model approaches advance video-conditioned action prediction, task unification, and historical context (Zhang et al., 2024b;a; Wei et al., 2026). NaVILA and DualVLN connect high-level reasoning to locomotion skills or pixel-grounded trajectory control (Cheng et al., 2025; Wei et al., 2025). CEM-Agent addresses a complementary question: which observations and execution records jointly justify a compositional progress judgment?

Language-model planning and tool-using embodied agents. ReAct interleaves reasoning and interaction (Yao et al., 2023); SayCan, Code as Policies, and LM-Nav connect language planning to affordances, APIs, and pretrained navigation components (Ichter et al., 2023; Liang et al., 2023; Shah et al., 2023). VLN agents use explicit planning, expert consultation, spatial-temporal reasoning, or language-conditioned value maps (Zhou et al., 2024; Long et al., 2023; Qiao et al., 2025; Long et al., 2025; Yokoyama et al., 2023). AgenticNav is especially close: it exposes depth, visual recall, and pixel-grounded motion as tools (Li et al., 2026). Minimal-interface agents similarly let general-purpose models manage the interaction loop (Zhou et al., 2026). CEM-Agent builds on these interfaces by exposing claim-specific missing bindings to the planner and checking them before commitment. The component controls in Section 4 test this addition within a shared tool stack.

Navigation memory and structured environment representations. HAMT represents observation histories, while DUET, ETPNav, and GridMM organize navigation through topological or spatial memory (Chen et al., 2021; 2022; An et al., 2023; Wang et al., 2023). VLMaps and ConceptGraphs construct language-aligned maps and multi-view scene graphs (Huang et al., 2023; Gu et al., 2023). CEM builds on the value of relational structure but organizes records as witnesses for a bounded claim, exposing unresolved roles and bindings. Maps and target associations remain shared components. A Set Transformer (Lee et al., 2019) with the same records, metadata, schemas, and supervision tests whether explicit evidence factorization adds value beyond learned set interactions.

Progress estimation and evidence-based verification. Self-Monitoring and the Regretful Agent use progress estimates to monitor navigation and guide backtracking (Ma et al., 2019a;b). CEM instead exposes source-linked support, refutation, or insufficiency for an explicit claim. FEVER and HotpotQA provide conceptual precedents for separating claims from evidence and composing supporting facts (Thorne et al., 2018; Yang et al., 2018); embodied observations additionally have uncertain identities, encounter boundaries, and action-dependent coverage. Abstention connects to selective classification (Geifman & El-Yaniv, 2017). We evaluate empirical risk and coverage without claiming a distribution-free guarantee, and test whether missing-evidence feedback improves acquisition as well as judgment.

3 METHOD

Figure 1 summarizes CEM-Agent. At each control boundary, the planner receives the current view and an evidence packet for the active subgoal. It selects a read or motion tool; the resulting records extend the archive. Before progress is committed, a verifier checks the packet and the runtime checks the relevant execution facts.

3.1 TASK FORMULATION

At a physical control boundary t, the robot observes a forward RGB-D view $o_t = (I_t, D_t)$ and localization input $\hat{T}_t$. An instruction L induces bounded claims $q_{j,t}$ about subgoal j, including its referent, encounter, and time interval. History $\mathcal{H}_t$ contains only records available by time t. The planner selects a tool u and arguments ξ at internal round k:

$$(u_{t,k}, \xi_{t,k}) \sim \pi_\phi(\cdot \mid L, I_t, \ell_t, P_{t,k}, F_{t,k}, h_{t,k}), \tag{1}$$

where ℓ_t is the progress ledger, P an evidence packet, F tool facts, and h the current tool exchange. Read tools update context without moving the robot. An accepted motion ends deliberation and runs until completion or interruption. Read rounds cannot access future camera frames.

We distinguish world truth $y(q, t)$ from the status justified by an observed set S:

$$z(q, S) \in \{\mathsf{S}\ \text{(support)}, \mathsf{R}\ \text{(refute)}, \mathsf{I}\ \text{(insufficient)}\}. \tag{2}$$

An event can occur without being observed, and a completed command need not establish the intended object's identity. Refutation requires incompatible evidence or adequate coverage of the claim's interval; failure to retrieve support is insufficient, not refutation.

3.2 TOOL INTERFACE AND CAUSAL EVIDENCE ARCHIVE

The planner has access to geometric queries, target grounding, memory inspection, progress checks, and motion tools (Appendix A). It chooses their arguments and order, can skip unnecessary reads, and can re-plan after structured failures. The controller retains responsibility for local collision checking and execution. Tool responses include units, validity, uncertainty, and source IDs. A motion request receives a unique motion ID, with pending, completed, or interrupted status; a grounded target ID denotes an association hypothesis, not a ground-truth instance label.

Three linked stores serve distinct purposes: the occupancy map M_t supports reachable motion, the target/motion registry X_t maintains association and execution state, and the append-only archive $\mathcal{A}_t$ retains evidential records. A record includes its timestamp, source frame or motion event, transform estimate and quality, candidate targets, encounter association, and producer tool. Derived facts retain their parent record IDs. Versioned maps and associations prevent later corrections from silently changing the contexts used for earlier decisions. Evaluator labels, future records, and ground-truth object identities are never evidence inputs.

The runtime bounds the number of read calls before requiring an action or terminal request. An earlier accepted motion ends the loop. Exhaustion or rejection of the final request yields a hold with an explicit reason. This controls computation without imposing a fixed semantic sequence of perception, retrieval, and movement.

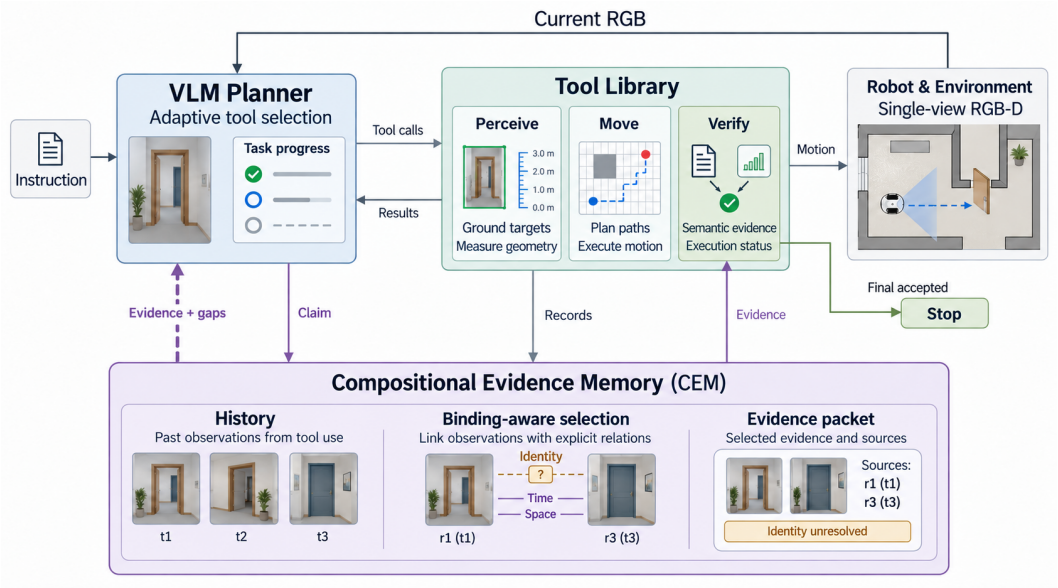


Figure 1: CEM-Agent connects progress judgments to information acquisition. CEM supplies source-linked evidence and unresolved bindings; the planner chooses subsequent tools, and semantic verification plus execution checks govern commitment. The doorway example is schematic.

3.3 Compositional Evidence Memory

Roles and bindings. A frozen query compiler maps claim q to support and refutation witness schemas $G_q^y = (U_q^y, R_q^y)$, for $y \in \{\mathsf{S}, \mathsf{R}\}$. Roles U specify required observations or events; edges R specify instance, encounter, temporal-order, or spatial-consistency relations. Alternative schemas permit different sufficient witnesses. For "currently back at the original doorway," support requires an original encounter, a later arrival, and an identity connection. An incompatible current location may refute that bounded claim, whereas an earlier failed attempt cannot refute eventual return.

An assignment $a : U \to C_t \cup \{\perp\}$ fills roles from a causal candidate pool $C_t \subset \mathcal{A}_t$ or marks them missing. Unary scores $u_\theta(q, r, e)$ assess record suitability; pairwise scores $b_\theta(q, r, r', e, e')$ assess required bindings from record embeddings and uncertain metadata. Same-instance and same-encounter are separate relations. IDs index association hypotheses; equality of IDs is not an oracle identity feature.

For a complete assignment, let $F_\theta(a)$ contain every required unary and relation score. We aggregate them with a smooth bottleneck:

$$c_\theta(a) = -\tau \log \left(\frac{1}{|F_\theta(a)|} \sum_{f \in F_\theta(a)} e^{-f/\tau} \right), \quad \tau > 0. \tag{3}$$

This discourages a high average relevance score from masking a weak necessary binding. It is an inductive bias rather than a logical guarantee, and finite-temperature scores may depend on schema size.

Witness search. A shared retriever combines relevance and recency to form a bounded candidate pool. Beam search expands high-scoring records per role plus a null assignment, separately for each schema and polarity. Search enforces causal constraints, deduplicates frames, and respects shared visual context limits. Among complete assignments, CEM ranks $c_\theta(a) - \lambda \,\mathrm{cost}(S(a))$, where cost measures serialized evidence length. Partial assignments remain available for deficit diagnosis but do not compete with complete witnesses through confidence in insufficiency. Thus, an empty packet receives no advantage merely because it is uninformative.

The returned packet is

$$P(q, t) = (S, a, \Delta, \Gamma, \rho), \tag{4}$$

where S contains source records, a their proposed roles, Δ missing or uncertain requirements, Γ detected conflicts, and ρ provenance and search-coverage metadata. If credible support and refutation both survive, their union is checked within the same budget; an oversized union or unresolved conflict yields insufficiency. If no complete witness survives, CEM returns an informative partial assignment. Search failure means that the available evidence and search budget did not establish a conclusion, not that the event is absent from the world.

3.4 Evidence-guided acquisition and progress commitment

CEM supplies context at the start of a boundary, answers explicit evidence queries, and constructs packets for subgoal checks. The planner can use Δ to choose a further observation: an identity gap may motivate recall or re-grounding, and uncertain execution status may call for motion inspection. A new view can clarify a current relation but cannot reconstruct an unobserved past event; if the required historical evidence is absent, the claim remains unresolved. Tool choices are model-directed, without a learned information-gain policy or a fixed deficit-to-action mapping.

A separate verifier reads source records and proposed assignments, with CEM's classification logits withheld. It returns a checked label $\hat{z}$, confidence $\hat{p}$, and unresolved requirements Δ_V, together with source citations. Low-confidence decisions and unresolved credible conflicts become insufficient. The runtime accepts a requested subgoal commitment only when

$$\mathrm{commit}_{j,t} = 1 \iff \mathrm{schemaOK} \wedge \mathrm{causalOK} \wedge \mathrm{depsOK} \wedge g_{j,t} \wedge [\hat{z}_{j,t} = \mathsf{S}], \tag{5}$$

where $g_{j,t}$ checks applicable execution or geometry predicates, such as completion of the relevant motion and membership in an approach region. A historical-event claim does not require current proximity unless the instruction does. Final Stop additionally requires every mandatory subgoal committed, a supported final claim, and no pending motion.

All central baselines use the same gate with their respective evidence packets. Accepted commitments are tied to claim, cutoff, association, and map versions; dependency changes trigger rechecking. The gate constrains commitment, but it does not certify perceptual correctness. An independent evaluator measures task success from the executed trajectory.

3.5 Learning

Frozen image/text features feed trainable role, relation, packet-classification, and missing-factor heads. Supervision includes alternative sufficient witnesses and relevance-preserving corruptions of instance, encounter, temporal, or spatial bindings. The objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{wit}} + 0.5\mathcal{L}_{\mathrm{rank}} + 0.5\mathcal{L}_{\mathrm{del}} + 0.5\mathcal{L}_{\mathrm{gap}}. \tag{6}$$

The witness loss marginalizes over valid assignments, ranking separates correct witnesses from corrupted alternatives, deletion training supervises insufficiency only when no valid witness remains, and gap prediction labels missing roles or relations. Missing-factor labels describe the retained evidence; they do not guarantee that a chosen tool can recover it.

The planner, compiler, grounder, and verifier remain frozen; only the evidence heads are trained. The packet classifier supplies auxiliary supervision; the separate verifier determines the progress gate's checked label. Support/refutation thresholds are calibrated on development data and frozen before testing. A 5% development error target is not a test-time risk guarantee. Appendix B gives the loss definitions and search details.

4 Experiments

The evaluation follows three questions: whether tool scheduling and evidence composition interact, whether evidence diagnosis and gap-guided acquisition explain the pattern, and whether the interaction persists with estimated localization. We separate the primary interaction contrast from supporting component and behavioral comparisons.

4.1 Experimental setup

Tasks and matched episodes. The study uses a navigation-only compositional slice of OmniNavBench (Sun et al., 2026), with static wheeled-robot instructions requiring target approach, land-

mark order, turns, or revisits. Eligibility is fixed from instruction/tool compatibility and scene availability before execution. Human-following, social interaction, and answer-generation tasks requiring other tools are outside this slice. All methods run once on the same 100 instruction episodes per comparison; the primary oracle- and estimated-pose comparisons reuse the same episode manifest. Ordinary failures and timeouts remain in the denominator. Scene and task identifiers preserve the pairing across methods and localization tracks.

Models, observations, and supervision. Frozen Qwen3.6-27B (Qwen Team, 2026) serves as planner, compiler, grounder, and verifier with separate prompts. Only evidence heads are trained. The offline dataset comprises 4,000 queries with a 3,000/500/500 training/development/test split; scenes and all derivatives of a trajectory remain within one partition. Agents share a forward RGB-D view, visual context rules, grounding, motion tools, executor, verifier, and progress gate. They receive neither panoramas nor ground-truth instance IDs. The oracle-pose track supplies shared simulator localization; the estimated-pose track replaces all pose-derived inputs, including mapping, association, and motion feedback.

Factorial controls. Scheduling is fixed or autonomous; memory is a generic Set Transformer (Lee et al., 2019) or CEM. The fixed scheduler follows a development-tuned sequence of spatial/motion inspection, grounding, evidence checking, and movement. Autonomous scheduling chooses the next legal tool from the same interface. Generic memory receives the same features, schemas, metadata, witness supervision, hard negatives, and missing-factor labels. Both memories can provide gap feedback; only its representation changes. The fixed scheduler does not adapt tool order to gap content. Thus, the central comparison changes scheduling and evidence organization without introducing a different verifier or action space.

Outcomes and uncertainty. Strict compositional success (SCS) is $100 \sum_i s_i/N$ for $N = 100$, where $s_i \in \{0, 1\}$ requires every necessary instruction predicate, dependency, and terminal condition. Episodes receive equal weight and no partial credit. Official evaluator scores are reported separately on this slice. False progress commitments (FPC) use accepted transitions as the denominator; no-progress timeouts (NPT) use episodes. The primary contrast is the interaction on the SCS percentage-point scale. We report aggregate point estimates, without paired confidence intervals or significance tests. Marginal success counts determine the interaction estimate but not its paired uncertainty; the interaction findings are therefore descriptive.

4.2 Interaction between tool scheduling and evidence composition

Let $p^{\ell}_{a,m}$ be SCS for scheduler $a \in \{\mathrm{fix}, \mathrm{auto}\}$, memory $m \in \{\mathrm{set}, \mathrm{CEM}\}$, and localization ℓ. The CEM effect within each scheduler and their interaction are

$$\delta^{\ell}_{a} = p^{\ell}_{a,\mathrm{CEM}} - p^{\ell}_{a,\mathrm{set}}, \tag{7}$$

$$\Delta^{\ell}_{\mathrm{int}} = \delta^{\ell}_{\mathrm{auto}} - \delta^{\ell}_{\mathrm{fix}}. \tag{8}$$

A positive interaction means that CEM's benefit is larger under autonomous scheduling on this scale. It does not require either component to have zero benefit in isolation, and it is not equivalent to an interaction coefficient on the log-odds scale.

Table 1: Oracle-pose factorial comparison on the same 100 episodes. Scores are percentages; successes are integer counts. Official is the benchmark evaluator score on this slice only. Bold marks the best observed score, not statistical significance.

Scheduling	Memory	Successes/100	SCS ↑	Official ↑	FPC ↓	NPT ↓
Fixed	Generic set	34	34	55	20.0	13
Fixed	CEM	40	40	58	13.1	15
Autonomous	Generic set	43	43	61	16.5	10
Autonomous	CEM	**51**	**51**	**65**	**9.4**	**8**

Interaction estimate. In Table 1, CEM raises SCS by 6 points under fixed scheduling (34 to 40) and 8 points under autonomous scheduling (43 to 51). Their difference gives $\Delta^{\mathrm{oracle}}_{\mathrm{int}} = +2$ points.

The full agent achieves the highest observed SCS and official score, but its 8-point gain over autonomous generic memory is a within-scheduler effect, not the interaction. The positive interaction estimate is small relative to the component gains and does not establish statistical significance.

Commitment and progress. FPC and NPT reveal a tradeoff hidden by success alone. Under fixed scheduling, CEM reduces FPC from 20.0% to 13.1%, while NPT rises from 13% to 15%. Under autonomous scheduling, FPC falls from 16.5% to 9.4% and NPT from 10% to 8%. This pattern is consistent with stricter evidence checking exposing deficits that a fixed schedule cannot resolve, whereas autonomous acquisition can address them. The mechanism tests below examine this interpretation directly.

4.3 Mechanism: from evidence diagnosis to information acquisition

We examine a two-stage explanation for the observed pattern: CEM first improves evidence judgment, and feedback about unresolved requirements then helps the planner obtain useful information. Fixed histories isolate judgment from navigation, and a feedback intervention tests the behavioral contribution of explicit deficits.

Stage 1: evidence judgment on fixed histories. Generic memory and CEM receive the same held-out histories, claims, schemas, context constraints, and verifier. Joint-F1 measures three-way classification on the common, independently identified multi-record query subset. Witness recovery tests whether a selected packet contains a sufficient witness on answerable queries. Packet-level false support and calibrated risk measure whether the cited evidence actually warrants the decision. Coverage uses all 500 held-out queries; Appendix A defines the conditional metrics. These targets distinguish answering the claim from supplying evidence that justifies it.

Table 2: Fixed-history mechanism comparison (%). Joint-F1 uses the same multi-record query subset for both models; coverage uses all 500 test queries. Witness and error metrics use their stated conditional denominators. Thresholds are frozen on development data.

Memory	Joint-F1 ↑	Witness ↑	False S ↓	Cov. ↑	Risk ↓
Generic set, matched labels	71.2	69.0	14.0	65.4	5.2
CEM	**80.0**	**79.0**	**8.5**	**74.8**	**4.8**

CEM improves Joint-F1 by 8.8 points and witness recovery by 10.0 points, while false support falls by 5.5 points (Table 2). Thus the classification gain accompanies better recovery of sufficient evidence, rather than merely more decisive answers. Coverage rises from 327/500 queries (65.4%) to 374/500 (74.8%), while observed decisive risk decreases from 5.2% to 4.8%. The 5% development calibration target is not a test guarantee: generic memory slightly exceeds it. These aggregate comparisons support better evidence selection but do not isolate which individual binding type accounts for the gain.

Stage 2: intervene on feedback to the planner. Under autonomous scheduling and oracle pose, the gap-hidden variant retains the CEM selector, source packets, verifier, thresholds, and final gate, while suppressing missing-role/relation feedback on every planner-visible path. This includes CEM fields, verifier messages, and semantic explanations returned by progress checks; indispensable execution-validity feedback remains available. The full variant exposes these deficits. The two policies start from the same 100 episodes, while their subsequent histories may diverge. Autonomous generic memory from Table 1 supplies context for the size of the CEM effect, without requiring another run.

Exposing gap feedback raises SCS from 47% to 51%, improves three-boundary resolution from 48.0% to 62.0%, and reduces the wasted-read fraction from 27.0% to 19.5% (Table 3). FPC also decreases from 11.0% to 9.4%. The joint improvement in resolution and success supports a behavioral contribution from explicit gap feedback beyond CEM's evidence selection. The lower wasted-read fraction does not imply fewer total tool calls or lower compute. Recoverability is labeled independently of subsequent policy success, but on-policy histories can diverge; these comparisons do not establish complete mediation of the factorial interaction.

Table 3: Feedback intervention with autonomous scheduling (%). Resolve: correct resolution within three boundaries among initially insufficient decisions independently labeled recoverable. Wasted reads: reads adding no new valid fact. Generic and full-CEM rows reuse primary runs.

Variant	SCS ↑	FPC ↓	Resolve ↑	Wasted reads ↓
Generic set, learned gap head	43	16.5	44.0	28.5
CEM, gap feedback hidden	47	11.0	48.0	27.0
CEM, gap feedback visible	**51**	**9.4**	**62.0**	**19.5**

4.4 Does the interaction persist under estimated localization?

We repeat the complete 2×2 comparison with shared estimated localization on the same episode manifest, measuring both scheduler-specific CEM effects. No simulator-pose-derived geometry, association, or completion feedback is retained in the estimated track.

Table 4: Estimated-pose factorial comparison on the same episodes as Table 1. Repeating all four cells identifies the interaction under estimated localization; percentages and integer success counts are reported separately.

Scheduling	Memory	Successes/100	SCS ↑	Official ↑	FPC ↓	NPT ↓
Fixed	Generic set	28	28	49	23.5	17
Fixed	CEM	33	33	52	16.8	19
Autonomous	Generic set	36	36	55	20.2	14
Autonomous	CEM	**43**	**43**	**59**	**13.0**	**12**

Under estimated pose, CEM improves SCS by 5 points with fixed scheduling and 7 points with autonomous scheduling. The interaction is therefore $(43 - 36) - (33 - 28) = +2$ points, matching the oracle point estimate. The full agent remains the best observed configuration, but its SCS falls from 51% to 43%; the other configurations lose 6–7 points. The fixed-schedule FPC/NPT tradeoff also remains: CEM reduces false commitments while increasing timeouts by 2 points. Thus the observed complementarity is not confined to oracle pose, yet composition does not remove localization sensitivity. Equal interaction point estimates establish neither statistical equivalence nor a reliably positive interaction in either track.

5 Conclusion

CEM-Agent connects evidence composition to subsequent tool selection through source-linked witnesses and explicit unresolved bindings. The full agent achieves the highest observed success in both localization tracks, with a small $+2$-point interaction in each. Fixed-history evidence recovery and the gap-feedback intervention provide more direct support for the proposed connection: better diagnosis accompanies more successful gap resolution and navigation. The results motivate coupling evidence checking with adaptive acquisition, while leaving the statistical reliability of the interaction open.

Limitations. The study concerns selected simulated navigation tasks, not full OmniNavBench coverage or hardware safety. Evidence supervision, association errors, shared-model errors, incomplete search, and irrecoverable observations can limit the method. One rollout on each of 100 episodes and aggregate-only reporting limit assessment of interaction uncertainty and scene dependence; no significance or equivalence claim is made. On-policy feedback interventions encounter different histories and cannot fully establish mediation. Shared tools do not imply equal compute, and the reported wasted-read fractions do not establish an overall efficiency advantage.

AI USE STATEMENT

Generative AI was utilized as an assistive tool in this work. Its applications included: (1) assisting in brainstorming the conceptual framework; (2) suggesting experimental design directions and generating schematic figure drafts, which the authors independently refined and executed; and (3) aiding literature search, bibliography checking, and language/structural revision. The authors are responsible for the final content.

REPRODUCIBILITY STATEMENT

The main text specifies the task scope, sensing and localization tracks, matched comparisons, and interaction estimands. The appendix defines metric denominators, evidence labels, tool semantics, and learning objectives. Reproduction requires the task manifest, executed configurations, model and prompt revisions, paired episode outcomes, and query- and decision-level records. Aggregate scores alone cannot reconstruct the paired interaction uncertainty.

ETHICS STATEMENT

The evaluation concerns simulated navigation. A progress gate does not guarantee collision avoidance or correct perception; hardware deployment requires separate validation. Any release of camera observations must respect permissions and privacy.

REFERENCES

Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. arXiv preprint arXiv:2304.03047, 2023. URL https://arxiv.org/abs/2304.03047.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History Aware Multimodal Transformer for Vision-and-Language Navigation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5834–5847, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2e5c2cb8d13e8fba78d95211440ba326-Abstract.html.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think Global, Act Local: Dual-Scale Graph Transformer for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16537–16547, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Think_Global_Act_Local_Dual-Scale_Graph_Transformer_for_Vision-and-Language_Navigation_CVPR_2022_paper.html.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. In *Robotics: Science and Systems*, 2025. URL https://navila-bot.github.io/.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. arXiv preprint arXiv:2309.16650, 2023. URL https://arxiv.org/abs/2309.16650.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual Language Maps for Robot Navigation. In *IEEE International Conference on Robotics and Automation*, 2023. URL https://arxiv.org/abs/2210.05714.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pp. 287–318, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/6018_ECCV_2020_paper.php.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020. doi: 10.18653/v1/2020.emnlp-main.356. URL https://aclanthology.org/2020.emnlp-main.356/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Yijian Li, Changze Li, Hantian Shi, Jiaying Luo, Jiyuan Cai, Ming Yang, and Tong Qin. AgenticNav: Zero-Shot Vision-and-Language Navigation as a Tool-Calling Harness. arXiv preprint arXiv:2606.10577, 2026. URL https://arxiv.org/abs/2606.10577.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. In *IEEE International Conference on Robotics and Automation*, 2023. URL https://code-as-policies.github.io/.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss Before Moving: Visual Language Navigation via Multi-expert Discussions. arXiv preprint arXiv:2309.11382, 2023. URL https://arxiv.org/abs/2309.11382.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 2049–2060, 2025. URL https://proceedings.mlr.press/v270/long25b.html.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. Self-Monitoring Navigation Agent via Auxiliary Progress Estimation. In *International Conference on Learning Representations*, 2019a. URL https://arxiv.org/abs/1901.03035.

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The Regretful Agent: Heuristic-Aided Navigation through Progress Estimation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019b. URL https://arxiv.org/abs/1903.01602.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Qi_REVERIE_Remote_Embodied_Visual_Referring_Expression_in_Real_Indoor_Environments_CVPR_2020_paper.html.

Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-Nav: Exploring Zero-Shot Vision-and-Language Navigation in Continuous Environment with Open-Source LLMs. In *IEEE International Conference on Robotics and Automation*, 2025. URL https://arxiv.org/abs/2409.18794.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Dhruv Shah, Błażej Osiński, Brian Ichter, and Sergey Levine. LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pp. 492–504, 2023. URL https://proceedings.mlr.press/v205/shah23b.html.

Samson Sun, Tianyi Yang, Tengyue Wang, Yikai Xue, Zhengjie Xu, Lingming Zhang, Qichen Zhang, Chao Liang, and Zhipeng Zhang. Beyond Isolation: A Unified Benchmark for General-Purpose Navigation. In *Robotics: Science and Systems*, 2026. URL https://arxiv.org/abs/2605.09441.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. GridMM: Grid Memory Map for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15625–15636, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_GridMM_Grid_Memory_Map_for_Vision-and-Language_Navigation_ICCV_2023_paper.html.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground Slow, Move Fast: A Dual-System Foundation Model for Generalizable Vision-and-Language Navigation. arXiv preprint arXiv:2512.08186, 2025. URL https://arxiv.org/abs/2512.08186.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming Vision-and-Language Navigation via SlowFast Context Modeling. In *IEEE International Conference on Robotics and Automation*, 2026. URL https://arxiv.org/abs/2507.05240.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. arXiv preprint arXiv:2312.03275, 2023. URL https://arxiv.org/abs/2312.03275.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks. arXiv preprint arXiv:2412.06224, 2024a. URL https://arxiv.org/abs/2412.06224.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation. arXiv preprint arXiv:2402.15852, 2024b. URL https://arxiv.org/abs/2402.15852.

Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7641–7649, 2024. doi: 10.1609/aaai.v38i7.28597. URL https://ojs.aaai.org/index.php/AAAI/article/view/28597.

Jian Zhou, Xunyi Zhao, Gengze Zhou, Zerui Li, Sihao Lin, Jiajun Liu, and Qi Wu. Embodied Agents Take Control: Minimal-Interface Zero-Shot Agents Rival Industrial-Scale Policies in Vision-and-Language Navigation. arXiv preprint arXiv:2607.26148, 2026. URL https://arxiv.org/abs/2607.26148.

A EVALUATION DEFINITIONS AND SHARED INTERFACES

Strict success and decision metrics. For episode trajectory τ_i, instruction predicates P_j, and required dependencies D_{ab}, the binary success indicator is

$$s_i = \mathbf{1}\left[\bigwedge_j P_j(\tau_i) \wedge \bigwedge_{(a,b)} D_{ab}(\tau_i) \wedge \mathrm{validTerminal}(\tau_i)\right]. \tag{9}$$

We report $\mathrm{SCS} = 100\sum_{i=1}^{N} s_i/N$ with $N = 100$ per evaluation condition. Each episode has equal weight, without partial credit; valid alternative routes are allowed. Ground-truth identities are evaluator-only. The official outcome is separate. FPC is the fraction of accepted progress transitions with false predicates or dependencies; its denominator is transitions, not episodes. NPT is the fraction of episodes ending in a no-progress timeout. The metrics must be read together because declining to commit can reduce accepted errors.

For checked prediction $\hat{z}_i$ and independent selected-packet label z_i^{pkt}, decisive risk R and coverage C are

$$R = \frac{\sum_i \mathbf{1}[\hat{z}_i \neq \mathsf{I},\ \hat{z}_i \neq z_i^{\mathrm{pkt}}]}{\sum_i \mathbf{1}[\hat{z}_i \neq \mathsf{I}]}, \qquad C = \frac{1}{N_q}\sum_i \mathbf{1}[\hat{z}_i \neq \mathsf{I}], \quad N_q = 500. \tag{10}$$

A decisive prediction on an insufficient packet counts as an error; risk is undefined at zero coverage. False support uses uncalibrated support predictions as its denominator. Joint-F1 instead compares predictions with the evidence status justified by the common full history, on an independently fixed multi-record subset. Thus a missed available witness affects query judgment, while an unsupported correct guess still fails packet verification. Witness recovery is measured on answerable queries against all valid sufficient sets. These conditional rates have different denominators and cannot be converted to counts using the full 500-query test set.

Resolve uses independently labeled recoverable, initially insufficient decisions as its denominator and requires correct resolution of the same bounded claim within three subsequent boundaries. Changed referents or temporal scope do not count as repairs. Termination without resolution counts as non-resolution. Wasted reads add no new valid fact, divided by all reads. A cache hit that supplies a fact absent from the planner's context is not automatically waste.

Scheduling and memory controls. Both schedulers use the same motion primitives, target registry, collision handling, and commitment gate. Generic memory receives the same features, schemas, witness supervision, negatives, and missing-factor labels. Fixed scheduling does not adapt tool order to deficit content. Hiding gap feedback masks every planner-visible semantic deficit channel, while retaining raw evidence, gate outcomes, and necessary execution-validity facts. It does not prevent the planner from inferring a gap from raw sources; the intervention isolates explicit feedback, not all awareness of uncertainty.

Evidence labels and interventions. World truth, common-history evidence status, and selected-packet evidence status are annotated separately. A true unobserved event can remain insufficient. Recoverability is audited from the initial state and accessible evidence independently of the tested policy's eventual outcome. Relevance-preserving substitutions require a verified invalid binding; deletion implies insufficiency only after every valid witness is removed. All trajectory derivatives stay in one partition; online retrieval is causal and restricted to the current episode.

Tool and execution contract. Read tools inspect depth, measure a pixel, ground a target, inspect spatial/motion state, query evidence, or check a subgoal. Motion tools move a bounded distance, turn, approach a pixel, or navigate toward a target's traversable approach region. A finish request passes through the shared progress gate. Each response carries a timestamp, source IDs, validity, units, and relevant map/association version. Read calls use the current physical observation; obtaining a new view requires motion. Invalid arguments return structured errors, and non-finite depth remains unknown. Motion completion and semantic identity are checked separately. In the estimated-pose track, map fusion, association, and motion feedback all use estimated localization; tracking loss freezes fusion and makes spatial facts unknown.

B LEARNING AND EVIDENCE SEARCH

Claim schemas and provenance. Each bounded claim specifies a predicate, object and encounter references, a time interval, dependencies, and alternative support/refutation schemas. Predicate families include target identity, approach arrival, passage, signed turn after passage, landmark order, and return. For "turn left after passing doorway A," a witness must connect identity, passage, and completed turn records to the same encounter in the required order. An earlier failed return cannot refute a claim about eventual return. Unsupported clauses remain unresolved rather than being dropped. Records retain their source frame or motion event, transform quality, producer tool, and parent IDs; subsequent association updates do not rewrite past decision contexts.

Evidence heads and objectives. Frozen visual/text features feed trainable role and relation heads. Role scores evaluate (q, r, e); relation scores evaluate (q, r, r', e, e') using visual/text features, time differences, spatial residuals, validity masks, and association uncertainty. A packet classifier and missing-factor head supply auxiliary supervision. The generic-set control receives the same underlying information without explicit role/relation bottleneck aggregation.

Let $\mathcal{W}_q^y$ contain valid complete assignments of polarity y and $\mathcal{B}_q^y$ contain training candidates, including hard negatives. The terms in Eq. 6 are

$$\mathcal{L}_{\text{cls}} = -\log p_\theta(z(q, S) \mid q, S), \tag{11}$$

$$\mathcal{L}_{\text{wit}} = -\log \frac{\sum_{a \in \mathcal{W}_q^y} \exp c_\theta(a)}{\sum_{a \in \mathcal{B}_q^y} \exp c_\theta(a)}, \tag{12}$$

$$\mathcal{L}_{\text{rank}} = \max(0, m - c_\theta(a^+) + c_\theta(a^-)), \tag{13}$$

$$\mathcal{L}_{\text{del}} = -\log p_\theta(\mathsf{I} \mid q, S \setminus D). \tag{14}$$

The witness term applies only when a decisive witness exists; the deletion term applies only when D destroys every valid alternative. $\mathcal{L}_{\text{gap}}$ is mean binary cross-entropy over missing roles and relations. Credible support/refutation conflicts carry a conflict flag and cannot force a decisive commitment.

Search and context. Candidate retrieval combines relevance and recency. Beam search expands role assignments separately for each schema and polarity, using cached relation scores and retaining null assignments to diagnose missing evidence. Complete witnesses are ranked by bottleneck score minus serialized evidence cost; partial packets prioritize reliable factor coverage, then consistency and cost. The current view is always included, historical frames are deduplicated, and every selector follows the same image-processing and context rules. Support/refutation conflicts are checked jointly; an unresolved conflict or a union that cannot fit in context yields insufficiency. This finite search can miss a valid witness, so search failure is not refutation.

For n records, r roles, e required relations, beam width w, and k candidates per role, cached scoring requires at most $O(rn + en^2)$ head evaluations, and search makes $O(rwk)$ extensions per schema. These costs are additional to model inference and shared navigation computation.

Verification and calibration. The verifier checks sources and assignments with CEM logits withheld, and cites the required roles and bindings. Unsupported or conflicting conclusions remain insufficient. Development calibration selects support/refutation thresholds to increase coverage subject to a 5% error target, then freezes them. Test risk and coverage are reported as observed; the target is not a guarantee and thresholds are not retuned on the test set.

Interaction estimate and uncertainty. For each episode define $d_i^\ell = s_{i,\text{auto,CEM}}^\ell - s_{i,\text{auto,set}}^\ell - s_{i,\text{fix,CEM}}^\ell + s_{i,\text{fix,set}}^\ell$. Then $\Delta_{\text{int}}^\ell = 100\bar{d}^\ell$. The marginal success counts suffice to compute this mean, giving $+2$ points in each localization track. They do not determine the variance of d_i^ℓ, which depends on within-episode outcome correlations. We therefore report no confidence interval or significance test from these aggregates. Quantifying uncertainty requires retaining paired outcomes across all methods and localization tracks and accounting for shared scenes. Likewise, the zero observed cross-localization difference does not establish equivalence.