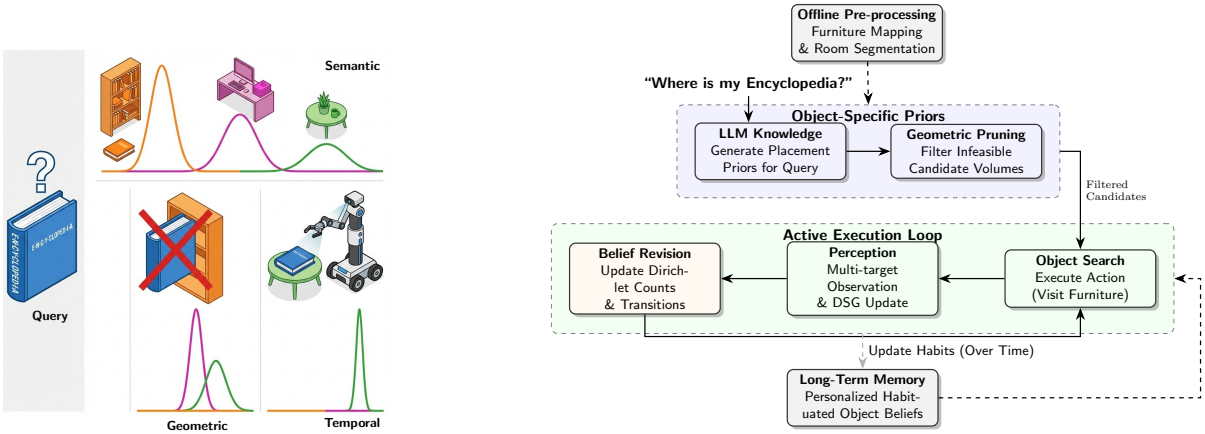
O-STaR: Effective Object Search through Spatio-Temporal Reasoning on Dynamic Scene Graphs

Rohit Menon Yasmin Schmiede Gokul Krishna Chenchani Hermann Blum* Maren Bennewitz*



(a) Proposed Spatio-Temporal Reasoning: Semantic → Geometric → Temporal

(b) Active Search Pipeline: Initialization → Search Loop → Personalization

Fig. 1. **O-STaR Framework Overview.** Our approach enables personalized object search by bridging conceptual reasoning (a) with a robot active search pipeline (b). When queried for encyclopedia's location (a), LLM priors suggest a bookshelf or desk. Geometric grounding prunes the undersized bookshelf, and physical observations during the system loop (b) trigger a temporal belief update, shifting search toward a learned household habit: the coffee table. Our framework unifies offline processing of static furniture and room segmentation as scene graph anchors, open-vocabulary perception for dynamic scene graph updates, and long-term memory refinement using temporal belief adaptation. (Fig 1 (a) generated using Gemini Nano Banana's generative modeling)

I. Introduction

Service robots must reliably locate objects that are relocated or concealed in dynamic households [1]. Many existing systems rely on static representations and fixed semantic assumptions, which often fail under long-term variability [2], [3].

Recent approaches leverage Large Language Models (LLMs) to provide common-sense knowledge about object placement [4]. While effective as generic priors, such knowledge reflects an *average user behavior* and fails to capture the personalized routines of individual households. For example, while generic priors might associate an encyclopedia with a bookshelf or desk, geometric constraints might rule out an undersized shelf, while temporal observations reveal a unique household habit: keeping the volume on a coffee table.

In this work, we present **O-STaR** (illustrated in Fig. 1), an integrated reasoning framework that combines semantic common sense reasoning, geometric feasibility, and adaptive temporal belief updates over dynamic scene graphs [5].

*Equal Supervision. All authors are with the University of Bonn, Germany, and are further affiliated with the Lamarr Institute and the Center for Robotics, Bonn. R. Menon and M. Bennewitz are also with the Humanoid Robots Lab. H. Blum is also with the Robot Perception and Learning Lab. This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob, and by the BMBF within the Robotics Institute Germany, grant No. 16ME0999.

II. Related Work

Object search has advanced through semantic priors, spatial memory, and belief planning [6]. While recent modular systems use open-vocabulary perception and scene graphs to generalize to novel objects [7], [8], they often rely on population-level LLM priors that lack personalization [9], [4]. DynaMem [10] explores dynamic adaptation, but lacks integration with geometric constraints. In contrast, O-STaR unifies semantic, geometric, and temporal reasoning to enable personalized search.

III. O-STaR: Unified Reasoning Framework

O-STaR represents the search for a target object as a sequential belief update process over an incrementally constructed open-vocabulary 3D semantic scene graph [2], initialized through offline RGB-D scanning and open-vocabulary segmentation [11], [12]. The scene graph explicitly models the hierarchical relationship between rooms, furniture, compartments, and objects. We employ a YOLO-based drawer detector [13] to partition storage volumes into distinct nodes annotated with centroids and axis-aligned bounding boxes to ground search in physical feasibility. This structure enables O-STaR to perform unified reasoning across semantic, geometric, and temporal dimensions.

A. Semantic–Geometric Reasoning for Physical Search

O-STaR leverages common-sense LLM knowledge strictly for Day-0 initialization, seeding the initial object-location

belief across the household. This semantic prior is grounded in physical feasibility by filtering candidate compartments through volume exclusion tests in the 3D [2]. By ensuring that the target object physically fits within proposed furniture volumes, O-STaR prevents the exploration of semantically likely but geometrically impossible locations. Once initialized, our pipeline performs active search, where the belief is progressively refined through physical observations and temporal adaptation.

B. Temporal Belief Adaptation and Personalization

To personalize search over time, we maintain a Dirichlet-Categorical belief over an object's possible locations $L = \{l_1, \ldots, l_N\}$. The belief is initialized at Day-0 using an LLM-derived prior p_{LLM} with a total pseudo-count mass S_0:

$$\alpha_{0,i} = S_0 \cdot \frac{p_{LLM,i}}{\sum_j p_{LLM,j}} \tag{1}$$

When a location l_k is visited and the object is not detected (*negative observation*), we redistribute a fixed evidence weight w_{miss} to all other candidates:

$$\alpha_{t+1,i} = \begin{cases} \alpha_{t,i} + \frac{w_{miss}}{N-1} & \text{if } i \neq k \\ \alpha_{t,i} & \text{if } i = k \end{cases} \tag{2}$$

Conversely, a successful detection adds w_{hit} to the specific location.

To account for object dynamics between search episodes, we employ a 'Stay+Leak' transition model that diffuses the belief:

$$\alpha_{t+1} = (1-\gamma)\alpha_t + \gamma \frac{\sum_j \alpha_{t,j}}{N} \mathbf{1} \tag{3}$$

where γ is the leakage rate. This mechanism enables O-STaR to "unlearn" incorrect semantic priors through repeated negative evidence, shifting the search toward observed household patterns. This adaptive transition model enables systematic belief shift away from incorrect semantic priors toward household-specific patterns observed over time. We further employ **relaxed inference** to learn transitions from sparse observations, bridging the gap between population-level common sense and personal experience. Furthermore, the framework leverages opportunistic multi-target perception during search to observe non-target re-locations i.e., detecting previously queried objects while searching for a different target, thereby accelerating the convergence of the temporal belief toward true household habits.

IV. Experimental Evaluation

We evaluate O-STaR through a series of real-world trials and long-term simulation sweeps designed to isolate physical feasibility and cognitive robustness. Individual configurations (R1–R5) are compared across varying levels of temporal plasticity, initial prior corruption, and observation noise.

ID	Transition Model	Obs. Noise (p_{fn})	Corruption	Success Rate
R1	Frozen (Stay)	0.1	0%	74.67%
R2	**Adaptive (Stay+Leak)**	0.1	0%	**82.67%**
R3	Frozen (Stay)	0.1	30%	48.67%
R4	**Adaptive (Stay+Leak)**	0.1	30%	**72.67%**
R5	**Adaptive (Stay+Leak)**	0.5	0%	**82.67%**

TABLE I

Object search success rates over 300 simulated episodes for different belief transition models, observation noise levels (p_{fn}), and degrees of prior corruption.

A. Real-World Validation of Semantic–Geometric Reasoning

In **44** real-world trials on a Stretch mobile manipulator, our geometric reasoning reduced concealed-space search time by **68%** compared to exhaustive search. Throughout these trials, the robot successfully navigated the environment and interacted with various storage compartments. Furthermore, we successfully detected **80%** of objects hidden within furniture components such as drawers and cabinets. These physical experiments demonstrate that O-STaR successfully grounds semantic hypotheses in real-world geometric constraints.

B. Long-Term Simulation of Temporal Adaptation

To quantify long-term adaptation, we conducted 300-episode simulation sweeps using the HOMER+ dataset [14], which captures multi-day object relocation patterns across three dynamic households. The results, summarized in Tab. IV-B, indicate that **temporal adaptation is essential for robustness** in dynamic environments. An adaptive model (R2) achieved a success rate of **82.67%**, outperforming a stationary belief model (R1: 74.67%). Notably, the adaptive model (R4) maintained a **72.67%** success rate under **30%** prior corruption, whereas the frozen model (R3) degraded to 48.67%. These results demonstrate that temporal adaptation is critical: the adaptive model not only improves baseline performance but provides resilience, recovering **72.67%** success from **30%** prior corruption through systematic belief redistribution. Furthermore, we observed a **synergy between perception and adaptation**: using multi-target perception to opportunistically observe non-target objects provided a **13%** absolute increase in success rate under corrupted priors. Performance remained stable even under high observation noise ($p_{fn} = 0.5$) when adaptation was enabled (R5).

V. Conclusion

O-STaR demonstrates that effective long-term object search requires reasoning that is personalizable over time through active belief management. By explicitly combining semantic priors, geometric feasibility, and temporal adaptation on dynamic scene graphs, with probability redistribution enabling adaptation from sparse experience, our framework enables long-term, personalized object search in dynamic human environments.

References

[1] A. J. Miller, "Understanding clutter: Geographies of everyday homes and objects," Ph.D. dissertation, University of Leeds, 2018.

[2] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[3] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint*, 2024.

[4] D. Qiu, W. Ma, Z. Pan, H. Xiong, and J. Liang, "Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps," *arXiv preprint*, 2024.

[5] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From slam to spatial perception with 3d dynamic scene graphs," *Intl. Journal of Robotics Research (IJRR)*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[6] K. Zheng, A. Paul, and S. Tellex, "A system for generalized 3d multi-object search," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*. IEEE, 2023.

[7] O. Lemke, Z. Bauer, R. Zurbrügg, M. Pollefeys, F. Engelmann, and H. Blum, "Spot-compose: A framework for open-vocabulary object retrieval and drawer manipulation in point clouds," *arXiv preprint*, 2024.

[8] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. W. Clegg, J. Turner, *et al.*, "Homerobot: Open-vocabulary mobile manipulation," *arXiv preprint*, 2023.

[9] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[10] P. Liu, Z. Guo, M. Warke, S. Chintala, N. M. M. Shafiullah, and L. Pinto, "Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation," *arXiv preprint*, 2024.

[11] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[12] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*. IEEE, 2023.

[13] T. Engelbracht, R. Zurbrügg, M. Pollefeys, H. Blum, and Z. Bauer, "Spotlight: Robotic scene understanding through interaction and affordance detection," *arXiv preprint*, 2024.

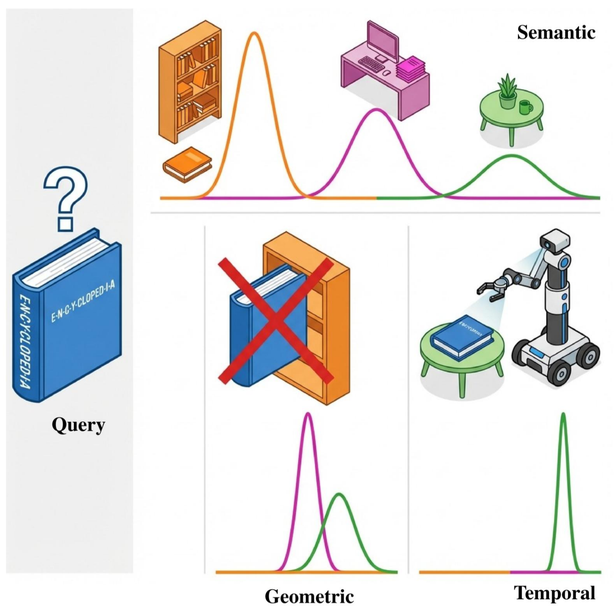
[14] M. Patel, A. G. Prakash, and S. Chernova, "Predicting routine object usage for proactive robot assistance," in *Proc. of Conf. on Robot Learning (CoRL)*, 2023.

O-STaR: Open-Vocabulary Object Search through Spatio-Temporal Reasoning on Dynamic Scene Graphs

Rohit Menon, Yasmin Schmiede, Gokul Chenchani,
Hermann Blum*, Maren Bennewitz*
University of Bonn, Lamarr Institute

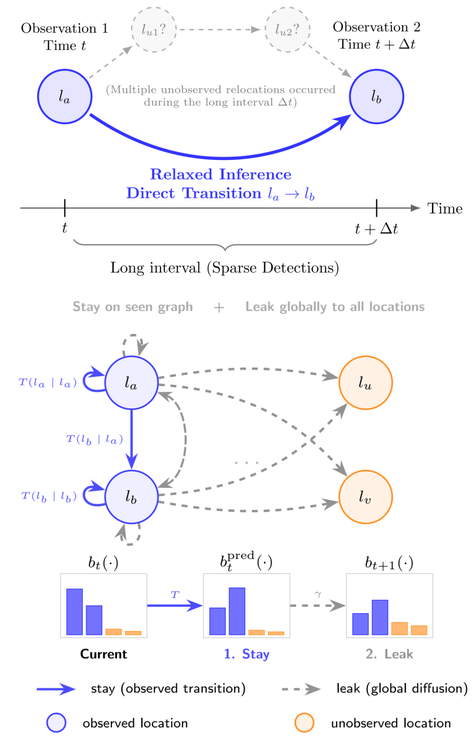
Overview

- **Goal:** Efficient **open-vocabulary object search** in **dynamic** homes
- **Problem: Outdated world models** from relocated or concealed objects
- **Challenge:** Integrate **semantic** priors, **geometric** feasibility, and **temporal** dynamics under **sparse episodic evidence**



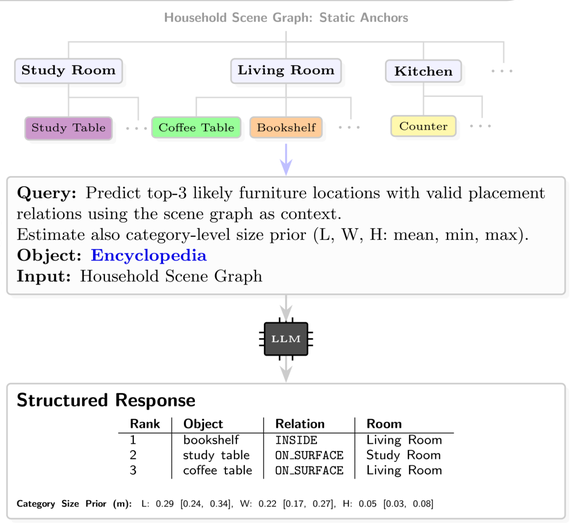
Temporal Adaptation

- **Relaxed inference** from sparse observations
- Direct **transition learning** across time gaps
- Propagate belief via learned transitions
- Leak probability globally to all locations
- **Gradual diffusion** prevents stale overconfidence
- **Personalize** search via **learned relocation** habits



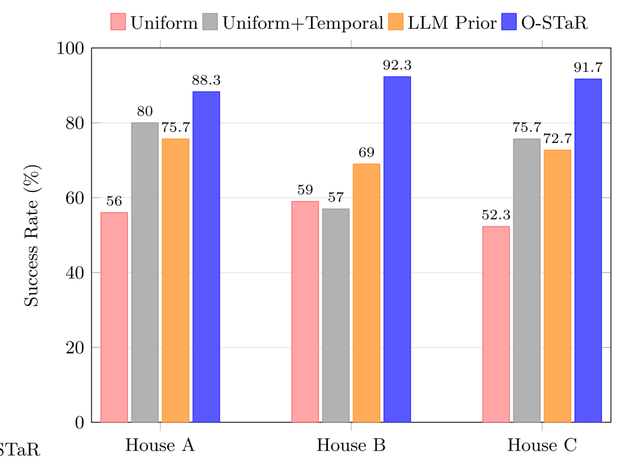
Semantic Priors from Scene Graph

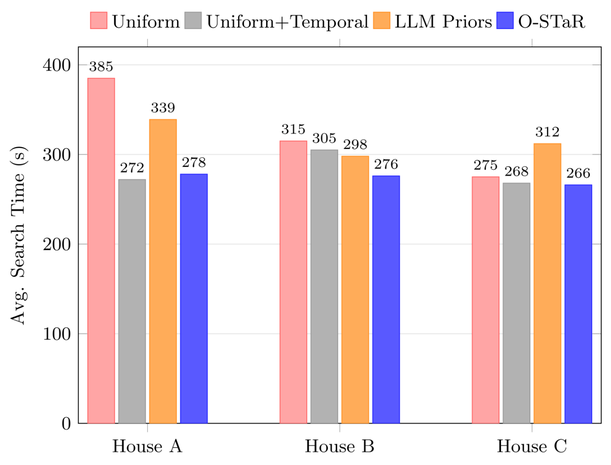
- Scene graph with room-furniture anchors
- LLM **commonsense placement** priors with ranking
- Category-level **object size** priors
- Rank-to-probability belief initialization



Experimental Results

- **HOMER+** dynamic household simulation dataset
- Sparse episodic observations across time
- **Baselines:** Uniform, Uniform+Temporal, LLM priors
- O-STaR: **Highest search success rate** across all households
- Temporal learning enables high success rates with uniform prior as well

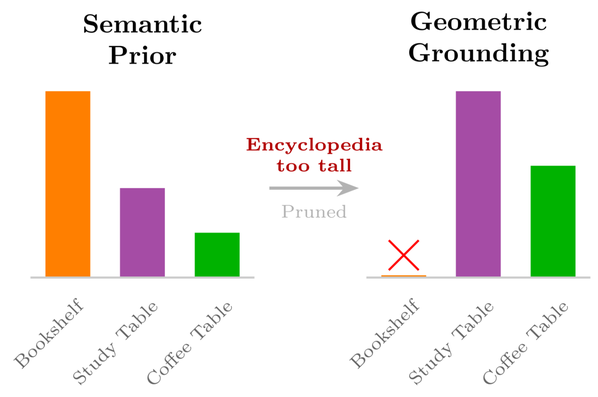




- Consistently **low search time** for O-STaR
- Geometric pruning prevents costly concealed storage interactions
- Cost-aware search policy enables efficient search

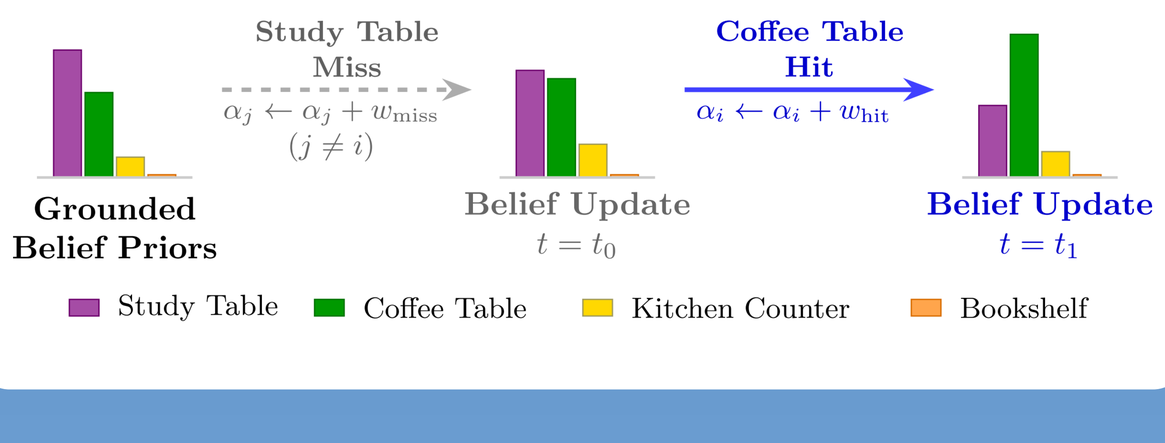
Semantic Grounding

- Check **object-compartment feasibility**
- Prune infeasible locations
- Refine belief over feasible candidates
- Avoid expensive articulated manipulations



Probabilistic Object Belief Update

- **Dirichlet–Categorical** location belief
- **Bayesian evidence update** from sparse observations
- **Uncertainty-aware** state estimation
- **Robust** to noisy observations



Summary

- **Geometric** grounding of LLM **semantic priors → Reduces** search space
- **Uncertainty-aware** Dirichlet belief over locations → Robust to noisy observations
- **Temporal** adaptation for **personalized** search
- Efficient **cost-aware** active search policy→ **Highest** success rate with **low** search times

Contact

Rohit Menon
menon@cs.uni-bonn.de
Humanoid Robots Lab
University of Bonn
Germany