DELTA-MAP BELIEF UPDATES FOR STABLE SPATIAL REVISION IN VISION-LANGUAGE MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Vision-language models can extract spatial information from images but struggle to maintain and revise spatial beliefs over time. Existing approaches either regenerate cognitive maps from scratch, losing valuable prior context, or regenerate fully, which is wasteful when changes are sparse. We propose delta-map updates, a sparse belief revision mechanism that preserves unchanged spatial beliefs while selectively updating only elements affected by new observations. On the Theory of Space benchmark, providing prior map context with explicit preserve/overwrite rules improves false-belief identification F1 by +16.7 percentage points over scratch regeneration. Delta-map updates achieve equivalent performance to full regeneration (F1 = 0.479 vs 0.477) while producing 52–63% smaller structured outputs. Our analysis validates the sparse evidence premise: only $\sim$30% of objects require updating per step, supporting the efficiency of targeted updates over full regeneration.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Spatial understanding is fundamental to embodied AI systems, from household robots to navigation assistants. Recent vision-language models (VLMs) have demonstrated impressive capabilities in extracting spatial information from images (Chen et al., 2024; Cheng et al., 2024), yet a critical challenge remains: how should these models *maintain and revise* spatial beliefs over time as environments change?

The Theory of Space benchmark (Zhang et al., 2026) isolates this challenge by evaluating whether foundation models can construct coherent spatial beliefs through active exploration. The benchmark reveals severe limitations: even strong VLMs exhibit high belief inertia, failing to update their spatial representations when objects move or rotate. A key question emerges: is this failure due to fundamental perception limitations, or does the *interface* by which beliefs are externalized contribute to the problem?

We hypothesize that repeatedly regenerating complete cognitive maps from long observation histories introduces transcription errors and inconsistent conflict resolution. When most observations only affect a small subset of objects, full regeneration is both wasteful and error-prone. This motivates our approach: **delta-map updates**, a sparse belief revision mechanism that preserves unchanged spatial beliefs while selectively updating only elements affected by new observations.

Our experiments on the Theory of Space benchmark validate this hypothesis. Providing prior map context with explicit preserve/overwrite rules improves false-belief identification F1 by +16.7 percentage points over scratch regeneration. Delta-map updates achieve equivalent performance to full regeneration (F1 = 0.479 vs 0.477) while producing 52–63% smaller structured outputs. Analysis confirms the sparse evidence premise: only $\sim$30% of objects require updating per step.

This work makes three contributions. First, we propose a delta-map belief update interface for VLMs that treats cognitive maps as external state and updates them through sparse delta operations.

[1] https://gitlab.com/fars-a/tos-delta-map-updates

Second, we provide empirical validation on the Theory of Space benchmark demonstrating that prior context dramatically improves change detection (+54% relative improvement). Third, we present evidence that sparse updates match full regeneration performance with substantial efficiency gains, validating the sparse evidence premise for spatial belief revision.

2 Related Work

Spatial Reasoning in Vision-Language Models. Recent advances in multimodal large language models have enabled impressive spatial understanding capabilities. SpatialVLM (Chen et al., 2024) endows VLMs with spatial reasoning through large-scale synthetic data generation, enabling quantitative spatial understanding such as distance estimation and size comparison. SpatialRGPT (Cheng et al., 2024) extends this by grounding spatial reasoning in region-level representations, achieving more precise localization. Yang et al. (2024) investigate how MLLMs perceive, remember, and recall spatial information, revealing fundamental limitations in spatial memory. These works focus primarily on spatial *extraction* from static scenes, whereas our work addresses the distinct challenge of spatial *belief revision* over time.

Cognitive Maps and Spatial Memory. Cognitive maps provide structured representations of spatial environments that support navigation and reasoning (Stoewer et al., 2023). The Theory of Space benchmark (Zhang et al., 2026) evaluates whether foundation models can construct spatial beliefs through active exploration, using structured JSON cognitive maps to represent object positions, orientations, and facing directions. Vision-to-geometry approaches (Cai et al., 2025) leverage 3D spatial memory for embodied reasoning, while VLM-based navigation systems (Goetting et al., 2024) transform spatial reasoning into question-answering. Our work builds on the cognitive map representation from Theory of Space but focuses specifically on how to efficiently update these representations when environments change.

Belief Revision and State Tracking. Understanding how language models maintain and update internal beliefs is an active research area. Li et al. (2025) investigate mechanisms by which LLMs track state across sequences, while Herrmann & Levinstein (2024) propose standards for evaluating belief representations in LLMs. In embodied settings, OpenEQA (Majumdar et al., 2024) evaluates question answering that requires maintaining beliefs about environments. Our delta-map approach contributes to this literature by proposing sparse, targeted belief updates that preserve unchanged information while selectively revising only affected elements—a strategy motivated by the observation that most spatial changes are sparse.

3 Method

3.1 Problem Formulation

We consider the problem of spatial belief revision in vision-language models. An agent explores a multi-room environment through a sequence of observations $O_1, O_2, \ldots, O_T$, where each observation O_t consists of first-person visual images and associated metadata. The agent maintains a cognitive map M_t representing its spatial beliefs about object locations and orientations. At each timestep t, given the prior map M_{t-1} and new observation O_t, the agent must produce an updated map M_t that integrates new information while preserving unchanged beliefs.

The key challenge is *belief revision*: when the environment changes (objects move or rotate), the agent must detect these changes and update its beliefs accordingly. This is particularly difficult because (1) the agent cannot observe all objects simultaneously, (2) new observations may conflict with prior beliefs, and (3) repeatedly regenerating the entire map from scratch can introduce transcription errors.

3.2 Cognitive Map Representation

Following the Theory of Space benchmark (Zhang et al., 2026), we represent cognitive maps as structured JSON objects containing spatial beliefs about observed entities:

$$M = \{e_i : (\mathbf{p}_i, d_i) \mid e_i \in \mathcal{E}\} \tag{1}$$

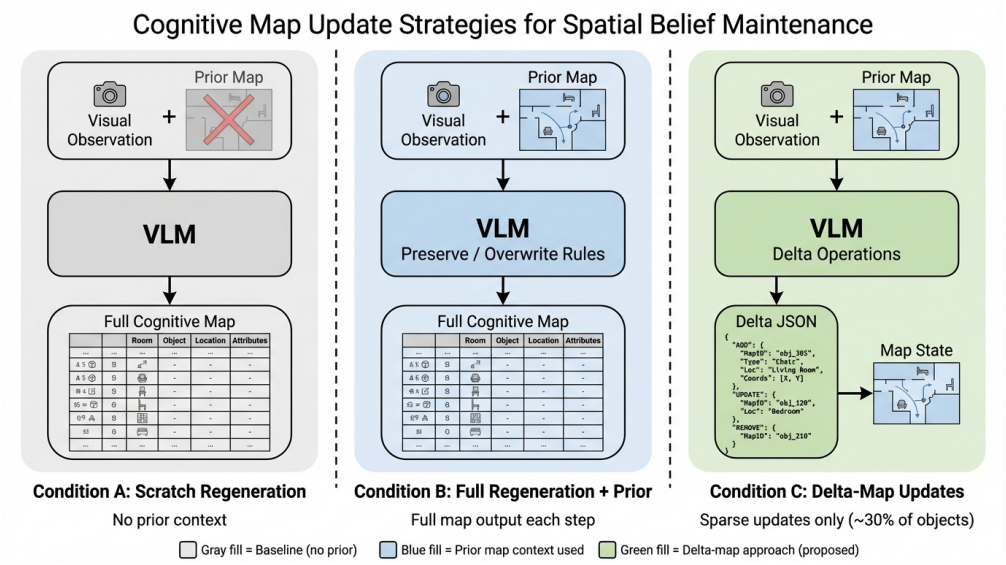

Figure 1: Delta-Map belief update framework for spatial revision in VLMs. The system maintains a structured JSON cognitive map that is updated through sparse delta operations, preserving unchanged spatial beliefs while selectively revising only the elements affected by new visual observations.

where $\mathcal{E}$ is the set of observed entities (agent, objects, gates), $\mathbf{p}_i = (x_i, y_i)$ is the position, and $d_i \in \{\text{north}, \text{south}, \text{east}, \text{west}\}$ is the facing direction. This structured representation enables precise evaluation of spatial beliefs across three dimensions: positional accuracy, directional accuracy, and facing accuracy.

3.3 Belief Update Conditions

We compare three approaches to cognitive map updates, illustrated in Figure 1:

Condition A: Scratch Regeneration. The baseline approach regenerates the complete cognitive map from the full observation history at each timestep. Given observations $O_1, \ldots, O_t$, the model outputs a complete map M_t without access to prior beliefs. This mirrors the original Theory of Space evaluation protocol.

Condition B: Full Regeneration with Prior. The model receives the prior map M_{t-1} alongside the current observation O_t and explicit update rules: (1) *preserve* all entries from M_{t-1} unchanged unless current observation provides contradicting evidence, (2) *restrict updates* to objects visible in O_t, and (3) *overwrite* entries that conflict with current observation. The model outputs a complete updated map M_t.

Condition C: Delta-Map Updates. The model receives the same inputs and rules as Condition B, but outputs only a sparse delta Δ_t containing changed entries:

$$\Delta_t = \{e_i : (\mathbf{p}'_i, d'_i) \mid e_i \text{ changed}\} \tag{2}$$

The updated map is computed programmatically: $M_t = \text{Apply}(M_{t-1}, \Delta_t)$. This approach reduces output length and isolates the model's generation to the minimal set of updates, potentially reducing transcription errors.

3.4 Evaluation Metrics

We evaluate belief updates using three complementary metrics:

Table 1: False-belief revision performance across update conditions. Higher F1 indicates better change detection; lower inertia indicates better belief revision. Best results per metric in **bold**.

Condition	ID F1 ↑	Pos F1 ↑	Ori F1 ↑	Pos Inertia ↓	Ori Inertia ↓
A (Scratch)	0.310 ± 0.042	0.168 ± 0.024	0.144 ± 0.031	$\mathbf{0.516 \pm 0.053}$	$\mathbf{0.222 \pm 0.073}$
B (Full Regen)	0.477 ± 0.019	$\mathbf{0.284 \pm 0.024}$	$\mathbf{0.299 \pm 0.029}$	0.563 ± 0.038	0.320 ± 0.076
C (Delta-Map)	$\mathbf{0.479 \pm 0.019}$	0.255 ± 0.025	0.274 ± 0.029	0.608 ± 0.015	0.220 ± 0.058

Identification F1. For false-belief revision, we measure how well the model identifies which objects have changed. Precision captures whether reported changes are correct; recall captures whether actual changes are detected.

Belief Inertia. Following Theory of Space, we measure the tendency to retain outdated beliefs. Positional inertia quantifies how often the model maintains incorrect positions for changed objects; orientation inertia measures the same for facing directions. Lower inertia indicates better belief revision.

Cognitive Map Accuracy. We evaluate the final map M_T against ground truth across three dimensions: positional accuracy (object locations), directional accuracy (pairwise spatial relationships), and facing accuracy (object orientations). The overall score is the weighted average of these components.

4 Experiments

4.1 Experimental Setup

We evaluate our delta-map belief update approach on the Theory of Space benchmark (Zhang et al., 2026), which tests spatial belief construction and revision in multi-room environments. The benchmark provides 25 scenes (3-room layouts) with pre-rendered first-person images and ground-truth object positions. Each scene involves 8–12 exploration steps using a SCOUT-like trajectory with 360-degree sweeps at each location.

For false-belief revision, the benchmark modifies $k = 4$ objects per scene (position or orientation changes) after initial exploration. The agent must re-explore using post-change images and update its cognitive map accordingly. We use Gemini 2.0 Flash (Team et al., 2025) as the base model with temperature 0.5 for Conditions B and C, and temperature 1.0 for Condition A (matching the original benchmark protocol).

4.2 False-Belief Revision Results

Table 1 presents results on the false-belief revision task, which evaluates how well models detect and revise beliefs about changed objects.

Providing prior map context with explicit preserve/overwrite rules (Conditions B and C) dramatically improves false-belief identification compared to scratch regeneration (Condition A). The overall identification F1 increases from 0.310 to 0.477–0.479, representing a +16.7 percentage point improvement (54% relative gain). This demonstrates that explicit prior context helps models detect environmental changes more reliably.

Delta-map updates (Condition C) achieve equivalent identification performance to full regeneration (Condition B), with F1 scores of 0.479 versus 0.477. However, the two approaches show different trade-offs in belief revision. Condition B achieves lower positional inertia (0.563 vs 0.608) but higher orientation inertia (0.320 vs 0.220). Notably, delta-map updates uniquely preserve low orientation inertia (0.220), matching the baseline (0.222), while full regeneration shows elevated orientation inertia (0.320). This suggests that targeted updates better preserve fine-grained spatial information.

Table 2: Cognitive map probing accuracy across update conditions. Results show mean accuracy ± standard error. Best results per metric in **bold**.

Condition	Overall	Position	Direction	Facing
A (Scratch)	0.219 ± 0.014	0.305 ± 0.017	0.112 ± 0.010	0.239 ± 0.027
B (Full Regen)	0.225 ± 0.008	**0.341 ± 0.011**	**0.126 ± 0.008**	0.207 ± 0.021
C (Delta-Map)	**0.236 ± 0.011**	0.335 ± 0.010	0.119 ± 0.008	**0.253 ± 0.026**

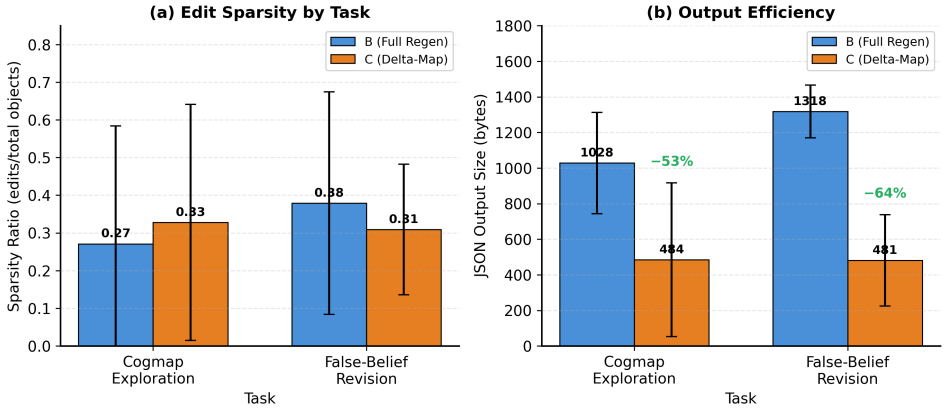

Figure 2: Edit magnitude analysis comparing full regeneration (Condition B) and delta-map updates (Condition C). Left: Sparsity ratio showing proportion of objects updated per step. Right: JSON output size demonstrating efficiency gains of delta-map approach (52.9% reduction for cogmap, 63.5% reduction for false-belief tasks).

Despite improvements in change detection, belief inertia remains high across all conditions (positional inertia 0.52–0.61), indicating that VLMs still struggle to actually revise their spatial beliefs even when they detect changes. This represents a fundamental challenge in spatial belief revision that warrants future investigation.

4.3 Cognitive Map Probing Results

Table 2 presents results on cognitive map probing, which evaluates the accuracy of spatial beliefs at the end of exploration.

Delta-map updates (Condition C) achieve the highest overall accuracy (0.236), outperforming both full regeneration (0.225) and scratch regeneration (0.219). The improvements are modest but consistent, with delta-map updates showing particular strength in preserving facing information (0.253 vs 0.207–0.239). Full regeneration (Condition B) achieves the best positional and directional accuracy, suggesting that complete map regeneration may help maintain spatial relationships. All conditions show similar performance ranges, reflecting the inherent difficulty of the spatial reasoning task.

4.4 Edit Magnitude Analysis

Figure 2 validates the sparse evidence premise underlying delta-map updates. During exploration, only 32.8% of objects are updated per step on average (4.75 out of 15.0 objects), and during false-belief revision, this ratio is 30.9% (4.73 out of 15.4 objects). This confirms that most spatial updates are sparse, supporting the efficiency of targeted delta updates over full regeneration.

The efficiency gains are substantial: delta-map JSON outputs are 52.9% smaller than full map outputs during exploration (484 vs 1028 bytes) and 63.5% smaller during false-belief revision (481 vs 1318 bytes). While raw response sizes are similar due to reasoning text, the structured output reduction demonstrates that delta-map updates achieve equivalent task performance with significantly more compact belief representations.

5 CONCLUSION

We introduced delta-map updates, a sparse belief revision mechanism for spatial reasoning in vision-language models. By treating cognitive maps as external state and updating only changed entries, delta-map updates achieve equivalent false-belief identification performance to full regeneration (F1 = 0.479 vs 0.477) while producing 52–63% smaller structured outputs. Our experiments validate the sparse evidence premise: only $\sim$30% of objects require updating per step, making targeted updates more efficient than full regeneration.

However, belief inertia remains high across all conditions (positional inertia 0.52–0.61), indicating that VLMs struggle to actually revise their spatial beliefs even when detecting changes. This fundamental challenge in belief revision warrants future investigation into the underlying mechanisms by which VLMs maintain and update internal representations.

REFERENCES

Zhongyi Cai, Yi Du, Chen Wang, and Yu Kong. Vision to geometry: 3d spatial memory for sequential embodied mllm reasoning and exploration. *ArXiv*, abs/2512.02458, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *ArXiv*, abs/2406.01584, 2024.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *ArXiv*, abs/2411.05755, 2024.

Daniel A. Herrmann and B. A. Levinstein. Standards for belief representations in llms. *Minds and Machines*, 35, 2024.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? *ArXiv*, abs/2503.02854, 2025.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, S. Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent-Pierre Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and A. Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024.

Paul Stoewer, Achim Schilling, Andreas K. Maier, and Patrick Krauss. Multi-modal cognitive maps based on neural networks trained on successor representations. *ArXiv*, abs/2401.01364, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2024.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Fei-Fei Li, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? 2026.