Cognitive Computing in Sociotechnical Systems, Vol. 205, 2026, 159–168
https://doi.org/10.54941/ahfe1007370

Human-AI Co-Navigation for Indoor Object Search Under Uncertainty

Ahmed Ghita[1], Qiuyi Cao[2], Daniel Watzenig[2,3], and Stefan K. Ehrlich[1]

[1]SETLabs Research GmbH, Elsenheimerstrase 55, 80686 Muenchen, Germany
[2]Virtual Vehicle Research GmbH, Inffeldgasse 21a, 8010 Graz, Austria
[3]Institute of Visual Computing, Graz University of Technology, Inffeldgasse 16/2, 8010 Graz , Austria

ABSTRACT

Assistive technologies for people with visual impairments increasingly use artificial intelligence to support object-finding and navigation in indoor environments. Yet fully autonomous perception remains unreliable in such settings, as indoor spaces are visually complex, only partially observable from the user's current viewpoint, and subject to continuous change. Our work takes the position that effective assistive navigation is inherently collaborative; the system performs continuous perceptual processing, while the user provides occasional natural-language guidance when the search becomes uncertain or inefficient. To this end, we propose a human–AI collaboration framework that utilizes a Vision-Language Model (VLM) as the perceptual and semantic backbone of a navigation agent. A human user, modeled by a simulated intervention controller, provides sparse and structured guidance, which is integrated with the VLM to update its semantic search hypotheses toward the likely location of the target object. Evaluation is conducted in the Habitat simulator on photorealistic scenes from the Habitat-Matterport3D dataset. Experiments analyze how human guidance affects task success and navigation efficiency, showing that guidance is most effective when it corrects the VLM's misaligned semantic search hypotheses, providing insights into the role of minimal human input in VLM-based assistive navigation systems.

Keywords: Assistive technology, Human-AI collaboration, Vision-Language models, Indoor navigation

INTRODUCTION

Assistive technologies for people with visual impairments increasingly incorporate computer vision and AI systems that interpret visual scenes and translate them into actionable feedback for users navigating everyday environments (Dakopoulos and Bourbakis, 2010; Tapu et al., 2013). In practice, however, autonomous perception alone is often insufficient. Indoor environments are visually complex, partially observable, and continuously changing, making it difficult for an AI system to reliably locate a target object. Navigation in such settings frequently becomes a collaborative process in which a human user provides occasional guidance when uncertainty or failure risk increases. Figure 1 illustrates this framing.

Prior work in assistive technology for people with visual impairments has explored wearable and mobile systems for obstacle avoidance and scene-aware

Received March 26, 2026; Revised April 29, 2026; Accepted May 19, 2026; Available online July 20, 2026

support (Dakopoulos and Bourbakis, 2010; Tapu et al., 2013). These systems establish the practical importance of assistive perception, but they typically emphasize direct environmental feedback rather than collaborative indoor object-finding under partial observability. More recent language-guided navigation work such as SayNav shows the growing use of language-based planning for embodied navigation (Rajvanshi et al., 2024), but it is not developed for assistive object-finding by visually impaired users or for studying sparse human guidance as belief-updating evidence during search. In contrast, the present work focuses on lightweight human--AI collaboration for indoor object finding, where occasional natural-language guidance updates the agent's internal hypothesis without retraining the navigation backbone or introducing a heavier mapping or policy-learning stack.

Supervisory control frameworks describe policies in which an operator monitors autonomous behavior and intervenes selectively when task uncertainty rises (Goodrich and Schultz, 2007). In shared autonomy systems, human inputs bias action selection or update task beliefs without replacing the underlying policy (Javdani et al., 2015), and sparse feedback can guide agent behavior as probabilistic evidence shaping decision-making (Griffith et al., 2013). Indoor object-finding provides a natural application: a user can provide concise guidance such as *"It should be next to the bed"* without specifying a trajectory. The challenge is to incorporate such sparse natural-language guidance into an autonomous navigation system while preserving the simplicity of the underlying policy.

Recent advances in VLMs have enabled joint reasoning over visual observations and natural language (Radford et al., 2021; Li et al., 2023; Liu et al., 2023), with VLMs increasingly explored as perceptual backbones for embodied agents navigating complex environments (Majumdar et al., 2022; Shen et al., 2023). In the *object-goal navigation* task, an agent must find a target object category using only egocentric visual observations and a semantic goal (Anderson et al., 2018). VLMs can provide coarse semantic cues about object presence and approximate direction, but translating these signals directly into navigation actions can produce oscillatory turning, inefficient exploration, or repeated collisions (Chaplot et al., 2020).

Many approaches address these limitations through persistent memory representations, spatial maps, or learned navigation policies (Chaplot et al., 2020; Majumdar et al., 2022), but these increase system complexity and training requirements. For assistive applications, where systems may need to run on wearable or mobile hardware with limited latency and power budgets, interpretability and computational efficiency are critical, making lightweight alternatives desirable. We explore whether sparse human guidance can correct navigation behavior by directly influencing the agent's internal beliefs about target location, without modifying the navigation architecture. We model human assistance as a structured intervention policy emitting short natural-language guidance during navigation, adopting a three-type taxonomy: informational, expressive, and directive, grounded in Bühler's Organon model (Bühler, 1934) and speech-act theory (Searle, 1969). The agent maintains a belief distribution over scene landmarks derived from VLM semantic ranking; human guidance modifies this belief distribution through re-ranking rather than direct action control. We conduct an empirical study in the Habitat ObjectNav simulator across twenty photorealistic HM3D

validation scenes, analyzing whether guidance corrects the VLM's semantic prior, whether belief updates translate into improved navigation, and how robust this reasoning process is to imperfect guidance. All guidance in this study is generated by a simulated intervention controller rather than real participants; the goal is to isolate the causal effect of structured guidance on VLM belief updating and navigation under reproducible conditions. The key contributions of this work are threefold. First, we propose a lightweight human--AI collaboration framework for indoor object-finding in which sparse natural-language guidance updates a VLM-based agent's landmark belief without retraining the navigation backbone. Second, we introduce a controlled simulation methodology for studying sparse structured guidance under reproducible conditions using a finite-state intervention controller. Third, we provide an empirical evaluation on twenty HM3D validation scenes and introduce Interaction Efficiency (IE) to measure how effectively guidance is utilized after it is provided. Section 2 describes the framework, Section 3 the experimental setup, Section 4 the results, and Section 5 concludes.

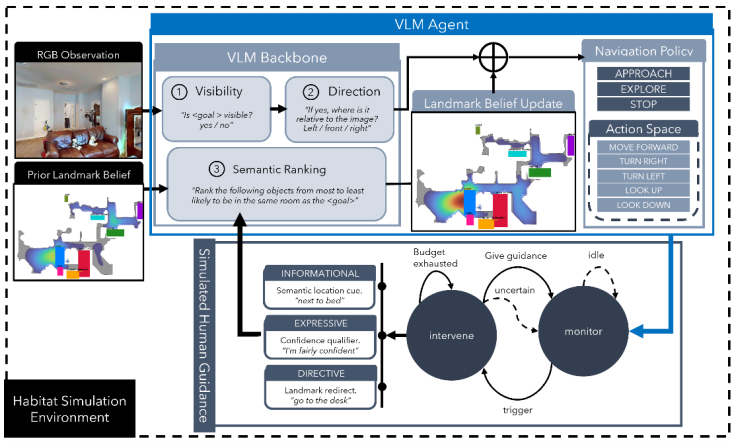

Figure 1: System overview of the human–AI collaboration framework. The VLM agent processes egocentric RGB observations through three structured queries: visibility detection, coarse direction estimation, and semantic landmark ranking, to maintain a landmark belief distribution and select navigation actions. A simulated human user monitors agent behavior and emits sparse structured guidance (informational, expressive, or directive) via the intervention controller, which triggers semantic re-ranking of candidate landmarks to update the agent's navigation hypothesis without modifying the underlying policy.

DEVELOPING A LIGHTWEIGHT HUMAN–AI COLLABORATION FRAMEWORK FOR INDOOR OBJECT-FINDING

Problem Formulation

We study collaborative indoor object-finding in an assistive setting, motivated by scenarios in which a visually impaired user seeks support in locating an everyday object in an unfamiliar or partially observed indoor space. Because real-user studies require substantial safety controls and introduce sources of variability that make early-stage mechanism analysis difficult, we conduct

this work as a controlled simulation study in Habitat. As shown in Figure 1, this setup isolates the interaction between the VLM navigation backbone, the landmark belief distribution, and the simulated human intervention policy under reproducible conditions. We formulate the task as *object-goal navigation* (ObjectNav), in which an agent must navigate to an instance of a specified object category using only egocentric visual observations and a semantic goal label. At each time step , the agent receives an RGB observation and selects a discrete action:

$$a_t \in \{\texttt{move\_forward}, \texttt{turn\_left}, \texttt{turn\_right}, \texttt{look\_up}, \texttt{look\_down}, \texttt{stop}\}.$$

The objective is to reach the target object within a bounded episode horizon while minimizing unnecessary exploration. An episode is successful if the agent issues *stop* within a predefined success radius of the goal.

VLM Navigation Policy

We use Qwen2-VL-2B-Instruct (Wang et al., 2024) as the VLM backend. As illustrated in Figure 1, the VLM supports three query functions that are used at different points in the navigation loop. At each step, it is asked whether the target object is currently visible in the RGB observation using a binary visibility prompt, for example:

"*Is the chair visible in this image? Answer only yes or no.*"

If the object is judged visible, the model predicts a coarse relative direction with respect to the camera view (e.g., left, front, right) using a directional query such as:

"*Where is the chair relative to the camera view? Answer with one word: left, front, or right.*"

Separately, the model receives a text-only ranking prompt over the set of scene landmarks to estimate semantic proximity between the goal category and potential proxy locations, for example:

"*Rank the following objects from most to least likely to be in the same location as a chair: bed, couch, kitchen cabinet, desk, bathtub. Return the ranked list.*"

The ranking is performed over the landmark set extracted from a scene-specific sparse prior map of the environment. In the present study, this prior map is precomputed once per HM3D scene from ground-truth semantic scene annotations and represented as a JSON landmark list containing category labels and associated navigable landmark positions. At episode start, this landmark vocabulary is loaded for the current scene, and the VLM ranks these landmarks conditioned on the goal category g, yielding the initial landmark distribution $b_0 = \text{VLMRank}(L, g)$. As shown in Figure 1, the same ranking mechanism is invoked again during exploration when progress stalls or when human guidance is received.

The navigation policy is a priority-ordered state machine with three modes:

1. **Stop:** If the agent is within the success distance of the goal, it issues *stop*.
2. **Approach:** If the goal is visible, the agent queries its coarse relative direction (left, front, right) and converts this into a turning or forward action; when sufficiently close, it uses a geometric controller for the final approach.
3. **Explore:** Otherwise, the agent selects the highest-ranked proxy landmark, $\ell_t^* = \arg\max_{\ell \in \mathcal{L}} b_t(\ell)$, and navigates towards it. If progress stalls, it refreshes the landmark distribution using a new ranking query and continues exploration.

Progress is considered stalled when the distance-to-goal does not decrease over the last k steps (here $k = 10$). In that case, the agent refreshes its belief by re-invoking $\mathbf{b}_{t+1} = \text{VLMRank}(\mathcal{L}, g)$, or if human guidance is available, applies guidance-conditioned re-ranking $\mathbf{b}_{t+1} = \text{VLMRank}(\mathcal{L}, g, u_t)$.

HUMAN GUIDANCE MODULE

Motivation and Communication Taxonomy

The collaboration layer shown in Figure 1 is motivated by the observation that failures of the baseline VLM agent are often evident at the behavioral level. High belief entropy indicates uncertainty over where to search low progress indicates that the agent is stuck or exploring inefficiently and repeated turning or proxy exhaustion indicate that current semantic guidance is not yielding useful progress. These are precisely the kinds of signals that, in supervisory interaction, justify selective human intervention (Goodrich and Schultz, 2007). To structure the content of human assistance, we adopt a three-part taxonomy grounded in established theories of language function. Bühler's Organon model distinguishes representational, expressive, and appellative functions of language (Bühler, 1934), and related distinctions are operationalized in speech act theory (Searle, 1969). We map these functions to three guidance types:

- **Informational guidance** provides semantic world knowledge about likely target location, e.g. *"It should be next to the bed."*
- **Expressive guidance** communicates confidence or uncertainty about preceding guidance, e.g. *"I am fairly confident."*
- **Directive guidance** instructs the agent to redirect behavior toward a named landmark, e.g. *"Go to the bathroom cabinet."*

In the present study, these guidance types are implementation choices within the simulated intervention controller rather than separate task conditions. Guidance is generated from a small set of predefined natural language templates selected according to the current interaction state and instantiated with the current landmark hypothesis where needed.

Intervention Policy

Human guidance is modeled as a lightweight intervention policy that monitors the agent's internal navigation context and emits structured guidance only when needed. Formally, the human user is modeled as a finite-state controller:

$$H = (S, s_0, \pi_H, E, B),$$

where S is the set of interaction states, s_0 is the initial monitoring state, π_H is the transition policy, E maps the current interaction state to a guidance type and template utterance, and B is the maximum number of guidance interventions allowed per episode. The controller observes a compact context vector consisting of belief entropy, progress over recent steps, proxy failures, turning oscillation, normalized episode progress, and budget remaining. These quantities are computed directly from the agent's current belief distribution and recent navigation history. Based on these signals, the controller alternates between monitoring and intervention states. In the deterministic version, high uncertainty with low progress triggers informational guidance, while repeated failure or late-stage urgency triggers directive guidance. An uncertain variant introduces stochastic transitions, occasional landmark noise, and reliability variation to model imperfect human input. The deterministic and uncertain variants are intended to span the space between ideal and noisy human communicators, not to replace empirical user behavior.

Semantic Re-Ranking Under Human Guidance

As depicted in Figure 1, human guidance influences navigation by conditioning the VLM's landmark re-ranking over the candidate set $\mathrm{L} = \{\ell_1, \ldots, \ell_N\}$. The agent maintains a landmark hypothesis as a categorical distribution $\mathrm{b}_t \in \Delta^{N-1}$. At episode start, this belief is initialized by ranking the landmarks in the sparse prior map with respect to the goal category, $\mathbf{b}_0 = \mathrm{VLMRank}(\mathrm{L}, g)$. When guidance u_t arrives, it is appended to the ranking prompt, yielding an updated distribution $\tilde{\mathbf{b}}_t = \mathrm{VLMRank}(\mathrm{L}, g, u_t)$, which replaces $\mathbf{b}_t$ for subsequent proxy selection. In all cases, the core mechanism is the same: guidance is fused into the landmark-ranking prompt as additional language context, and the updated ranking is then used to revise the agent's current search hypothesis. The different guidance forms described above differ only in the content of the added text, for example by providing location evidence, expressing confidence, or specifying a landmark-level redirect.

Table 1: Evaluated conditions. All agents share the same zero-shot Qwen2-VL-2B-Instruct navigation backbone; only the collaboration layer differs.

Tag	Condition	Description
C1	Vanilla VLM	Unassisted baseline without human guidance.
C2	Fixed guidance	Single informational guidance using an oracle landmark hint, injected once at a uniformly random step during the episode.
C3	FSM (deterministic)	Deterministic finite-state intervention controller that monitors the agent state and emits informational, expressive, or directive guidance according to predefined trigger rules.
C4	FSM (uncertain)	Stochastic variant of the FSM controller with noisy transitions, occasional landmark noise, and sampled reliability to model imperfect human guidance.

EXPERIMENTAL SETUP AND EVALUATION

Experiments are conducted in Habitat using the Habitat-Matterport3D (HM3D) validation split containing 20 validation scenes (Ramakrishnan et al., 2021). Episodes are drawn from a fixed multi-goal benchmark with six object categories: `bed`, `chair`, `plant`, `sofa`, `toilet`, and `tv_monitor`. The navigation backbone is Qwen2-VL-2B-Instruct used zero-shot without finetuning. We compare four agent conditions (tagged C1–C4) that share the same VLM navigation backbone and differ only in the collaboration layer (Table 1). C1 is the unassisted baseline. C2 injects a single informational guidance at a uniformly random step and uses an oracle landmark selector, where the oracle landmark is the landmark geodesically closest to the goal object from the shared vocabulary. C3 uses a deterministic finitestate machine (FSM) intervention policy with up to three interventions per episode. C4 uses a stochastic FSM variant with transition noise, landmark noise, and sampled reliability, modeling imperfect human input. In C3 and C4, the intervention budget is $B = 3$ per episode. The uncertain FSM in C4 uses transition noise $\epsilon = 0.15$, landmark noise $\eta = 0.20$, and sampled reliability. Within the FSM-based conditions, the different guidance forms are used as part of the intervention policy and are not evaluated as separate experimental conditions; therefore, the main comparison in this study is at the level of collaboration strategy rather than individual guidance type. We evaluate on a total of 1,817 episodes per condition. Episodes are limited to 200 steps. For each scene, episodes are held fixed across conditions to ensure direct comparability. Moreover, both the agent and the simulated human controller access the same precomputed landmark vocabulary derived from the scene's semantic annotations.

We evaluate both task performance and collaboration mechanism. Standard ObjectNav metrics are Success Rate (SR), Success weighted by Path Length (SPL), mean path length, and mean number of steps. To isolate post-intervention behavior, we define *Interaction Efficiency* (IE) as SPL recomputed over the trajectory segment after the guidance step, thereby removing exploration cost incurred before assistance arrives. We further report three belief-level metrics: entropy reduction proxy redirect rate, and pre-guidance oracle rank. The pre-guidance oracle rank is the rank assigned to the oracle landmark within the agent's belief distribution at the moment guidance arrives; a high rank indicates that the agent had deprioritized the correct search target before the intervention. These metrics measure whether guidance changes what the agent believes and whether this change is behaviorally adopted.

RESULTS AND DISCUSSION

Table 2 summarizes aggregate performance across all 20 HM3D validation scenes. Relative to the unassisted baseline (C1), all collaborative conditions improve both success rate and SPL. The best aggregate condition is C2, which increases SR from 78.8% to 87.8% and SPL from 0.637 to 0.764, while reducing mean path length from 15.06m to 10.57m and mean steps from 68.9 to 44.6. C3 and C4 remain close behind, indicating that sparse guidance is effective at scale and that the collaboration mechanism is robust to moderate uncertainty in the guidance channel.

Table 2: Aggregate navigation performance across all 20 HM3D validation scenes.

Condition	N	SR (%)	SPL	Path(m)	Steps
C1	1817	78.8	0.637	15.06	68.9
C2	1817	87.8	0.764	10.57	44.6
C3	1817	87.0	0.707	12.51	54.9
C4	1817	86.6	0.705	12.54	55.2

Table 3: Belief-level collaboration metrics over guidance episodes. Redundancy denotes the fraction of episodes in which the oracle landmark was already ranked first before the guidance.

Cond.	N	ΔH	Redirect (%)	Pre-Guidance Rank	Redundancy (%)
C2	170	0.029	**100.0**	4.49	7.1
C3	182	0.027	0.0	4.70	2.7
C4	187	0.034	9.6	4.42	9.1

Several limitations bound the scope of these findings. Human guidance in this study is generated by a simulated intervention controller rather than real users; whether the observed interaction properties generalize to actual human users, including users with visual impairments, remains to be established. Additionally, the prior landmark map is precomputed from ground-truth scene semantics, an assumption that would require a semantic mapping stage in a deployed system. Future work should extend evaluation to studies with real human participants, and investigate how human guidance interacts with the VLM's internal semantic knowledge during landmark re-ranking. Specifically, what linguistic properties of guidance drive belief correction, and how the model weighs human-provided evidence against its own spatial priors when updating the landmark belief distribution.

CONCLUSION

This paper studied how sparse simulated human guidance affects a zero- shot VLM navigation agent in assistive indoor object-finding. Across 20 HM3D validation scenes, all collaborative conditions improved over the unassisted baseline, showing that lightweight language-based collaboration can substantially improve navigation without retraining the backbone or adding a heavier mapping architecture. The results further indicate that the main value of guidance lies in correcting mis-ranked semantic search hypotheses, and that its effectiveness depends on how quickly those belief updates influence action selection. Overall, the study supports sparse human–AI collaboration as a viable design direction for assistive indoor object-finding systems.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the COMET K2 Competence Centers for Excellent Technologies programme, funded by the Austrian Federal Ministry for Climate Action (BMK), the Austrian Federal

Ministry for Labour and Economy (BMAW), the Province of Styria (Department 12), and the Styrian Business Promotion Agency (SFG). Programme management is carried out by the Austrian Research Promotion Agency (FFG). OpenAI GPT-5 was used for text refinement and the generation of tables.

REFERENCES

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A. (2018) "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments". https://arxiv.org/abs/1711.07280

Bühler, K. (1934). Sprachtheorie: Die Darstellungsfunktion der Sprache. Gustav Fischer, Jena.

Chaplot, D.S., Gandhi, D., Gupta, S., Salakhutdinov, R. (2020) "Object goal navigation using goal-oriented semantic exploration", in: Advances in Neural Information Processing Systems (NeurIPS).

Dakopoulos, D., Bourbakis, N.G. (2010) "Wearable obstacle avoidance electronic travel aids for blind: A survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(1), 25–35.

Goodrich, M.A., Schultz, A.C. (2007) "Human–robot interaction: A survey", Foundations and Trends in Human–Computer Interaction 1(3), 203–275.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C., Thomaz, A. (2013) "Policy shaping: Integrating human feedback with reinforcement learning", in: Advances in Neural Information Processing Systems (NeurIPS).

Javdani, S., Glas, D.F., Kollar, T., Srinivasa, S.S. (2015) "Shared autonomy via hindsight optimization", in: Robotics: Science and Systems (RSS).

Li, J., Li, D., Savarese, S., Hoi, S.C. (2023) "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", in: Proceedings of the International Conference on Machine Learning (ICML).

Liu, H., Li, C., Wu, Q., Lee, Y.J. (2023) "Visual instruction tuning", in: Advances in Neural Information Processing Systems (NeurIPS).

Majumdar, A., Aggarwal, R., Chen, Y., Chen, J., Bisk, Y., Batra, D., Kembhavi, A. (2022) "Vlmaps: Building scalable visual-language maps for robot navigation", in: Proceedings of the Conference on Robot Learning (CoRL).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021) "Learning transferable visual models from natural language supervision", in: Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763.

Rajvanshi, A., Sikka, K., Lin, X., Lee, B., Chiu, H.-P., Velasquez, A. (2024) "Saynav: Grounding large language models for dynamic planning to navigation in new environments". https://arxiv.org/abs/2309.04077

Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y., Batra, D. (2021) "Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI", in: NeurIPS Datasets and Benchmarks Track.

Searle, J.R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

Shen, W., Yang, R. et al. (2023) "Vlm-nav: Vision-language models for object goal navigation", in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Tapu, R., Mocanu, B., Bursuc, A. and Zaharia, T. (2013), A smartphonebased obstacle detection and classification system for assisting visually impaired Human–AI Co-Navigation for Indoor Object Search under Uncertainty 11 people, in '2013 IEEE International Conference on Computer VisionWorkshops' pp. 444–451.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J. and Lin, J. (2024), 'Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution', arXiv preprint arXiv:2409.12191 .