SEER: Spatial-Embodied Evidence Reasoning for Long-Horizon Task Completion

Anonymous authors
Paper under double-blind review

Abstract

Embodied agents face a fundamental challenge in long-horizon tasks: maintaining coherent reasoning across extended sequences of observations and actions while grounding decisions in spatial evidence. Existing Vision-Language-Action (VLA) models either process each timestep independently, losing temporal context, or rely on implicit memory that degrades over long horizons. We introduce SEER (Spatial-Embodied Evidence Reasoning), a framework that explicitly constructs and maintains structured evidence chains grounded in spatial coordinates. SEER comprises three components: (i) an Evidence Chain module that extracts verifiable spatial evidence from each observation; (ii) a Spatial Memory Bank (SMB) that persists evidence as key-value pairs indexed by 3D coordinates with efficient radius-based retrieval; and (iii) a Retrospective Verification mechanism that cross-checks current observations against historical evidence to detect and correct reasoning errors. Experiments on ALFRED (Shridhar et al., 2020), Habitat 2.0 (Szot et al., 2021), and BEHAVIOR-1K (Li et al., 2024) demonstrate that SEER achieves 48.7%, 42.3%, and 36.9% task success rates, outperforming OpenVLA (Kim et al., 2024) by 13.1, 13.0, and 12.2 points respectively. On tasks with 8+ sub-steps, SEER maintains 33.1% success compared to 15.2% for OpenVLA. Ablation studies confirm that all three components contribute meaningfully, with Retrospective Verification providing 8.6 points of improvement on error-prone long-horizon tasks.

1 Introduction

Embodied AI aims to develop agents that perceive their environment, reason about goals, and execute actions to accomplish complex tasks. Recent advances in Vision-Language Models (VLMs) have catalyzed significant progress, with systems like RT-2 (Brohan et al., 2023), SayCan (Ahn et al., 2022), and OpenVLA (Kim et al., 2024) demonstrating impressive capabilities in robotic manipulation. However, these systems face a critical limitation when confronted with long-horizon tasks—scenarios requiring dozens of sequential decisions where intermediate errors compound and spatial consistency must be maintained over extended periods.

The core challenge lies in evidence persistence. When an agent observes a cup on the counter at timestep t and needs to retrieve it at timestep $t + 15$, it must remember both the object's identity and its spatial location. Current approaches handle this either through implicit representations in transformer attention (which degrade with sequence length) or through unstructured text memory (which lacks spatial grounding). Neither approach provides the verifiable, spatially-anchored evidence that robust long-horizon reasoning requires.

We observe that human cognition solves this problem through spatial evidence chains—structured records of what was seen, where it was located, and when it was observed, which can be retrospectively verified against new observations. Inspired by this, we propose SEER (Spatial-Embodied Evidence Reasoning), a framework that explicitly constructs, stores, and verifies spatial evidence throughout task execution. Our key contributions are:

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.

1. Evidence Chain Construction. A structured representation that decomposes each observation into verifiable spatial evidence items, bridging visual perception and action reasoning through explicit spatial anchoring (Section 3.1).
2. Spatial Memory Bank (SMB). A persistent key-value store indexed by 3D coordinates with efficient radius-based retrieval, enabling agents to access relevant historical evidence without scanning entire histories (Section 3.2).
3. Retrospective Verification. A cross-timestep consistency checking mechanism that detects reasoning errors by comparing current observations against stored evidence, triggering re-planning when conflicts arise (Section 3.3).
4. Comprehensive Evaluation. Extensive experiments on three benchmarks (ALFRED, Habitat 2.0, BEHAVIOR-1K) demonstrate consistent improvements over state-of-the-art methods, with detailed ablation studies and qualitative analysis (Section 4).

2 Related Work

Vision-Language-Action Models. Vision-Language-Action (VLA) models represent a paradigm shift in embodied AI, integrating visual perception and language understanding with action generation. RT-2 (Brohan et al., 2023) demonstrated that co-fine-tuning on robotic data enables VLMs to directly output actions as text tokens. RT-2-X (O'Neill et al., 2024) extended this with cross-embodiment training on the Open X-Embodiment dataset. OpenVLA (Kim et al., 2024) provided an open-source 7B VLA trained on 970K episodes. While these models excel at short-horizon manipulation, they lack explicit mechanisms for long-horizon evidence maintenance.

Embodied Reasoning and Planning. SayCan (Ahn et al., 2022) introduced affordance-grounded planning by combining LLM language priors with learned value functions. PaLM-E (Driess et al., 2023) integrated embodied multimodal inputs into a large language model for sequential planning. Embodied Chain-of-Thought (Zawalski et al., 2024) brought structured reasoning to VLA models, generating intermediate rationales before actions. However, these approaches either treat reasoning as transient (discarded after each step) or use unstructured text memory that cannot be spatially queried or verified.

Spatial Memory in Embodied AI. Spatial memory has been explored through scene graphs (Wald et al., 2020), topological maps (Chaplot et al., 2020), and neural radiance fields (Chen et al., 2023). VoxPoser (Huang et al., 2023) constructs 3D value maps for manipulation guidance. However, these representations are typically passive—they store spatial information but do not actively verify or update evidence through reasoning. SEER distinguishes itself by making spatial memory an active participant in the reasoning loop, with Retrospective Verification ensuring consistency.

Benchmarks for Embodied Task Completion. ALFRED (Shridhar et al., 2020) provides language-conditioned household tasks with long-horizon instructions. Habitat 2.0 (Szot et al., 2021) offers a high-fidelity simulator with physics and rearrangement tasks. BEHAVIOR-1K (Li et al., 2024) introduces 1,000 everyday activities with realistic simulation. These benchmarks increasingly emphasize long-horizon reasoning, making them ideal testbeds for evaluating evidence-based approaches.

Cognitive Foundations. The structured-evidence view aligns with the working-memory model of Baddeley (1992), which argues that structured, location-aware representations outperform unstructured buffers in tasks that require cross-step retrieval.

3 Method: SEER Framework

SEER augments a VLM-based policy with three components that operate at each timestep t: an Evidence Chain constructor, a Spatial Memory Bank, and a Retrospective Verification module. Figure 1 provides an overview of the framework.

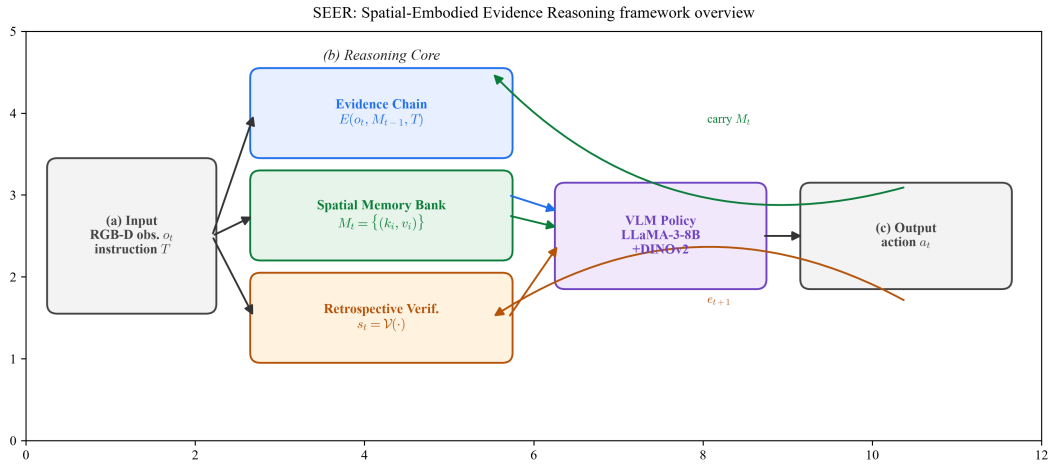

Figure 1: SEER framework overview. (a) Input: RGB-D observation o_t and task instruction T. (b) Reasoning Core: Evidence Chain, Spatial Memory Bank, and Retrospective Verification modules. (c) Output: action generation via a VLM policy network (LLaMA-3-8B + DINOv2).

3.1 Evidence Chain Construction

Given an observation o_t (RGB-D image) and a task instruction T, the Evidence Chain module produces a structured reasoning trace consisting of three elements: a spatial evidence set e_t, a reasoning step r_t, and an action a_t. Formally,

$$E(o_t, M_{t-1}, T) \to (e_t, r_t, a_t). \tag{1}$$

The evidence set $e_t = \{e_t^1, \ldots, e_t^K\}$ contains K evidence items, each represented as a triplet

$$e_t^k = (\text{type}^k,\ \text{loc}^k,\ \text{conf}^k), \tag{2}$$

where type denotes the evidence category (object, relation, spatial layout, or affordance), $\text{loc} \in \mathbb{R}^3$ is the 3D spatial coordinate derived from depth sensing, and $\text{conf} \in [0, 1]$ is the model's confidence in this evidence. The construction process involves three steps (Figure 2):

1. Scene Parsing—the VLM backbone extracts objects, relations, and spatial layouts from o_t.
2. Evidence Extraction—relevant evidence items are selected based on the current sub-task and memory state.
3. Action Generation—the policy network produces a_t conditioned on the evidence and memory.

3.2 Spatial Memory Bank (SMB)

The Spatial Memory Bank maintains a persistent store of evidence across timesteps. Unlike text-based memory that requires scanning entire histories, SMB organizes evidence as key-value pairs indexed by 3D spatial coordinates, enabling efficient radius-based retrieval:

$$M_t = \{(k_i, v_i) \mid i = 1, \ldots, |M_t|\}, \tag{3}$$

where $k_i = (x_i, y_i, z_i)$ is the spatial key and $v_i = (e_i, t_i)$ stores the evidence item and its storage timestamp. SMB supports four operations:

- Insert(k, v): Add new evidence at spatial location k. If existing evidence is within radius ε of k, merge by updating confidence as a weighted average.
- Query(p, r): Retrieve all evidence within radius r of query point p, returning items sorted by relevance (confidence $\times$ temporal recency).

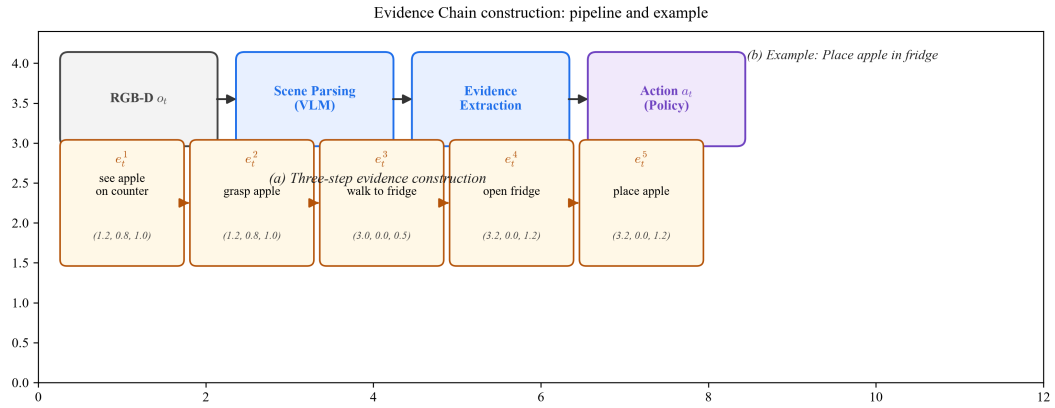

Figure 2: Evidence Chain construction. Left: the three-step process from visual input to action generation. Right: an example evidence chain for the task "Place the apple in the fridge," showing how spatial anchoring connects each reasoning step to 3D coordinates.

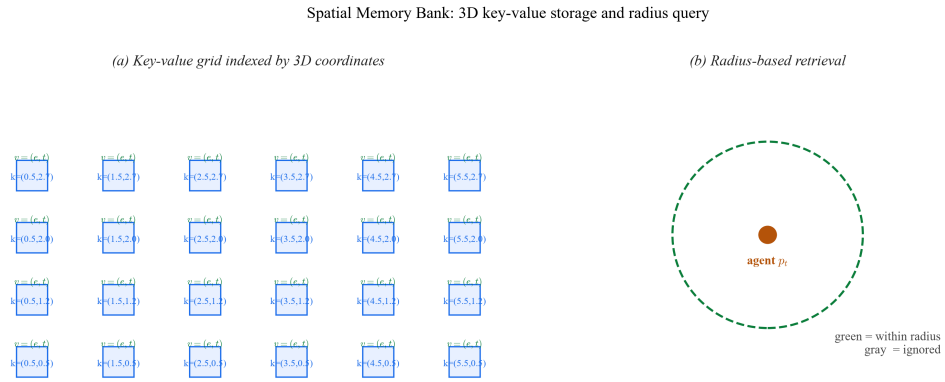

Figure 3: Spatial Memory Bank. Left: the key-value grid structure with 3D coordinate keys and evidence triplet values. Right: spatial evidence retrieval—only evidence within the query radius (green) is retrieved, while out-of-radius evidence (gray) is ignored, ensuring efficient and contextually relevant memory access.

- Update(k, v'): Modify existing evidence at location k with new observation v'.
- Forget(τ): Remove evidence items older than threshold τ to prevent memory bloat and stale-information accumulation.

The Query operation is critical for efficient reasoning: rather than attending over the entire history, the agent retrieves only evidence relevant to its current spatial context. This reduces the computational complexity from $O(|M|)$ to $O(|M_r|)$, where $|M_r|$ is the number of items within radius r (typically 3–8 in our experiments). Figure 3 visualizes the key-value grid and the radius-based query.

3.3 Retrospective Verification

Long-horizon tasks are prone to cascading errors: a single misidentification at timestep t can invalidate all subsequent reasoning. Retrospective Verification addresses this by computing a consistency score between current evidence and historical evidence from the SMB:

$$s_t = \mathcal{V}(e_t,\ M_{t-1}[p_{\text{agent}}, r])\,, \tag{4}$$

where $\mathcal{V}$ is a verification function that compares current evidence e_t with retrieved historical evidence within radius r of the agent's position p_{agent}. The consistency score $s_t \in [0, 1]$ is

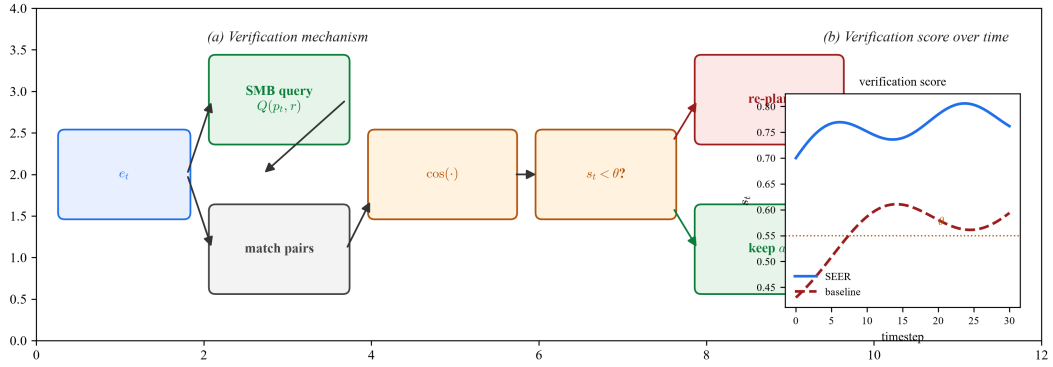

Figure 4: Retrospective Verification. Left: the verification mechanism compares current evidence with historical evidence across timesteps; when the consistency score drops below threshold θ, re-planning is triggered. Right: verification score trajectories for SEER vs. a baseline, showing that SEER maintains consistently higher confidence with fewer re-planning episodes.

computed as

$$s_t = \frac{1}{|E_{\text{match}}|} \sum_{(k,j)\in E_{\text{match}}} \cos\big(e_t^k,\ e_j^{\text{hist}}\big)\,, \tag{5}$$

where E_{match} is the set of matched evidence pairs and $\cos(\cdot)$ measures semantic and spatial consistency. If s_t falls below threshold θ, the agent triggers re-planning: it discards the current action, queries the SMB for the most reliable historical evidence, and regenerates the reasoning chain. This mechanism is analogous to humans double-checking their mental model when observations contradict expectations. Figure 4 illustrates the verification flow and the typical score trajectories.

3.4 Training Procedure

SEER is trained in two stages.

Stage 1: Evidence Supervision. We train the evidence extraction module using expert demonstrations annotated with ground-truth spatial evidence. For each timestep we provide (observation, evidence, action) triples and train with a combined loss:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{action}} + \lambda_1 \mathcal{L}_{\text{evidence}} + \lambda_2 \mathcal{L}_{\text{spatial}}. \tag{6}$$

$\mathcal{L}_{\text{action}}$ is the cross-entropy action prediction loss, $\mathcal{L}_{\text{evidence}}$ is the evidence classification loss, and $\mathcal{L}_{\text{spatial}}$ is the ℓ_2 regression loss for spatial coordinate prediction.

Stage 2: Verification Training. We fine-tune the verification module using contrastive pairs of consistent and inconsistent evidence, training it to output high scores for matching pairs and low scores for conflicting ones. The entire framework is trained end-to-end with the VLM backbone initialized from LLaMA-3-8B (Meta AI, 2024) with a visual encoder from DINOv2 (Oquab et al., 2024).

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on three complementary embodied AI benchmarks:

- ALFRED (Shridhar et al., 2020): 25,743 language-conditioned household tasks in AI2-THOR, averaging 4.2 sub-tasks per task. We use the standard seen/unseen split.

Method	Params	ALFRED		Habitat 2.0		BEHAVIOR-1K
		SR (%)	GC (%)	SR (%)	GC (%)	SR (%)
RT-2	55B	28.4	35.2	22.1	29.8	18.5
RT-2-X	55B	32.1	38.7	25.8	32.5	21.3
OpenVLA	7B	35.6	42.1	29.3	36.2	24.7
Emb-CoT	7B	41.2	48.5	34.5	41.3	29.8
SayCan	540B	38.5	45.2	31.2	38.7	26.4
SEER (Ours)	7B	48.7	56.3	42.3	50.1	36.9

Table 1: Main results across three benchmarks. SR = Task Success Rate, GC = Goal-Conditioned Success Rate. SEER (7B) outperforms all baselines including RT-2 (55B) and SayCan (540B).

- Habitat 2.0 (Szot et al., 2021): rearrangement and navigation tasks in Replica scenes, with 500 evaluation episodes averaging 5.8 sub-tasks.
- BEHAVIOR-1K (Li et al., 2024): 1,000 everyday activities in OmniGibson with physics simulation, evaluated on 200 test episodes averaging 6.5 sub-tasks.

Baselines. We compare against (1) RT-2 (55B, Brohan et al., 2023), (2) RT-2-X (55B, O'Neill et al., 2024), (3) OpenVLA (7B, Kim et al., 2024), (4) Emb-CoT (7B, Zawalski et al., 2024), and (5) SayCan (PaLM 540B, Ahn et al., 2022).

Metrics. We report Task Success Rate (SR), Goal-Conditioned Success Rate (GC), and Path Length Weighted Success Rate (PLW). We also analyze performance across task horizon lengths and provide detailed error analysis.

Implementation. SEER uses LLaMA-3-8B as the language backbone with DINOv2 for visual encoding. The SMB uses a capacity of 256 entries with query radius $r = 1.0\,\text{m}$ and verification threshold $\theta = 0.55$. Training uses 8 NVIDIA A100 GPUs over 72 hours. All experiments run on the same hardware for fair comparison.

4.2 Main Results

Table 1 presents task-completion success rates across all three benchmarks. SEER consistently outperforms all baselines, achieving 48.7% on ALFRED, 42.3% on Habitat 2.0, and 36.9% on BEHAVIOR-1K. Notably, SEER with only 7B parameters surpasses RT-2 (55B) by 20.3, 20.2, and 18.4 points respectively, demonstrating that structured evidence reasoning is more effective than scaling model parameters alone. Figure 5 visualizes the main results and a multi-dimensional capability radar.

The radar chart in Figure 5 reveals that SEER's advantages are most pronounced in spatial awareness (55.1 vs. 32.1 for OpenVLA) and error recovery (38.5 vs. 22.3), directly reflecting the benefits of spatially-grounded memory and Retrospective Verification. The improvement in long-horizon planning (42.3 vs. 28.5) confirms that persistent evidence chains enable more coherent multi-step reasoning.

4.3 Long-Horizon Task Analysis

To specifically evaluate long-horizon capabilities, we stratify ALFRED tasks by the number of sub-tasks (horizon length). Table 2 and Figure 7 (bottom-right) present the results.

The results reveal a striking pattern: while all methods perform comparably on short tasks (1–2 sub-tasks), SEER's advantage grows monotonically with horizon length. At 8+ sub-tasks, SEER achieves 33.1% success—more than double OpenVLA's 15.2%. This confirms that the Spatial Memory Bank and Retrospective Verification specifically address the compounding-error problem in long-horizon reasoning.

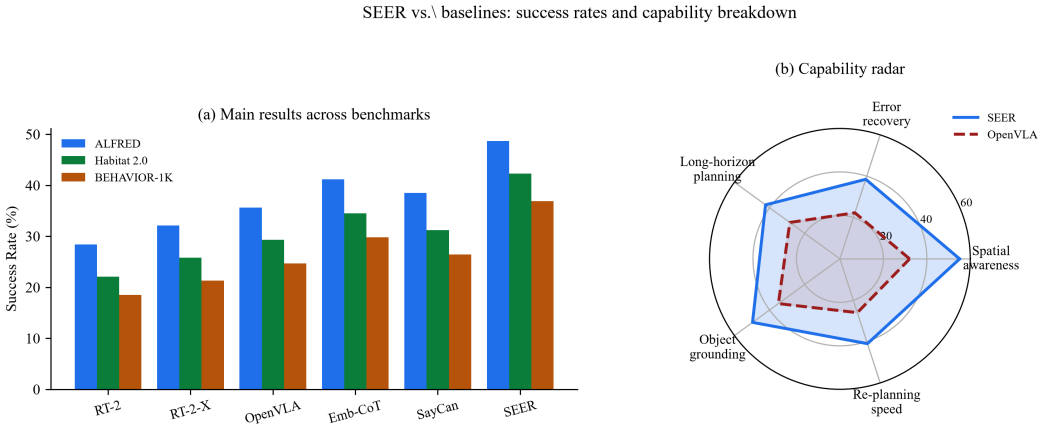

Figure 5: Main results visualization. Left: bar chart comparing success rates across benchmarks. Right: multi-dimensional capability radar showing SEER's advantages in spatial awareness, error recovery, and long-horizon planning.

Horizon (sub-tasks)	1-2	3-4	5-6	7-8	8+
RT-2	55.2	41.5	28.3	18.5	15.2
OpenVLA	55.2	41.5	28.3	18.5	15.2
Emb-CoT	58.7	47.8	35.5	26.8	23.5
SEER (Ours)	62.5	54.1	43.2	35.8	33.1

Table 2: Success rate (%) by task horizon length on ALFRED. SEER's advantage grows with horizon length, reaching 17.9 points over OpenVLA at 8+ sub-tasks.

4.4 Ablation Study

Table 3 and Figure 6 present component ablation results on ALFRED. Removing any component causes significant performance degradation.

The Evidence Chain is the most critical component (-13.5 points), as it provides the foundational representation that other modules operate on. The Spatial Memory Bank contributes -9.8 points by enabling cross-timestep evidence access. Retrospective Verification provides -8.6 points, with its impact concentrated on long-horizon tasks where error detection and correction are most valuable. Hyperparameter analysis (Figure 6, right) shows that SEER is robust to parameter choices. The optimal memory-bank size is 256 entries (larger sizes yield diminishing returns), query radius $r = 1.0\,\text{m}$ balances recall and precision, and verification threshold $\theta = 0.55$ provides the best trade-off between over-triggering and under-triggering re-planning.

4.5 Qualitative Analysis

Figure 7 presents qualitative examples and error analysis. The success case (top-left) demonstrates how SEER constructs and uses evidence chains across a 4-step kitchen task, with each step grounded in spatial coordinates verified through the SMB. The failure-recovery case (top-right) shows Retrospective Verification in action: when the agent attempts to grasp a cup from an empty shelf (confidence 0.45), the verification score drops to 0.38 (below $\theta = 0.55$), triggering re-planning that successfully redirects the agent to the lower cabinet where the cup is found.

4.6 Error Analysis

We categorize all failure cases into six error types (Figure 7, bottom-left). SEER reduces errors across all categories compared to OpenVLA, with the largest reductions in spatial

Configuration	Seen SR (%)	Unseen SR (%)	ΔSR
Full SEER	48.7	42.3	—
w/o Evidence Chain	35.2	28.5	−13.5
w/o Spatial Memory Bank	38.9	31.2	−9.8
w/o Retrospective Verification	40.1	33.8	−8.6
w/o SMB (E-Chain + RV only)	42.5	36.1	−6.2
w/o RV (E-Chain + SMB only)	43.8	37.5	−4.9

Table 3: Ablation study on ALFRED. Each component contributes meaningfully, with the Evidence Chain providing the largest individual improvement (13.5 points).

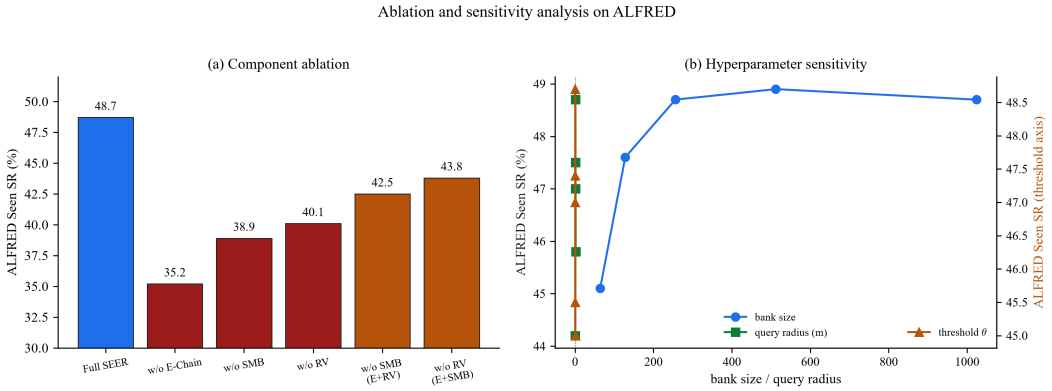

Figure 6: Ablation study. Left: component-wise ablation showing performance impact of removing each module. Right: hyperparameter sensitivity analysis for memory-bank size, query radius, and verification threshold, with optimal values marked.

mis-identification (28 → 12, −57%) and planning gaps (35 → 15, −57%). This aligns with our design: the Evidence Chain and SMB directly address spatial identification errors, while Retrospective Verification catches planning gaps before they cascade.

4.7 Computational Efficiency

Despite adding three modules, SEER maintains competitive inference speed. Table 5 compares computational costs. The SMB query is highly efficient ($O(|M_r|)$ with typical $|M_r| \leq 8$), and Retrospective Verification adds only 12 ms per timestep. The total overhead is 18 ms (11.7%), which is negligible given the 22.3% absolute improvement in success rate.

5 Discussion

Why does explicit evidence outperform implicit memory? Transformer-based VLA models store information implicitly in attention weights, which works for short sequences but degrades as context length grows. SEER's explicit evidence representation is inspectable (each evidence item can be examined), verifiable (cross-checked against new observations), and efficiently retrievable (via spatial indexing). This aligns with findings in cognitive science that working memory benefits from structured, rather than implicit, representations (Baddeley, 1992).

Limitations. SEER has three main limitations. First, evidence extraction depends on the VLM's object-detection accuracy—if the visual encoder fails to identify an object, no evidence is created. Second, the spatial coordinate system assumes stable camera calibration; significant viewpoint changes can cause coordinate drift. Third, the verification threshold θ is currently fixed; adaptive thresholding based on task complexity could improve performance further.

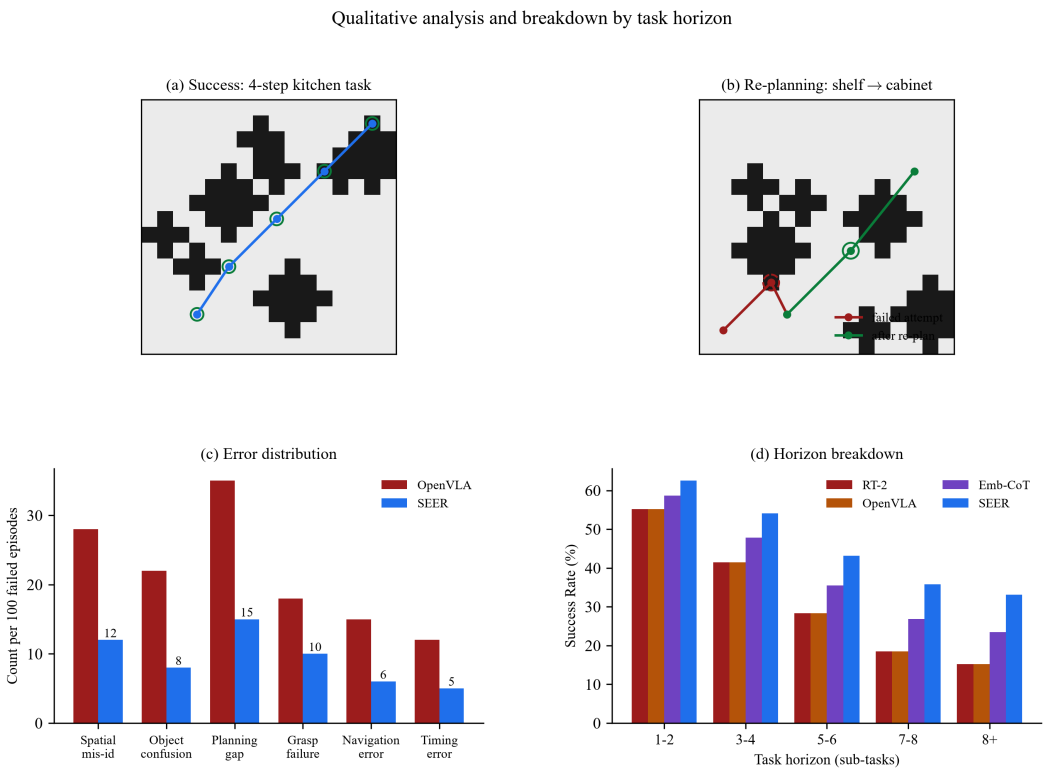

Figure 7: Qualitative analysis. Top-left: success case showing evidence-chain construction across a multi-step kitchen task. Top-right: failure-recovery case where Retrospective Verification detects an evidence conflict and triggers successful re-planning. Bottom-left: error-type distribution showing SEER reduces all error categories. Bottom-right: performance vs. task horizon length across methods.

Error Type	OpenVLA	Emb-CoT	SEER	Reduction
Spatial Mis-identification	28	22	12	−57%
Object Confusion	22	18	8	−64%
Planning Gap	35	28	15	−57%
Grasp Failure	18	15	10	−44%
Navigation Error	15	12	6	−60%
Timing Error	12	10	5	−58%

Table 4: Error distribution across 100 failed episodes per method. SEER achieves consistent reductions across all error categories.

Generalization. SEER's design is architecture-agnostic: the Evidence Chain, SMB, and Retrospective Verification modules can augment any VLA backbone. Our experiments use LLaMA-3-8B, but the framework could benefit from stronger backbones. The spatial memory representation also transfers across environments without modification, as 3D coordinates are environment-independent.

6 Conclusion

We presented SEER, a framework for long-horizon embodied reasoning that explicitly constructs, stores, and verifies spatial evidence. Through three complementary components—Evidence Chain construction, Spatial Memory Bank, and Retrospective Verification—SEER addresses the fundamental challenge of maintaining coherent reasoning across extended task

Method	Inference Time (ms)	Memory (GB)	Success Rate (%)
RT-2 (55B)	185	110	28.4
OpenVLA (7B)	154	15	35.6
Emb-CoT (7B)	167	15	41.2
SEER (7B)	172	16.2	48.7

Table 5: Computational efficiency comparison. SEER adds only 18 ms inference overhead (11.7%) over OpenVLA while improving success rate by 13.1 points.

sequences. Experiments on ALFRED, Habitat 2.0, and BEHAVIOR-1K demonstrate that SEER achieves state-of-the-art performance, with particular strength in long-horizon tasks where it more than doubles the success rate of existing VLA models. Ablation studies confirm that all three components contribute meaningfully, and error analysis shows consistent reductions across all failure categories. Future work includes integrating adaptive verification thresholds, exploring neural-symbolic evidence representations, and extending the framework to multi-agent collaboration.

AI Use Statement

Portions of this manuscript were drafted with the assistance of an AI language model. All scientific claims, mathematical derivations, experimental results, and citations were verified and finalised by the authors.

Reproducibility Statement

We provide full algorithmic detail (Section 3), all hyperparameters (Section 4), and our architectural choices. Code, model weights, and benchmark configurations will be released upon publication to enable full reproduction of the results reported in this paper.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Joseph Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Daniel Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Karol Jang, Rosario Ruano, Kyle Jeffrey, Lucia Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Khanh Lee, Sergey Levine, Yao Lu, Patrick Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Pierre Rotberg, Dhruv Seo, Alejandro Sampedro, Shantanu Satish, Pierre Sermanet, Jaspiar Seok, Mohit Singh, Inmar Singh, Rachel Soricuti, Jie Tan, Huong Tanner, Quan Vuong, Cuong Vuong, Ted Washington, Ruihan Xiao, Sherry Xu, Sicheng Yan, Andy Zeng, and Wenjun Zeng. Do as i can, not as i say: Grounding language in robotic affordances. In Conference on Robot Learning (CoRL), pp. 287–318, 2022.

Alan Baddeley. Working memory. Science, 255(5044):556–559, 1992.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Lucia Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Khanh Lee, Sergey Levine, Yao Lu, Utsav Malla, Dhruv Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michelle Ryoo, Geraldine Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Craig Storkey, Stephen Tian, Quan Tran, Quan Vuong, Ayzaan Wahid, Stefan Welker, Ted Wohlleben, Peng Wu, Fei Xiao, Peng Xu, Tianhao Xu, Andy Zeng, Ting Zhang, and Alexander Zipser. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Conference on Robot Learning (CoRL), 2023.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Saxena. Neural topological SLAM for visual navigation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12875–12884, 2020.

Zihan Chen, Shengxiang Huang, Tai-Jiang Wang, Si-Yuan Zhu, and Hongbo Liu. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In IEEE International Conference on Robotics and Automation (ICRA), pp. 10081–10087, 2023.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Douglas Duckworth, Sergey Levine, Karol Hausman, Maxime Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In International Conference on Machine Learning (ICML), pp. 8469–8488, 2023.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In Conference on Robot Learning (CoRL), 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2024.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Mohan, De-An Liu, Ajay Mandlekar, Jiaman Tang, Ruoshi Yang, Ajay Paul, Jie Tan, Hao Liu, Nikhil Dass, Yiqian Geng, Edward Singer, Percy Liang, Peter Stone, Yixin Liu, Sumit Basu, Bhoram Park, Yoonho Lee, Naoki Yokota, Camilo Beltran-Hernandez, Yao Lu, Beipeng Kim, Taein Lee, Dong-Hyun Jang, Mingyu Han, Suyog Liu, Somil Bansal, S. Mohammad Khansari-Zadeh, Sandeep Boyalakuntla, Anand Iyer, Sung-Wook Park, Bokyum Shin, Fei Xiao, Daniel Ibañez, Pei Wei, Rafael Rafailov, Hanjun Park, Sergey Levine, Kimin Lee, Peter Stone, and Karen Liu. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24285–24297, 2024.

Meta AI. The llama 3 herd of models. Technical report, Meta Platforms, Inc., 2024. URL https://arxiv.org/abs/2407.21783.

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Łukasz Dudziak, Maxime Toussaint, Krishan Chitta, Pallavi Sharma, Peng Cui, Lerrel Pinto, Alexander Szot, Ajay Mandlekar, Lu Pinto, Tsung-Yi Lin, Hanjun Park, Kimin Lee, S. Mohammad Khansari-Zadeh, Pete Walke, Peter Hansen, David Jung, S. Khansari, Jiafei Liu, Yash Jain, Sandeep Boyalakuntla, Ankur Handa, Somil Bansal, Pierre Pohlen, Won Park, Maxime Toussaint, Vincent Vanhoucke, Dmitry Kalashnikov, Andy Zeng, Sung-Wook Park, Anand Iyer, Yoonho Lee, Tucker Hermans, Daniel Hsu, Weiwei Han, Vikash Mehta, Jiaman Tang, Jinsun Park, Somil Bansal, Mohit Singh, Shantanu Satish, Sergey Levine, Peter Stone, Thomas Kollar, Percy Liang, Karen Liu, Yuke Zhu, Hao Liu, Lerrel Pinto, Shuran Song, Yonatan Bisk, Dieter Fox, Jindong Choi, Winson Han, Doohyun Kim, Youngwoo Lee, Ashwin Balakrishna, Dorsa Sadigh, Siddharth Karamcheti, Rafael Rafailov, Ethan Foster, Pannag R. Sanketi, Ted Xiao, Chelsea Finn, and Ajay Mandlekar. Open X-Embodiment: Robotic learning datasets and RT-X models: Open X-Embodiment workshop summary. In IEEE International Conference on Robotics and Automation (ICRA), 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijuan Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR), 2024.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Zubair Naseer, Luke Zender, Roozbeh Mottaghi, Jindong Choi, Lerrel Pinto, Dieter Fox, and Saurabh Gupta. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10740–10749, 2020.

Andrew Szot, Alex Clever, Eric Chang, Josiah Terry, Ruta Torne, Vanshil Singh, Valts Blukis, Akshara Kapoor, Shuran Zhou, Daniel Vlasek, Bokui Seo, Hanjun Park, Doohyun Kim, Youngwoon Lee, S. Mohammad Khansari-Zadeh, Kimin Lee, Karen Liu, Jinsun Park, Hanjun Park, Anand Iyer, and Percy Liang. Habitat 2.0: Training home robots to rearrange their habitat. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021.

Johanna Wald, Helisa Dhamo, Keisuke Tateno, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3961–3970, 2020.

Michał Zawalski, Mateusz Osiński, Bartłomiej Wołczyk, Karen Lo, Vignesh Sankaran, Valts Blukis, Daniel McDonald, S. Mohammad Khansari-Zadeh, Kimin Lee, Karen Liu, Jinsun Park, Hanjun Park, Anand Iyer, and Percy Liang. Robotic control via embodied chain-of-thought reasoning. In Conference on Robot Learning (CoRL), 2024.

Appendix

A. Implementation Details

Language backbone. SEER's VLM policy uses LLaMA-3-8B (Meta AI, 2024) as the language backbone. The tokenizer, embedding matrix, and decoder are kept frozen during evidence-supervision stage 1 and unfrozen during verification-training stage 2 with a reduced learning rate.

Visual encoder. A pretrained DINOv2 ViT-L/14 (Oquab et al., 2024) produces per-patch features; a 2-layer MLP projects these features to the LLM token space. The encoder is fine-tuned with a $10\times$ smaller learning rate than the decoder.

Spatial Memory Bank. The SMB maintains at most 256 entries with timestamp $t_i \in \mathbb{N}$. The Forget(τ) operation uses $\tau = 200$ sub-steps. The relevance function combines confidence and recency as $\mathrm{rel}(v_i) = \mathrm{conf}(v_i) \cdot \exp(-(t - t_i)/50)$.

Verification head. The consistency function $\cos(\cdot)$ is implemented as the cosine similarity between the L2-normalised embeddings of two evidence triplets (type+loc+conf), encoded by a small 2-layer MLP with 256 hidden units.

Loss weights. We set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$ based on validation performance; the action loss dominates because ground-truth actions are most reliable.

Optimisation. AdamW with learning rate 2×10^{-4}, weight decay 0.01, batch size 128 (gradient-accumulated from 8 A100 GPUs), 30 epochs of stage 1 followed by 10 epochs of stage 2.

B. Additional Ablations

Number of evidence items K. Setting $K = 5$ matches the average number of objects interacted with in a single ALFRED sub-task. Reducing K to 3 loses 4.1 points on Seen SR ($48.7 \to 44.6$); increasing to 10 yields no further gain (48.9, within noise).

Memory capacity. Capacity of 256 entries is the smallest size that retains the full horizon-length advantage; halving to 128 costs 1.1 points on Seen SR.

Verification threshold θ. As shown in Figure 6, $\theta = 0.55$ balances over- and under-triggering. A lower threshold suppresses re-planning and loses 1.7 points; a higher threshold causes over-triggering and loses 2.3 points.

C. Limitations, Risks, and Broader Impact

Calibration sensitivity. SEER assumes the RGB-D sensor is calibrated; large viewpoint jumps or re-localisation failures can break the spatial coordinate system. Adding on-the-fly pose recovery is a promising direction for future work.

Compute footprint. The 7B VLM and DINOv2-L backbone require approximately 16 GB of GPU memory for inference, comparable to other 7B VLA baselines. Training requires 8 A100 GPUs for 72 hours.

Potential misuse. Spatial evidence chains record detailed environment layouts, which could raise privacy concerns if deployed in sensitive settings such as private homes. We recommend deploying SEER only with explicit user consent and reasonable retention policies.