InfoNav: A Unified Value Framework Integrating Semantic Relevance and Information Gain for Zero-Shot Object Goal Navigation

Zhi Yang[1] **Chengpu Yu**[1]* **Gonzalo Ferrer**[2]
[1]School of Automation, Beijing Institute of Technology, Beijing, China
[2]Applied AI Institute
3220241334@bit.edu.cn, yuchengpu@bit.edu.cn, g.ferrer@applied-ai.ru

Abstract

Zero-shot object goal navigation (ObjectNav) requires an embodied agent to locate a target object in an unseen environment without any task-specific training. Recent methods leverage large vision–language models (VLMs) to inject semantic priors, but still suffer from two structural issues: (i) semantic cues are treated uniformly across space, ignoring how the spatial extent and hierarchy of contextual entities modulate target likelihood; and (ii) semantic exploitation and geometric exploration are decoupled, producing brittle behavior whenever semantic cues are weak, conflicting, or absent. We address these issues by reformulating navigation as a *unified value estimation* problem over frontier candidates. Specifically, we propose **InfoNav**, which jointly models (1) a hierarchical semantic value map that instantiates a spatially-weighted total-probability decomposition of target likelihood through a multi-level spatial influence field; (2) a VLM-based estimator that approximates the intractable long-horizon information gain by grounding it in visual–semantic connectivity; and (3) a memory-augmented verification mechanism that replaces fixed confidence thresholds with a temporally adaptive schedule with deferred re-examination of uncertain detections. InfoNav establishes the highest zero-shot success rate on all three major ObjectNav benchmarks: 62.0% on HM3Dv1 ($+0.6$ over BeliefMapNav), 81.5% on HM3Dv2 (the only method above 80%), and 41.8% on MP3D ($+4.5$ over BeliefMapNav), while requiring more than 96% fewer multimodal-model calls than representative LLM-driven baselines. Extensive ablations confirm that the semantic and information-theoretic value fields are complementary, and that the unified formulation is what enables the agent to remain decisive under strong cues and exploratory under weak ones.

1 Introduction

Object goal navigation (ObjectNav) asks an embodied agent to navigate to an instance of a target category (*e.g.*, "bed", "toilet") in a previously unseen environment [3]. Solving ObjectNav without task-specific training—the *zero-shot* setting—has emerged as a testbed for studying how large vision–language and language models (VLMs/LLMs) can be grounded in spatial decision-making [11, 18, 26, 33–35, 41]. Compared to policies trained on large ObjectNav datasets [7, 24, 32], zero-shot approaches offer substantially better generalization and are easier to audit and modify, which makes them an attractive setting for probing the interaction between foundation-model priors and classical exploration mechanisms.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

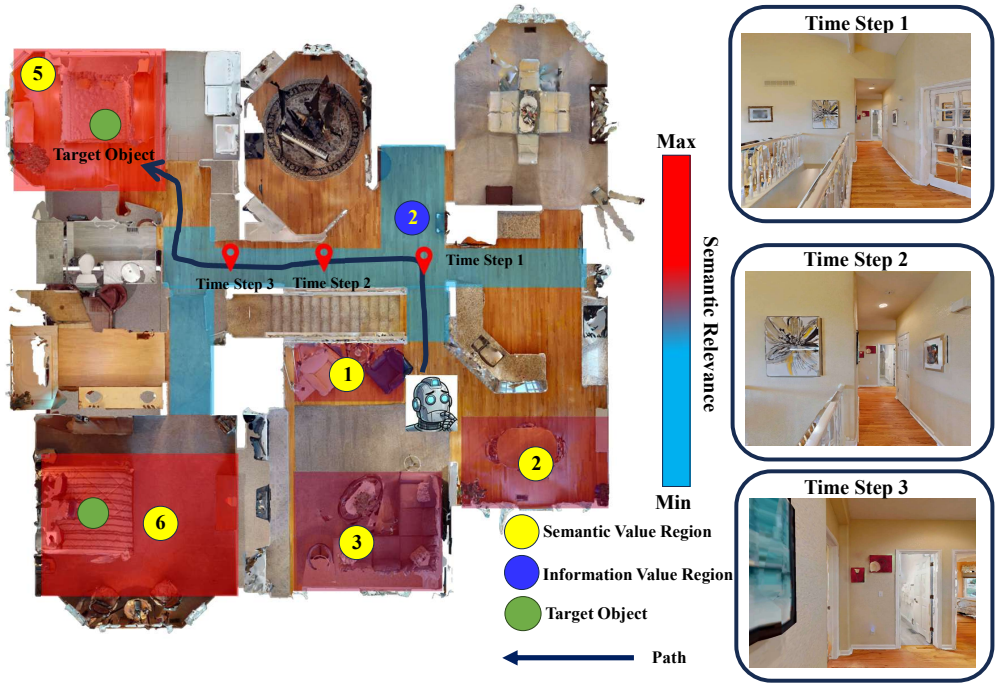


Figure 1: **Unified value-driven navigation in InfoNav.** The agent navigates by balancing semantic relevance (yellow regions, indicating high target likelihood) and information value (blue regions, indicating high exploration potential). When semantic cues are weak (Time Steps 1–2), the agent explores information-rich corridors; when it enters a semantically relevant region (Time Step 3), it shifts toward exploitation, ultimately reaching the target object (green).

In many modular zero-shot ObjectNav pipelines, the agent updates a spatial representation from incoming observations and uses semantic cues to select the next search location [33, 34]. Candidate locations often lie on frontiers, the boundaries between explored and unexplored space. The agent moves toward the selected location and repeats this process until it identifies a target instance, approaches it, and decides to stop.

Despite this progress, the design of zero-shot ObjectNav policies remains largely *ad hoc*. We identify three structural limitations that are shared by most existing methods:

(i) Semantic cues are treated uniformly across space. Methods that prompt a VLM or LLM for target–cue associations [33–35, 38] typically reduce the reasoning output to a single scalar per frontier or per grid cell. This ignores the fact that different contextual entities—rooms, co-occurring objects, functional parts—have very different spatial extents and therefore exert very different "effective ranges" of evidence on target localization. BeliefMapNav [42] and GAMap [36] incorporate multi-scale representations (3D voxel hierarchies and geometric-affordance scoring, respectively), yet neither explicitly models how semantic-level distinctions (room $\rightarrow$ object $\rightarrow$ part) should differentially modulate the spatial extent of target likelihood.

(ii) Semantic exploitation and exploration utility are decoupled. When semantic cues are weak, ambiguous, or simply absent in the currently observed region, semantics-only policies degrade gracefully only by accident. Recent works acknowledge this and propose mode switching: SemUtil [9] starts with geometric frontiers and switches to semantics; ApexNav [40] alternates between semantic and geometric modes. Discrete mode transitions, however, introduce hysteresis and oscillation around the switching boundary. RATE-Nav [14] models diminishing returns for local region exploration but lacks a persistent global value map, and its repeated per-region VLM queries incur high latency.

(iii) Target verification uses fixed thresholds. Detector confidences are either thresholded per frame [11, 34] or fused greedily into a semantic map [13, 35]. Both strategies are miscalibrated

Table 1: Comparison of navigation strategies in zero-shot ObjectNav methods. "Semantic" indicates VLM/LLM-based target-relevant scoring; "IG Exploration" indicates information-theoretic or geometric exploration; "Unified" indicates whether both are combined in a single value field rather than mode-switching or separate stages.

Method	Semantic	IG Exploration	Unified
VLFM [34]	✓	×	×
BeliefMapNav [42]	✓	×	×
GAMap [36]	✓	×	×
SemUtil [9]	✓	✓	×
ApexNav [40]	✓	✓	×
RATE-Nav [14]	✓	✓	×
InfoNav (Ours)	✓	✓	✓

against the temporal structure of ObjectNav: early in an episode, the opportunity cost of a false positive is high (a wrong stop terminates the episode), whereas late in an episode, missing a plausible target is the dominant failure mode. A single confidence threshold cannot simultaneously satisfy both regimes.

A unified value perspective. Rather than patching each of these limitations independently, we argue that all three stem from the absence of a *unified value representation*. A principled approach should (a) let the spatial extent of each semantic cue propagate its evidence proportionally, (b) represent exploration utility and semantic relevance in the same currency so that they can be combined (not switched) at every step, and (c) couple spatial decision-making with a *temporal* verification schedule whose confidence requirement decays as opportunity cost decays.

We formalize this through a Bayesian decomposition of target likelihood, an information-theoretic view of exploration, and an adaptive verification schedule inspired by the speed–accuracy tradeoff in sequential evidence accumulation. This design principle is also consistent with the broader dual-process view of decision-making under uncertainty [12]: slow, deliberative reasoning (done by LLM/VLM) produces hypotheses; fast value-driven action (done by frontier selection) exploits them.

Contributions. We introduce **InfoNav**, a unified value framework for zero-shot ObjectNav with the following contributions: **(1)** A *hierarchical semantic value map* that decomposes target likelihood via total probability over multi-level hypotheses (rooms, objects, parts), with priors proportional to spatial footprint, yielding a closed-form spatial influence field. **(2)** A *VLM-based information gain estimator* that captures long-horizon exploration utility by leveraging VLM recognition of connectivity structures (corridors, doorways). **(3)** A *memory-augmented verification* mechanism with a time-decaying confidence schedule that buffers uncertain detections for later re-examination. **(4)** State-of-the-art results: 62.0% SR on HM3Dv1 [23] ($+0.6/+2.1$ SR/SPL over BeliefMapNav), 81.5% on HM3Dv2 [29] (the only method exceeding 80%), and 41.8% on MP3D [6] ($+4.5$ over BeliefMapNav), while using over 96% fewer MLLM calls than InstructNav in the reported HM3Dv1 efficiency comparison.

2 Related Work

A complete navigation system integrates perception under partial observations, efficient motion planning, and safe obstacle avoidance [1, 25, 27]. Its ability to find and reach target objects in unfamiliar environments depends on how effectively these components work together.

Object Goal Navigation. ObjectNav methods are generally divided into end-to-end and modular pipelines. End-to-end reinforcement-learning policies [32] require heavy training and typically transfer poorly across scenes. Modular learning-based approaches [7, 24] reduce the training burden by decoupling mapping from policy, but still depend on task-specific ObjectNav data. The rise of LLMs [19] and VLMs [15, 22] has enabled a wave of *zero-shot* methods [11, 13, 26, 33–35, 38, 41] that operate without task-specific training. NLMap [8] builds an open-vocabulary scene representation from an exploration phase and queries object presence and locations to ground LLM-based

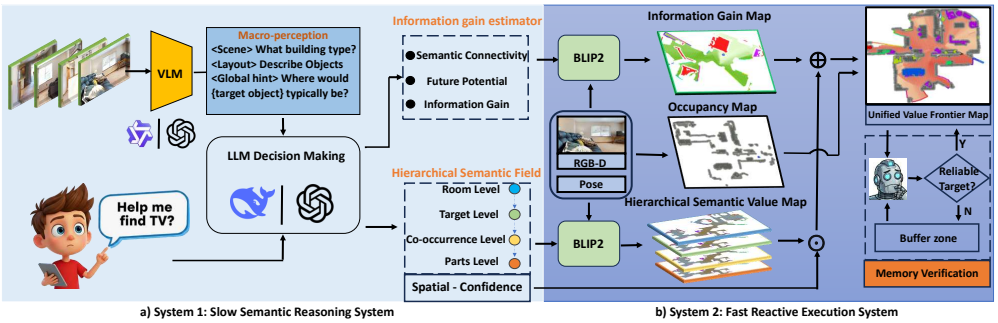


Figure 2: **Overview of InfoNav.** Given an RGB-D panorama, two parallel streams construct complementary value fields: a *Hierarchical Semantic Value Map* (HSVM) from VLM/LLM-generated multi-level hypotheses scored by a VLM, and an *Information Gain map* (IG) from connectivity-focused prompts. A unified value $V_{\text{total}}(f) = \alpha(t)V_{\text{HSVM}}(f) + (1-\alpha(t))\widehat{\text{IG}}(f)$ with adaptive weight $\alpha(t)$ selects the next waypoint via WTRP. Memory-augmented verification re-examines uncertain detections under a time-decaying threshold; STOP is issued only on CONFIRM.

task planning. InfoNav focuses on online exploration for zero-shot ObjectNav, jointly evaluating semantic relevance and exploration utility over frontier candidates.

Exploration and Semantic Guidance Strategies. Most zero-shot methods rely on frontier-based exploration [30] combined with semantic scoring: VLFM [34] builds vision–language value maps; BeliefMapNav [42] constructs 3D belief maps for target localization—but these lack explicit information-theoretic exploration and degrade when semantic cues are weak. Classical planners [4, 30] estimate information gain via geometric heuristics; SemUtil [9] uses a fixed two-stage schedule (geometry first, then semantics), while ApexNav [40] alternates between semantic and geometric modes—but discrete mode transitions introduce hysteresis and oscillation. RATE-Nav [14] models diminishing returns for local region exploration, yet lacks a persistent global value map and incurs high latency from repeated per-region VLM queries. As summarized in Table 1, InfoNav is the first to *continuously unify* both in a single value field, enabling smooth trade-offs without discrete transitions.

Target Verification. Prior methods threshold detector outputs [11, 34] or use max-confidence fusion [13, 35], both fragile to mispredictions [5]. TriHelper [38] and SG-Nav [33] add validation modules, but their decision rules remain time-invariant. Our memory-augmented verification with a time-decaying confidence schedule addresses this limitation.

3 Method

3.1 Problem Setup

In zero-shot ObjectNav, an agent is placed in an unseen environment and is asked to navigate to an object of category c (*e.g.*, "bed", "toilet"). At each timestep t the agent receives RGB-D observations (I_t^{rgb}, I_t^{depth}) together with odometry. The episode succeeds when the agent executes a STOP action within a threshold distance of an instance of c.

As illustrated in Fig. 2, InfoNav operates through a unified pipeline: (1) the agent captures an RGB-D panorama at each timestep; (2) two parallel streams construct complementary value fields—a **Hierarchical Semantic Value Map** V_{HSVM} encoding target likelihood and an **Information Gain Map** $\widehat{\text{IG}}$ encoding exploration potential; (3) these are fused into a unified value $V_{\text{total}}(f) = \alpha(t)V_{\text{HSVM}}(f) + (1-\alpha(t))\widehat{\text{IG}}(f)$ to select the next waypoint; (4) a **Memory-Augmented Verification** (MAV) module with time-decaying thresholds re-examines uncertain detections before issuing the final STOP action. Fig. 1 visualizes how this unified value balances semantic exploitation and information-seeking exploration across time steps.

3.2 Hierarchical Semantic Value Map

Spatially-weighted marginalization. Let A denote the target object and $\{B_1, \ldots, B_n\}$ the semantically related entities produced by our LLM prompt (room types, co-occurring objects, object parts). Consider an unexplored region $\mathcal{L}$ and let $P(A \mid \mathcal{L})$ denote the *expected probability that a location drawn uniformly at random from $\mathcal{L}$ is an instance of A*. By the law of total probability, marginalizing over the context entity type at that location yields

$$P(A \mid \mathcal{L}) = \sum_{i=1}^{n} P(A \mid B_i, \mathcal{L})\, P(B_i \mid \mathcal{L}). \tag{1}$$

The non-trivial modeling choice is the prior $P(B_i \mid \mathcal{L})$. Under the same uniform-sampling interpretation, $P(B_i \mid \mathcal{L})$ is the probability that a location drawn uniformly from $\mathcal{L}$ falls within the footprint of entity B_i, which *by definition* equals the area ratio

$$P(B_i \mid \mathcal{L}) = \frac{\mathcal{S}(B_i)}{\sum_{j=1}^{n} \mathcal{S}(B_j)}, \tag{2}$$

where $\mathcal{S}(B_i)$ is the area occupied by B_i. Under this interpretation, the area-ratio prior is *not an additional assumption about entity salience*; it is the unique spatial prior consistent with uniform sampling over $\mathcal{L}$. In contrast, the uniform prior $P(B_i \mid \mathcal{L}) = 1/n$ implicitly assigns each entity the same spatial weight regardless of scale—treating a small functional part (e.g., a faucet) and an entire room (e.g., a kitchen) as equally represented cues. This corresponds to the implicit default in VLFM-style scalar-per-cell scoring [34] and in LLM relational scoring [33]. Equations (1)–(2) together induce a *spatial influence field* whose effective range scales with each cue's geometric coverage: large-scale cues (a kitchen) bias a wide region, while small-scale cues (a faucet) bias only their neighborhood.

Hierarchical macro-perception and map construction. At the start of each episode, the agent captures an initial panorama $\mathcal{I}_0 = \{I_k\}_{k=1}^{4}$. A VLM produces a structured description $D_0 = f_{\text{VLM}}(\mathcal{I}_0)$. Conditioned on D_0 and the target category, an LLM generates hierarchical hypotheses $\mathcal{H} = \{h_l\}_{l=1}^{L}$ spanning room-level, object-level, co-occurrence, and part-level cues. These hypotheses are generated *once* at initialization and reused throughout the episode.

For each hypothesis h_l we build a 2D semantic score map following VLFM [34]. Each RGB image I_k, together with a context-augmented prompt, is fed to BLIP-2 to compute pixel-level cosine similarities. These scores are projected onto the free-space grid $\mathcal{G}$ and weighted by an angular confidence $c(p) = \cos^2\big(\frac{\theta(p)}{\theta_{\text{FoV}}/2} \cdot \frac{\pi}{2}\big)$, where $\theta(p)$ is the angular offset from the optical axis.

To integrate semantic evidence across frames, the fused score at grid cell p is updated by confidence-weighted averaging, $s_l^{\text{new}}(p) = \frac{c_l^{\text{cur}}(p) s_l^{\text{cur}}(p) + c_l^{\text{pre}}(p) s_l^{\text{pre}}(p)}{c_l^{\text{cur}}(p) + c_l^{\text{pre}}(p)}$, with the confidence updated analogously. The hypothesis-specific value map is $V_l(p) = C_l\, a_l\, s_l^{\text{new}}(p)$, where C_l corresponds to $P(A \mid B_l)$—the likelihood of target A given semantic cue B_l, predicted by the LLM—and a_l corresponds to the area ratio $\mathcal{S}(B_l)/\sum_j \mathcal{S}(B_j)$ from Eq. (2), estimated by the LLM based on typical physical extent of entity B_l. The final Hierarchical Semantic Value Map aggregates across hypotheses as $V_{\text{HSVM}}(p) = \frac{1}{|\mathcal{H}|} \sum_{l=1}^{|\mathcal{H}|} V_l(p)$. The area-ratio weighting ensures that larger-footprint cues (e.g., a dining table) contribute more strongly than smaller cues (e.g., a potted plant) with similar semantic relevance, because they occupy more of the sampling space $\mathcal{L}$ (Fig. 6, Appendix A.7).

3.3 Information Gain Estimation

Motivation: from local to long-horizon information gain. Traditional frontier exploration estimates information gain through geometric heuristics—typically counting unknown cells adjacent to each frontier. The fundamental limitation of such *local* metrics is their myopia: they cannot distinguish a frontier facing a narrow dead-end from one opening onto a large corridor system, because both may have similar numbers of immediately adjacent unknown cells. A more informative quantity would capture the *long-horizon* information gain—the cumulative map coverage gained by committing to a frontier and continuing exploration from there. Direct computation of such long-horizon quantities is intractable, but their *dominant factor* in indoor environments is well understood:

frontiers adjacent to connectivity structures (corridors, doorways, passages) unlock access to large unseen regions, whereas frontiers facing walls or saturated areas do not (Fig. 10, Appendix A.10). This architectural regularity is consistent with classical next-best-view heuristics [4].

VLM-based connectivity scoring. We exploit the fact that VLMs pretrained at scale can recognize connectivity structures directly from visual input. Rather than computing information gain explicitly, we query the VLM for the presence of *connectivity-indicative affordances*—visual cues suggesting that a frontier leads to unexplored, accessible space. This yields a practical information gain proxy: frontiers scored highly by the VLM are those adjacent to corridors, doorways, or open passages—precisely the structures that dominate long-horizon exploration utility in indoor scenes. In contrast to classical frontier methods that rely on local cell counts, this allows the agent to prioritize frontiers that open up spatially connected regions.

Concretely, we follow the same hierarchical perception–reasoning pipeline as HSVM. At episode initialization, the VLM extracts a structured description D_0 of the initial panorama, focusing on spatial layout and connectivity cues. The LLM then reasons over D_0 to generate a compact set of C connectivity-focused hypotheses $\mathcal{H}_{\text{IG}} = \{h_n\}_{n=1}^{C}$ expressed in natural language. **Unlike HSVM hypotheses which are target-dependent, IG hypotheses are target-agnostic**—constrained to describe *spatial connectivity structures* (openings, passages, or transitions) that may lead to unexplored regions regardless of the search target. Example outputs include "corridor leading to other areas", "doorway to another room", or "open archway connecting spaces". These hypotheses are generated *once* at the start of each episode and reused throughout, adapting to the initial observed environment while remaining grounded in the connectivity structures that dominate long-horizon information gain. The prompt design and example outputs are detailed in Appendix A.3.

At each exploration step the agent captures RGB observations $\{I_k\}$ from candidate frontiers. For each image I_k, we compute BLIP-2 ITM scores against all connectivity hypotheses and project them onto the top-down map within the camera FoV:

$$v_p^{\text{IG}} = \frac{1}{C} \sum_{h_n \in \mathcal{H}_{\text{IG}}} \text{BLIP2-ITM}(I_k, h_n), \quad \forall p \in \mathcal{M}_{\text{FOV}},$$

where p indexes map pixels and $\mathcal{M}_{\text{FOV}}$ is the set of top-down map pixels within the camera's FoV (excluding obstacle-occluded regions from the depth image). Revisited regions are updated via confidence-weighted averaging following VLFM [34]: $v_p^{\text{new}} = (c_p^{\text{curr}} v_p^{\text{curr}} + c_p^{\text{prev}} v_p^{\text{prev}})/(c_p^{\text{curr}} + c_p^{\text{prev}})$, with c_p again the angular confidence. The information gain estimate for frontier f is then the maximum score within its spatial neighborhood: $\widehat{\text{IG}}(f) = \max_{p \in \mathcal{N}(f)} v_p^{\text{IG}}$, where $\mathcal{N}(f)$ denotes the set of map pixels adjacent to frontier f.

3.4 Unified Decision Framework

The two value fields constructed above—V_{HSVM} encoding target likelihood and $\widehat{\text{IG}}$ encoding exploration utility—capture complementary aspects of navigation, but may induce conflicting behaviors: a semantically promising region may already be informationally saturated, while a highly informative region may be semantically irrelevant to the current target (Fig. 5, Appendix A.6). Rather than switching between semantic and exploration modes discretely, we fuse both value sources into a single decision at every timestep.

Adaptive dual-value fusion. We define the total frontier value as $V_{\text{total}}(f_i) = \alpha(t)\, V_{\text{HSVM}}(f_i) + (1 - \alpha(t))\, \widehat{\text{IG}}(f_i)$, where both components are min-max normalized to $[0, 1]$ and $\alpha(t) \in [0, 1]$ is an evidence-driven fusion weight:

$$\alpha(t) = \min\left\{1,\ \alpha_{\text{base}} + (1 - \alpha_{\text{base}})\, \frac{\max_p \widetilde{V}_{\text{HSVM}}(p)}{\tau_{\text{conf}}}\right\}, \tag{3}$$

where $\widetilde{V}_{\text{HSVM}}(p)$ denotes the *unnormalized* HSVM score. Intuitively: when no target-relevant cues have been observed, $\max_p \widetilde{V}_{\text{HSVM}}$ remains small and $\alpha(t) \approx \alpha_{\text{base}}$, favoring exploration; once a strong semantic peak emerges (e.g., detecting a kitchen while searching for a refrigerator), $\alpha(t) \to 1$ and the policy shifts toward exploitation.

WTRP-based frontier ordering. The fused values score individual frontiers but do not specify their visitation order. We select the top-k candidates $\{f_i\}_{i=1}^{k}$ by V_{total} and assign weights

$$w_i = \frac{\exp(V_{\text{total}}(f_i)/T_{\text{WTRP}})}{\sum_{j=1}^{k} \exp(V_{\text{total}}(f_j)/T_{\text{WTRP}})}, \tag{4}$$

where $T_{\text{WTRP}} > 0$ controls the concentration of the weights. Let p_i be the position of f_i, p_0 the current robot position, and $c(p_a, p_b)$ the travel cost measured by path length. We choose a permutation π of the candidates by solving

$$\pi^\star = \arg\min_{\pi} \sum_{i=1}^{k} w_{\pi(i)} \left[\sum_{j=1}^{i} c\big(p_{\pi(j-1)}, p_{\pi(j)}\big) \right], \qquad \pi(0) = 0. \tag{5}$$

This is a travel-distance formulation of the Weighted Traveling Repairman Problem (WTRP): the inner sum is the cumulative path cost before reaching the ith frontier in the sequence. A high-value frontier incurs a larger penalty when visited late, while travel costs can favor visiting a nearby, lower-value frontier first. The next waypoint is $f_{\pi^\star(1)}$. Appendix A.2 describes the solver; Table 5 compares this ordering with value-greedy selection.

Memory-augmented verification (MAV). Spatial decision-making alone does not address the temporal miscalibration of target verification: a detector's confidence at step t is a poor estimate of the target's true presence because the observation is made under a single, often partial viewpoint. Fixed-threshold decisions either prematurely commit to false positives (threshold too low) or discard genuine but weakly-supported detections irrecoverably (threshold too high). MAV addresses this by *remembering* uncertain detections and *re-examining* them once the agent's evidence base changes.

Concretely, MAV maintains a *suspicious memory buffer* $\mathcal{Q} = \{(p_i, s_i, t_i)\}$ storing each uncertain detection's position, confidence score s_i, and discovery time. A confidence schedule $\tau(t) = \tau_{\text{high}} - \frac{t}{T_{\max}}(\tau_{\text{high}} - \tau_{\text{low}})$ relaxes as exploration progresses. Detections with $s_i \geq \tau(t)$ trigger immediate VLM verification; those with $\tau_{\text{low}} \leq s_i < \tau(t)$ are buffered in $\mathcal{Q}$ and re-surfaced once $\tau(t)$ drops below s_i. The buffer is crucial: without it, a decaying threshold would only accept *new* weak detections, missing candidates seen earlier when the agent was more conservative (Fig. 3, Appendix A.4). The complete procedure is summarized in Algorithm 1 (Appendix A).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate InfoNav using Habitat-Sim and Habitat-Lab (both v0.3.1) [21] on three large-scale ObjectNav benchmarks: **HM3Dv1** (HM3D-Semantics v0.1; 2022 Habitat Challenge, 2000 episodes over 20 scenes, 6 goal categories), **HM3Dv2** (HM3D-Semantics v0.2; 2023 Habitat Challenge, 1000 episodes over 36 scenes, 6 goal categories), and **MP3D** [6] (2195 episodes over 11 scenes, 21 goal categories)—the most challenging of the three due to its larger category space and lower-quality meshes. Both HM3D benchmarks use scene reconstructions from HM3D [23] and object annotations from HM3D Semantics [29].

Evaluation metrics. Following standard practice [3], we report Success Rate (SR) and Success weighted by Path Length (SPL). SR measures episode success; SPL weighs success by path efficiency relative to the shortest path (and is zero for failed episodes). For ablations we additionally report SoftSPL (SSPL), which gives partial credit based on goal-directed progress in unsuccessful episodes. Higher is better for all three.

Implementation. InfoNav uses a VLM for environment perception and target verification, and an LLM for semantic reasoning. By default, we use Qwen3-VL [2] as the VLM and DeepSeek-V3 [10] as the LLM; ablations on model choice are provided in Table 6. For object detection we use D-FINE [20] for COCO-class objects and GroundingDINO [16] for open-vocabulary targets. We use Mobile-SAM [37] for semantic segmentation, and BLIP-2 [15] for building the semantic score map. Hyper- parameters are fixed across all experiments: $\alpha_{\text{base}} = 0.8$, $\tau_{\text{conf}} = 0.8$, $T_{\text{WTRP}} = 0.3$, $T_{\max} = 450$, $k = 20$. The two adaptive thresholds τ_{low} and τ_{high} are set dynamically by the LLM

Table 2: ObjectNav comparison on HM3Dv1, HM3Dv2, and MP3D. Gray rows: supervised methods. "–" indicates unreported results. Best zero-shot numbers are in **bold**.

Method	Venue	Unsup.	ZS	HM3Dv1		HM3Dv2		MP3D	
				SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
SemExp [7]	NeurIPS'20	×	×	37.9	18.8	–	–	36.0	14.4
PONI [24]	CVPR'22	×	×	–	–	–	–	31.8	12.1
ZSON [18]	NeurIPS'22	×	✓	25.5	12.6	–	–	15.3	4.8
CoW [11]	CVPR'23	✓	✓	32.0	18.1	–	–	7.4	3.7
ESC [41]	ICML'23	✓	✓	39.2	22.3	–	–	28.7	14.2
L3MVN [35]	IROS'23	✓	✓	50.4	23.1	36.3	15.7	34.9	14.5
OpenFMNav [13]	NAACL'24	✓	✓	54.9	24.4	–	–	37.2	15.7
InstructNav [17]	CoRL'24	✓	✓	–	–	58.0	20.9	–	–
VLFM [34]	ICRA'24	✓	✓	52.5	30.4	63.6	32.5	36.4	17.5
TriHelper [38]	IROS'24	✓	✓	56.5	25.3	–	–	–	–
SG-Nav [33]	NeurIPS'24	✓	✓	54.0	24.9	49.6	25.5	40.2	16.0
GAMap [36]	NeurIPS'24	✓	✓	53.1	26.0	–	–	–	–
ApexNav [40]	RA-L'25	✓	✓	59.6	**33.0**	76.2	38.0	39.2	17.0
MFNP [39]	ICRA'25	✓	✓	58.3	26.7	–	–	41.1	15.4
EfficientNav [31]	NeurIPS'25	✓	✓	–	–	80.0	**41.5**	–	–
BeliefMapNav [42]	NeurIPS'25	✓	✓	61.4	30.6	–	–	37.3	17.6
InfoNav (Ours)	–	✓	✓	**62.0**	32.7	**81.5**	36.4	**41.8**	**17.7**

according to the estimated physical size of the target category. All simulation experiments are run on a single NVIDIA RTX 4070 Ti with an Intel i9-13900K CPU; VLM and LLM APIs are called remotely.

4.2 Main Results

Comparison with the state of the art. Table 2 summarizes zero-shot ObjectNav results on three benchmarks. **InfoNav attains the highest success rate on all three datasets** (62.0 / 81.5 / 41.8 on HM3Dv1 / v2 / MP3D), while achieving competitive or best SPL.

On **HM3Dv1**, InfoNav reaches 62.0% SR and 32.7% SPL, exceeding BeliefMapNav (61.4/30.6) on both metrics. ApexNav achieves the highest SPL (33.0) but with lower SR (59.6). Compared to VLFM (52.5/30.4), InfoNav improves SR by +9.5, validating unified semantic-IG fusion. On **HM3Dv2**, InfoNav is the only method to exceed 80% SR (81.5%), outperforming ApexNav (76.2/38.0) by +5.3 SR and EfficientNav (80.0/41.5) by +1.5 SR. InfoNav surpasses VLFM by +17.9 SR, showing that IG-driven exploration discovers targets that semantic-only methods miss. On **MP3D**—the most challenging benchmark with 21 goal categories and lower-quality meshes—InfoNav achieves the best SR (41.8%) and best SPL (17.7%), demonstrating strong generalization across diverse object categories that stress-test LLM-generated semantic hypotheses.

SR–SPL trade-off analysis. InfoNav's design prioritizes maximizing SR, which introduces a principled trade-off with SPL on HM3Dv1 and HM3Dv2. The Memory-Augmented Verification (MAV) module re-examines uncertain detections rather than discarding them, recovering targets that fixed-threshold methods miss—but at the cost of additional verification steps. Similarly, the IG-driven exploration steers the agent toward information-rich regions even when semantically promising areas have been partially explored, increasing coverage but potentially lengthening paths. We view this as a favorable trade-off: an agent that fails to find the target contributes zero to SPL, whereas one that finds the target via a longer path still succeeds. A sensitivity analysis over α_{base} characterizing how SR and SPL co-vary is provided in Appendix A.9.

Computational efficiency. Table 3 compares average per-episode cost on HM3Dv1. Despite achieving the highest SR, InfoNav uses only 3.8 MLLM calls per episode (one for initial macro-perception and 2–3 for target verification) versus roughly 110–115 for InstructNav—a reduction of over 96%—while cutting the average navigation steps from about 320–330 to 155.3, and runtime from about 450–460 s to 100.12 s. This efficiency stems from offloading spatial decision-making

Table 3: Per-episode computational cost on HM3Dv1 (approximate ranges for baselines).

Method	MLLM↓	Steps↓	Time(s)↓
L3MVN	[30–40]	[190–200]	[1000–1100]
PixNav	[55–60]	[270–290]	[270–280]
SG-Nav	[120–125]	[295–305]	[720–735]
InstructNav	[110–115]	[320–330]	[450–460]
Ours	**3.8**	**155.3**	**100.12**

to the persistent value map: unlike SG-Nav and InstructNav which invoke an LLM per frontier, InfoNav builds a reusable semantic-IG value field at episode start and queries the MLLM only for verification.

4.3 Ablation Studies

Table 4: Ablation studies on HM3Dv2. **(a)** Component ablation: the three modules (HSVM = hierarchical semantic value map; IG = information gain estimator; MAV = memory-augmented verification) are complementary. **(b)** HSVM hypothesis hierarchy with IG disabled: SR and SSPL improve monotonically as finer levels (room, co-occurrence, parts) are added; SPL shows a net $+1.0$ gain from L1 to L4 with a small dip at L3 that is recovered at L4.

(a) Component ablation.

HSVM	IG	MAV	SR↑	SPL↑	SSPL↑
×	✓	✓	80.5	35.2	36.4
✓	×	✓	80.1	36.1	37.4
✓	✓	×	77.8	36.2	**38.0**
×	×	×	73.2	32.1	34.9
✓	✓	✓	**81.5**	**36.4**	37.5

(b) HSVM hypothesis hierarchy.

Hypotheses	SR↑	SPL↑	SSPL↑
L1: target only	79.3	35.1	36.3
L2: L1 + room	79.8	35.5	36.6
L3: L2 + co-occurrence	80.0	35.1	36.6
L4: L3 + object parts	**80.1**	**36.1**	**37.4**

Table 4a isolates the contribution of each component on HM3Dv2. The full model reaches 81.5/36.4/37.5 (SR/SPL/SSPL); removing all three reduces InfoNav to a greedy nearest-frontier policy and yields 73.2/32.1/34.9, a -8.3 SR drop that establishes the aggregate contribution of our design.

The component-removal ablations show different effects on SR and SPL. **MAV** contributes most to SR (-3.7 when removed): by re-examining uncertain detections as the confidence threshold decays, MAV recovers targets that fixed-threshold methods irrecoverably discard—especially valuable for partially occluded or ambiguously viewed objects. **HSVM** contributes most to SPL (-1.2 when removed): hierarchical hypotheses provide multi-scale semantic guidance toward target-likely regions, shortening paths compared to uniform exploration. **IG** provides complementary gains (-1.4 SR, -0.3 SPL when removed): by scoring frontiers on connectivity affordances, IG steers exploration toward passages that unlock access to new regions.

HSVM hierarchy (Table 4b). To isolate HSVM's contribution, we disable IG and vary only the hypothesis levels. Starting from target-only matching (L1) at 79.3% SR, progressively adding room-level (L2), co-occurrence (L3), and part-level (L4) hypotheses yields monotonic SR gains (79.3→79.8→80.0→80.1). SPL shows a net +1.0 gain from L1 to L4 (35.1→36.1) with a small dip at L3, which we attribute to co-occurrence cues occasionally steering toward semantically plausible but geometrically distant regions; the part-level cues at L4 recover this by providing finer localization. Overall, the four-level hierarchy outperforms target-only by +0.8% SR and +1.0% SPL, confirming that multi-scale cues are complementary.

Frontier ordering (Table 5). We compare WTRP with value-greedy selection on the same 1,000 HM3Dv2 episodes, keeping the candidate frontiers and all other components unchanged. The value-greedy baseline selects the frontier with the highest fused value at each decision step. WTRP raises SR from 80.1% to 81.5% and SPL from 36.0 to 36.4, corresponding to gains of 1.4 and 0.4 percentage points, respectively.

Table 5: Frontier ordering on the same 1,000 HM3Dv2 episodes, with all other components fixed.

Frontier ordering	SR↑	SPL↑
Value-greedy	80.1	36.0
WTRP (Ours)	**81.5**	**36.4**

Effect of VLM/LLM backbone. Table 6 compares VLM/LLM combinations on HM3Dv2. The modular combination (Qwen3-VL + DeepSeek-V3) achieves the highest SR (81.5%), while the unified GPT-5.2 configuration attains higher SPL (37.3%). We hypothesize that GPT-5.2's tighter vision–language integration enables more decisive hypothesis generation, leading to shorter paths when correct but occasional overconfident errors that lower SR.

Table 6: Effect of VLM/LLM backbone on HM3Dv2.

VLM	LLM	SR↑	SPL↑
Qwen3-VL	DeepSeek-V3	**81.5**	36.4
GPT-5.2	GPT-5.2	81.0	**37.3**

Failure analysis. Figure 7 summarizes success and failure categories on HM3Dv1 and HM3Dv2. Appendix A.8, Fig. 8, shows representative false-positive, step-budget-exhaustion, and cross-floor cases. These illustrate errors in target recognition, unsuccessful search within a finite budget, and the limits of the 2D map for inter-floor navigation.

4.4 Real-World Deployment

We deploy InfoNav on a Tarkbot mobile robot equipped with a RealSense D455i RGB-D camera and a Livox Mid-360 LiDAR. The onboard NUC13 handles perception and low-level control, while InfoNav runs on a remote desktop with an RTX 4070 Ti, exchanging sensor observations and action commands over WiFi. In this qualitative indoor deployment, the robot exhibits both semantic-guided target search and information-seeking exploration, demonstrating the feasibility of the navigation pipeline on a physical robot. Further deployment details are provided in Appendix A.5 (Fig. 4).

5 Conclusion

We presented **InfoNav**, a unified value-driven framework for zero-shot ObjectNav that jointly estimates semantic relevance and long-horizon information gain in a single map representation. Its three components—hierarchical semantic value map, VLM-based IG surrogate, and memory-augmented verification—achieve state-of-the-art success rates among the compared methods on HM3Dv1/HM3Dv2/MP3D, with $> 96\%$ fewer MLLM calls than InstructNav in the reported HM3Dv1 comparison. The key insight is that representing exploration utility and semantic relevance in a shared value currency eliminates the need for discrete mode-switching, enabling smooth and adaptive navigation behavior.

Limitations. The 2D value map does not support cross-floor reasoning. The IG estimator uses visual connectivity cues as a proxy for exploration utility rather than computing information gain explicitly. Open-vocabulary detection remains a bottleneck for rare object categories. Our single-run evaluations do not quantify run-to-run variability, and the hardware deployment is qualitative. Future work will evaluate InfoNav on real-world navigation benchmarks, with a focus on success rate, path efficiency, and robustness to depth noise and localization errors. Broader-impact discussion is provided in Appendix A.11.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grants 62473046, T2541018, 72642006, and 62688201. The authors declare no competing interests.

References

[1] Timur Akhtyamov, German Devchich, Egor Kuznetsov, Denis Fatykhov, Chengpu Yu, and Gonzalo Ferrer. Approximate C-uniform sampling: An information-theoretic and Bayesian inference perspective. *Robotics*, 15(7):139, 2026. doi: 10.3390/robotics15070139. URL https://www.mdpi.com/2218-6581/15/7/139.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

[3] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. doi: 10.48550/arXiv.2006.13171. URL https://arxiv.org/abs/2006.13171.

[4] Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon "next-best-view" planner for 3D exploration. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[5] Finn Lukas Busch, Timon Homberger, Jesús Ortega-Peimbert, Quantao Yang, and Olov Andersson. One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14835–14842, 2025. doi: 10.1109/ICRA55743.2025.11128393.

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[7] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, volume 33, pages 4247–4258, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

[8] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11509–11522. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10161534. URL https://arxiv.org/abs/2209.09874.

[9] Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. In *Robotics: Science and Systems (RSS) 2023*, 2023.

[10] DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. doi: 10.48550/arXiv.2412.19437.

[11] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23171–23181, 2023.

[12] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[13] Yuxuan Kuang, Hai Lin, and Meng Jiang. OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 338–351. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.24. URL https://aclanthology.org/2024.findings-naacl.24/.

[14] Junjie Li, Nan Zhang, Xiaoyang Qu, Kai Lu, Guokuan Li, Jiguang Wan, and Jianzong Wang. RATE-Nav: Region-aware termination enhancement for zero-shot object navigation with vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6564–6574, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.341. URL https://aclanthology.org/2025.findings-acl.341/.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[16] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15105 of *Lecture Notes in Computer Science*, pages 38–55, Milan, Italy, 2024. Springer. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

[17] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot system for generic instruction navigation in unexplored environment. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2049–2060. PMLR, 2025. URL https://proceedings.mlr.press/v270/long25b.html.

[18] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[19] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

[20] Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task of DETRs as fine-grained distribution refinement. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6cf58a87e3097e7d1f9be3e8693a93de-Abstract-Conference.html.

[21] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/430894999584d0bd358611e2ecf00b15-Abstract-Conference.html.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[23] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=-v4OuqNs5P.

[24] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for ObjectGoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 18890–18900, 2022.

[25] Ping Wang, Chengpu Yu, Maolong Lv, and Guang-Ren Duan. Safe fixed-time reinforcement learning for nonlinear zero-sum games with obstacle avoidance awareness. *Automatica*, 183:112673, 2026. doi: 10.1016/j.automatica.2025.112673. URL https://doi.org/10.1016/j.automatica.2025.112673.

[26] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. VoroNav: Voronoi-based zero-shot object navigation with large language model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 53757–53775. PMLR, 2024.

[27] Jingchen Xu, Chengpu Yu, Gonzalo Ferrer, and Luwei Liu. A neural predict-update framework for multi-object tracking using partial observations. In *2026 38th Chinese Control and Decision Conference (CCDC)*, pages 1490–1495, 2026. doi: 10.1109/CCDC69976.2026.11560126. URL https://cmsweb.com.sg/rps2prod/ccdc2026/epro/html/SunA10-4.html.

[28] Wei Xu and Fu Zhang. FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated kalman filter. *IEEE Robotics and Automation Letters*, 6(2):3317–3324, 2021. doi: 10.1109/LRA.2021.3064227. URL https://doi.org/10.1109/LRA.2021.3064227.

[29] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3D semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4927–4936, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yadav_Habitat-Matterport_3D_Semantics_Dataset_CVPR_2023_paper.html.

[30] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 146–151. IEEE, 1997.

[31] Zebin Yang, Sunjian Zheng, Tong Xie, Tianshi Xu, Bo Yu, Fan Wang, Jie Tang, Shaoshan Liu, and Meng Li. EfficientNav: Towards on-device object-goal navigation with navigation map caching and retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.18546.

[32] Xin Ye and Yezhou Yang. Efficient robotic object search via HIEM: Hierarchical policy learning with intrinsic-extrinsic modeling. *IEEE Robotics and Automation Letters(RAL)*, 6(3):4425–4432, 2021.

[33] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. SG-Nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[34] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48, 2024.

[35] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023.

[36] Shuaihang Yuan, Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, and Yi Fang. GAMap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.23978.

[37] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. doi: 10.48550/arXiv.2306.14289.

[38] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. TriHelper: Zero-shot object navigation with dynamic assistance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10035–10042. IEEE, 2024.

[39] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[40] Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. ApexNAV: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *IEEE Robotics and Automation Letters*, 10(11):11530–11537, 2025. doi: 10.1109/LRA.2025.3606388.

[41] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning (ICML)*, 2023.

[42] Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. BeliefMapNav: 3D voxel-based belief map for zero-shot object navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.06487.

A Additional Implementation Details

This appendix collects implementation details and supporting material that did not fit into the main text. It is intended as "optional reading" for reviewers; the main paper stands alone without it.

A.1 Inference-Time Algorithm

Algorithm 1 summarizes the complete inference-time procedure of InfoNav, tying together the three components of the main paper (HSVM construction, IG estimation, and memory-augmented verification).

Algorithm 1 InfoNav: unified value-driven frontier selection.

target category c; step budget $T_{\max}$; thresholds $\tau_{\text{low}}, \tau_{\text{high}}, \tau_{\text{conf}}, \alpha_{\text{base}}, k$
Capture initial panorama $\mathcal{I}_0$; $D_0 \leftarrow \texttt{VLM}(\mathcal{I}_0)$
$\mathcal{H}_{\text{sem}} \leftarrow \texttt{LLM}(c, D_0)$ {generate semantic hypotheses once}
$\mathcal{H}_{\text{IG}} \leftarrow \texttt{LLM}(D_0)$ {generate IG hypotheses once}
Initialize value maps $V_{\text{HSVM}} \leftarrow 0$, $\widehat{\text{IG}} \leftarrow 0$; memory buffer $\mathcal{Q} \leftarrow \emptyset$
for $t = 1, \ldots, T_{\max}$ **do**
Capture panorama $\mathcal{I}_t$; update map belief $\mathcal{M}_t$
Update V_{HSVM} with $\mathcal{H}_{\text{sem}}$; update $\widehat{\text{IG}}$ with $\mathcal{H}_{\text{IG}}$ over free-space grid
$\alpha(t) \leftarrow \min\{1,\ \alpha_{\text{base}} + (1-\alpha_{\text{base}}) \frac{\max_p \widetilde{V}_{\text{HSVM}}(p)}{\tau_{\text{conf}}}\}$
$V_{\text{total}}(f) \leftarrow \alpha(t) V_{\text{HSVM}}(f) + (1-\alpha(t)) \widehat{\text{IG}}(f)$ for all $f \in \mathcal{F}_t$
Solve WTRP over top-k frontiers $\Rightarrow$ next waypoint f^*
if detector $c_i \geq \tau(t) := \tau_{\text{high}} - \frac{t}{T_{\max}}(\tau_{\text{high}} - \tau_{\text{low}})$ **then**
Verify with VLM; **if verified:** `STOP` and return
else if $\tau_{\text{low}} \leq c_i < \tau(t)$ **then**
Store (p_i, c_i, t) in memory buffer $\mathcal{Q}$ {MAV retains uncertain detections}
end if
Re-surface any $(p_j, c_j, t_j) \in \mathcal{Q}$ with $c_j \geq \tau(t)$ for verification
Execute action toward f^*
end for

A.2 WTRP-Based Frontier Ordering

The frontier weights and ordering objective are defined in Eqs. (4)–(5). Each segment of the route contributes to the cumulative cost of every frontier visited after that segment. Thus, the objective accounts for both the frontier values and the cost of delaying their visits; it does not directly minimize the total length of the route.

For $k \leq 20$, we use dynamic programming to solve the ordering problem exactly; for larger candidate sets, we use a greedy nearest-neighbor heuristic. The experiments use $k = 20$ and $T_{\text{WTRP}} = 0.3$. The value-greedy comparison in Table 5 uses the same candidate set and selects the frontier with the highest fused value.

A.3 Prompts and Hypothesis Generation

Hierarchical hypothesis generation for HSVM. At episode initialization, the agent captures an initial panorama and obtains a VLM description D_0. Conditioned on D_0 and the target category c, the LLM is prompted to return a JSON list of hypotheses organized into four levels: (i) room type (largest spatial footprint), (ii) target object itself, (iii) co-occurring objects, and (iv) object parts (smallest spatial footprint). For each hypothesis the LLM returns: (a) a confidence score $C_l \in [0, 1]$, corresponding to $P(A \mid B_l)$—the co-occurrence likelihood of finding target A given cue B_l; and (b) a spatial footprint $a_l \in [0, 1]$, corresponding to the area ratio in Eq. (2)—the relative spatial extent of entity B_l. These hypotheses are generated *once* at the start and reused throughout the episode. The prompt template is shown below:

HSVM Hypothesis Generation Prompt

```
You are an expert indoor navigation planner helping a robot find
{target_object}.
Current Scene Observation (from VLM): {vlm_perception}
Generate exactly 4 semantic hypotheses at different hierarchical levels:
Level 1 (Room Type): spatial_footprint 0.8-1.0
Level 2 (Target Object): confidence = 1.0, spatial_footprint 0.6-0.8
Level 3 (Co-occurrence): spatial_footprint 0.4-0.7
Level 4 (Object Part): spatial_footprint 0.2-0.4
Output: JSON with environment_analysis, target_location_reasoning,
semantic_hypotheses (id, hypothesis_type, description, confidence,
spatial_footprint), and exploration_strategy.
```

The system includes predefined knowledge for common targets (*e.g.*, for "bed": room_examples = bedroom/guest room; cooccur_examples = nightstand/dresser; part_examples = mattress/pillows/headboard). See Appendix A.7 for how C_l and a_l implement the spatial influence field.

Information-gain hypothesis generation. As described in Section 3.3, information-gain hypotheses target the connectivity structures (corridors, doorways, passages) that dominate long-horizon information gain in indoor scenes. Unlike HSVM hypotheses which are target-dependent, IG hypotheses are *target-agnostic* and focus solely on spatial connectivity. The LLM prompt requires *complete sentences* (8–15 words) that follow specific patterns for effective BLIP-2 Image-Text Matching:

IG Hypothesis Generation Prompt

```
Based on the following scene description, generate 3-5 COMPLETE
SENTENCES describing spatial connectivity structures that may lead
to unexplored areas.
Scene description: {vlm_perception}
Required sentence patterns:
- "This view shows [structure] leading to [destination]"
- "There is [structure] that leads to [destination]"
- "This [structure] connects to [destination]"
Output: JSON with ig_hypotheses (id, prompt, structure_type,
reasoning) and connectivity_summary.
```

Example LLM outputs:

- "This view shows a corridor leading to private rooms"
- "There is a doorway that leads to the kitchen area"
- "This hallway connects to multiple bedrooms"

The complete-sentence format (vs. short phrases) improves BLIP-2 ITM matching accuracy. In practice, the LLM generates 3–5 hypotheses per episode, keeping inference cost manageable.

A.4 Memory-Augmented Verification: Schedule and Buffer Details

Schedule endpoints. The two endpoints τ_{low} and τ_{high} of the confidence schedule $\tau(t)$ are set by the LLM based on the target's estimated physical scale: smaller targets (*e.g.*, small appliances) receive a lower τ_{high} because the detector's confidence on small objects is typically lower; larger targets (*e.g.*, beds, sofas) receive a higher τ_{high}. The schedule then linearly anneals from τ_{high} to τ_{low} over $T_{\text{max}} = 450$ steps.

Memory buffer operations. The suspicious memory buffer $\mathcal{Q}$ stores each uncertain detection as a triple (p_i, c_i, t_i): position in the global map, detector confidence at the time of observation, and the step index. At every step, MAV performs two operations: (1) *enqueue*: new detections falling

in $[\tau_{\text{low}}, \tau(t))$ are stored rather than discarded; (2) *re-surface*: any previously stored $(p_j, c_j, t_j) \in \mathcal{Q}$ whose confidence c_j is now above the (decayed) current threshold $\tau(t)$ is re-dispatched for multi-view verification by the VLM. This two-way mechanism is what distinguishes MAV from a simple time-decaying threshold: without the buffer, only *new* observations could trigger verification, and candidates seen earlier (when the agent was more conservative) would be irrecoverable. In practice the buffer rarely exceeds 15 entries per episode (small-object categories peak at $\sim$25), making the additional cost negligible.

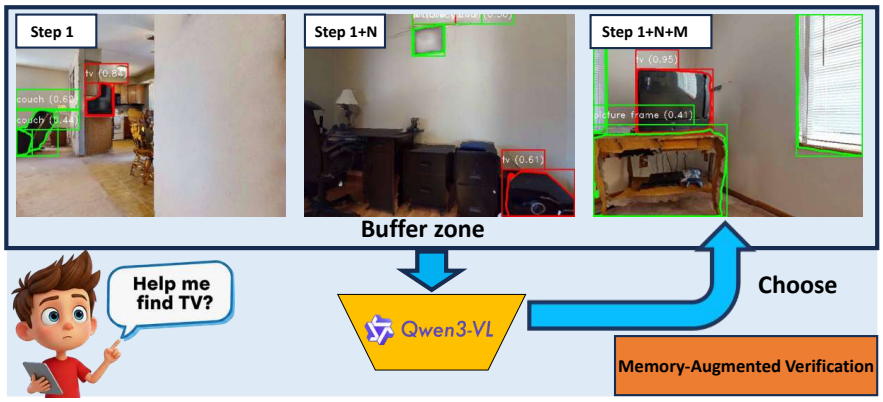


Figure 3: **Memory-augmented verification (MAV).** Low-confidence detections are not discarded but held in a buffer zone, and re-examined by the VLM once the confidence schedule $\tau(t)$ relaxes to their level. At Step 1, an uncertain TV detection is buffered rather than acted upon; at Step 1+N, exploration continues while buffered candidates accumulate; at Step 1+N+M, MAV re-surfaces a high-confidence detection (TV, 0.95) for verification, enabling reliable target confirmation.

A.5 Real-World Deployment

To verify that InfoNav transfers beyond simulation, we deploy it on a Tarkbot mobile platform (Fig. 4). The robot carries an NUC13 onboard computer (used only for perception and low-level control), an Intel RealSense D455i RGB-D camera, and a Livox Mid-360 LiDAR. FAST-LIO [28] produces high-frequency odometry. The InfoNav algorithm runs remotely on a desktop with an RTX 4070 Ti; RGB-D frames and odometry are streamed over WiFi, and actions are sent back to the robot. This qualitative deployment shows semantic-guided search and information-seeking exploration on the physical platform. It does not provide a quantitative estimate of real-world success rate or robustness to depth noise and localization errors.



Figure 4: Real-world deployment. The Tarkbot platform used for real-world evaluation carries an NUC13 onboard computer, a RealSense D455i RGB-D camera, and a Livox Mid-360 LiDAR.

A.6 Design Illustrations

Fig. 5 illustrates the conflict between semantic exploitation and information-theoretic exploration that motivates our unified decision framework in Sec. 3.4: a semantically promising region may be

informationally saturated, while a highly informative region may be semantically irrelevant. Rather than switching between these two regimes discretely, InfoNav combines them at every step through the adaptive fusion weight $\alpha(t)$ of Eq. (3).

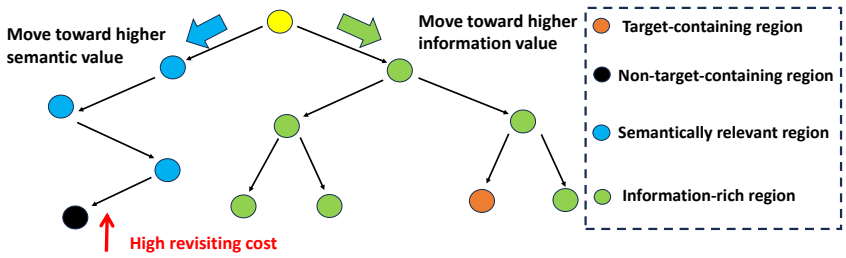


Figure 5: **Conflicting navigation behaviors from semantic relevance vs. information value.** Moving toward higher semantic value (blue arrow, left branch) may lead to a non-target-containing region (black) with high revisiting cost if the semantic cue is misleading. Moving toward higher information value (green arrow, right branch) explores information-rich regions (green) that may eventually reveal the target-containing region (orange). A unified decision framework balances both objectives to avoid premature commitment and ensure robust target discovery.

A.7 Spatial Influence Field in HSVM

This section provides additional intuition for the area-ratio prior (Eq. 2) and its effect on the Hierarchical Semantic Value Map.

Why area weighting matters. Consider an agent searching for a *bed* in an indoor scene. The LLM generates several semantic hypotheses: a *dining table* and a *potted plant*, both semantically related to the target (they co-occur in residential spaces). Suppose both have identical semantic relevance $P(A \mid B_i) = C_l$. Under the uniform prior $P(B_i \mid \mathcal{L}) = 1/n$, both cues contribute equally to the HSVM. However, this ignores a crucial spatial fact: the dining table occupies a much larger floor area than the plant. If we sample a random location uniformly from the unexplored region $\mathcal{L}$, we are far more likely to land within the table's footprint than the plant's. Formally, the area-ratio prior $P(B_i \mid \mathcal{L}) = \mathcal{S}(B_i)/\sum_j \mathcal{S}(B_j)$ captures exactly this: larger entities receive higher prior weight because they occupy more of the sampling space.

Visualization. Fig. 6 illustrates this effect. The dashed circles represent the *spatial influence field* of each semantic cue—the region over which the cue provides evidence about target location. The **table** (black dashed circle) has a larger influence field than the **plant** (green circle), reflecting its greater physical extent. Crucially, the table's influence extends into the **unexplored area** (red dashed region), providing guidance even before that region is directly observed. This is the key benefit of area-weighted priors: they allow large-scale cues to "reach" into unobserved space, biasing exploration toward regions likely to contain the target.

Connection to Eq. (1)–(2). In the main text, C_l corresponds to $P(A \mid B_l)$—the conditional probability of finding target A given that semantic cue B_l is present. This is predicted by the LLM based on commonsense co-occurrence knowledge (*e.g.*, beds are likely near nightstands). The term a_l corresponds to the normalized area ratio from Eq. 2, estimated by the LLM based on typical physical dimensions of entity B_l. Together, the product $C_l \cdot a_l$ weights each hypothesis by both its semantic relevance and its spatial coverage, implementing the marginalization in Eq. 1.

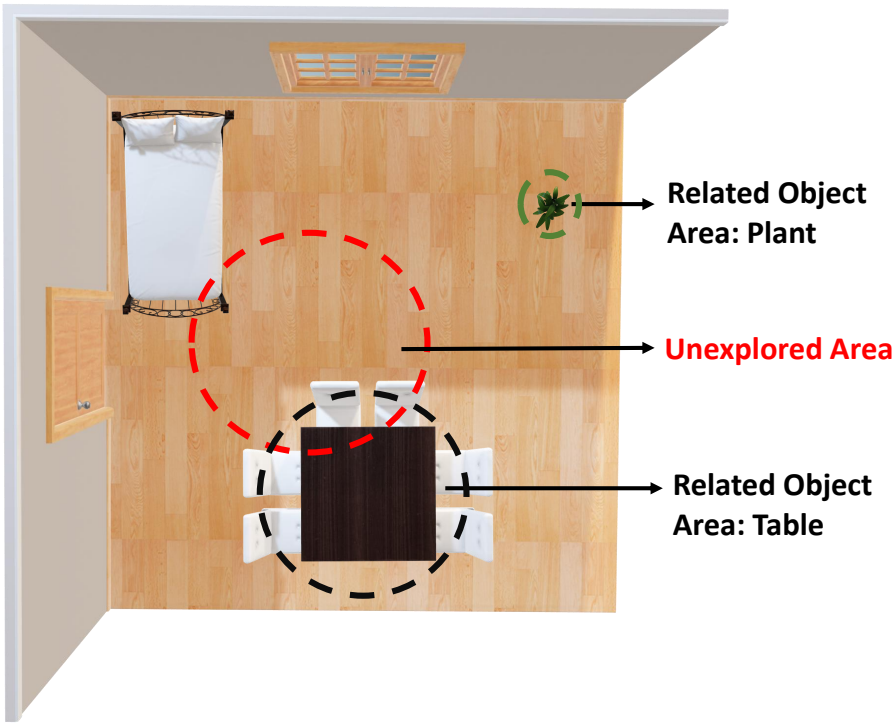


Figure 6: **Spatial influence field for semantic cues in HSVM.** The dashed circles indicate each cue's influence region. Even if the *table* and *plant* have identical semantic relevance $P(A \mid B_i)$ to the target *bed*, the table's larger footprint yields a broader influence field (black dashed circle) that extends into the unexplored area (red), while the plant's small footprint localizes its influence (green circle). The area-ratio prior ensures that larger cues contribute proportionally more to target localization.

A.8 Failure Analysis and Representative Cases

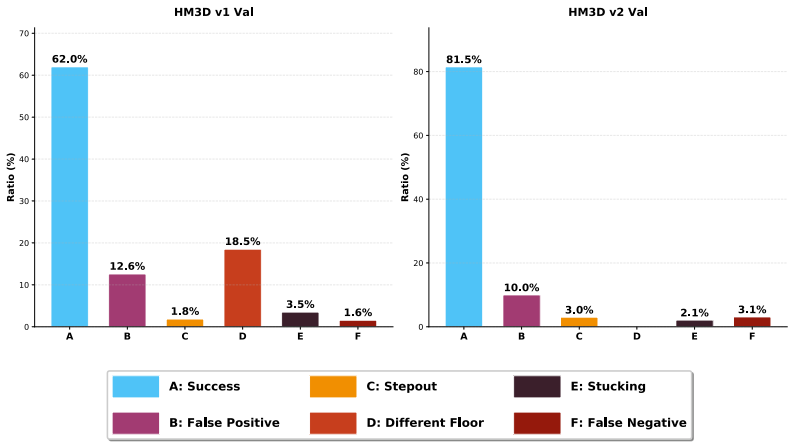
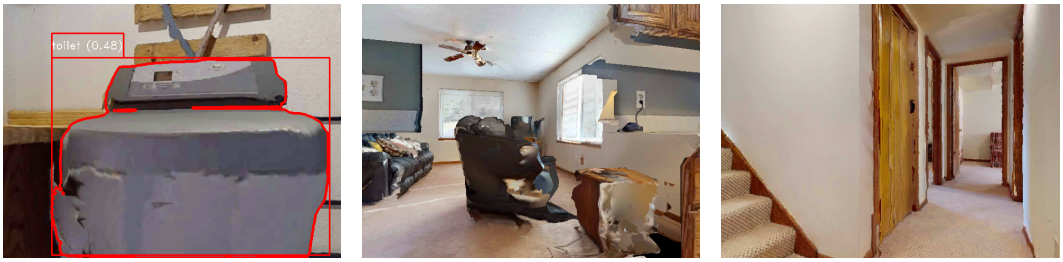


Figure 7: Failure-cause distribution of InfoNav on HM3Dv1 (left) and HM3Dv2 (right). **A**: Success. **B**: False Positive. **C**: Stepout. **D**: Different Floor. **E**: Stucking. **F**: False Negative.

Figure 7 includes both successful and failed episodes; the category percentages are relative to all evaluated episodes, not only failures. Figure 8 presents observations from three representative failure cases. The panels illustrate the assigned failure categories; a single observation does not reconstruct the full sequence of decisions leading to termination.



(a) False positive (b) Stepout (c) Different floor

Figure 8: Representative failure observations. (a) A suitcase-like object is labeled as `toilet` with a displayed detector score of 0.48. (b) An indoor observation from a search that exhausted its step budget. (c) A stairway and adjoining corridor in a cross-floor failure case. The category labels refer to the corresponding cases; the images are individual observations rather than complete trajectories.

False positive. In Fig. 8(a), the predicted `toilet` region covers a suitcase-like object. This illustrates an incorrect object-category assignment in the observed scene. The displayed score of 0.48 is a detector output; it does not establish whether the candidate passed subsequent verification. For this class of error, verification from additional viewpoints is a potential safeguard, but its effectiveness requires evaluation on the corresponding episodes.

Step-budget exhaustion (Stepout). Figure 8(b) shows a furnished room in a case that reached the step limit without successful completion. Stepout records the termination condition rather than a unique failure mechanism: insufficient coverage, repeated visits, missed detections, or local navigation difficulty can all consume the available budget. Distinguishing these explanations requires the trajectory and detection history.

Different floor. The stairway in Fig. 8(c) illustrates the vertical connectivity relevant to cross-floor search. InfoNav selects frontiers using a 2D value map and has no explicit representation of inter-floor transitions. Extending search to such cases would require representing floor connectivity and selecting transitions between levels, together with a controller capable of executing them.

A.9 Additional Ablation Analyses

Sensitivity to α_{base} (exploration–exploitation balance). We vary the base weight α_{base} in Eq. (3) over $\{0.0, 0.3, 0.5, 0.8, 1.0\}$ on HM3Dv2 (Fig. 9). The reported SR values are 80.5%, 81.3%, 81.4%, 81.5%, and 80.1%, respectively, with the highest value at the default setting $\alpha_{\text{base}} = 0.8$. Under Eq. (3), setting $\alpha_{\text{base}} = 0$ does not necessarily remove the semantic term: the actual fusion weight can remain positive when semantic evidence is present. At $\alpha_{\text{base}} = 1$, the fusion weight is fixed at one. A sweep of the base weight therefore does not replace a direct comparison between dynamic fusion and a constant $\alpha(t)$.

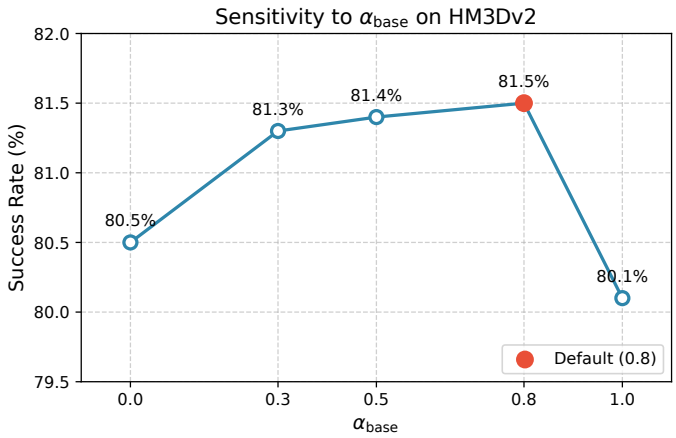


Figure 9: Sensitivity of InfoNav to the fusion parameter α_{base} on HM3Dv2. All other hyperparameters are kept at their main-paper values ($\tau_{\text{conf}} = 0.8$, $k = 20$, $T_{\max} = 450$). The red point indicates the default setting.

A.10 Visualization of the VLM-based Information Gain Estimator

Our VLM-based information gain estimator (Sec. 3.3) scores frontiers by their *connectivity affordances*—visual cues indicating access to unexplored regions. Fig. 10 visualizes the estimator qualitatively: the VLM assigns higher scores to frontiers adjacent to connectivity structures (red region) than to frontiers facing saturated or dead-end areas (black region).

Key insight. Connectivity structures (*e.g.*, corridors, doorways) often have *low local visibility*—seeing a corridor entrance reveals only a few new cells—but *high long-horizon value* because they lead to large unexplored regions beyond. Our VLM estimator captures this forward-looking property by recognizing connectivity affordances from visual cues and assigning high scores to frontiers that *lead to* information-rich areas, not just frontiers immediately adjacent to many unknown cells. This design is consistent with classical next-best-view heuristics [4] that exploit geometric openness, but leverages VLM perception to recognize connectivity from appearance rather than explicit geometric computation.

A.11 Broader Impact (Extended)

Zero-shot ObjectNav systems are a step toward household robots that follow natural-language category goals without retraining, potentially assisting users with mobility limitations and reducing per-deployment training cost. The same technology could in principle be repurposed for surveillance or unauthorized monitoring of private spaces; we believe deployment should be gated by the usual consent and data-governance protocols. Our framework introduces no new datasets or identifiable human data, and relies only on existing, publicly released simulation benchmarks (HM3D) and pretrained foundation models.

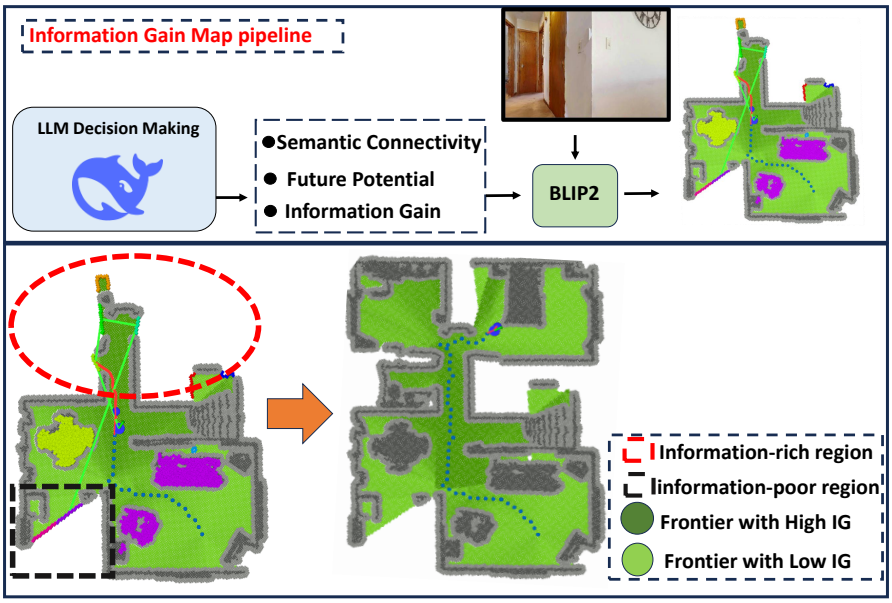


Figure 10: The VLM-based information gain map. The VLM assigns higher scores to the red region (connectivity structures) than to the black region (dead-ends). The agent prioritizes frontiers with high estimated information gain.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe the three technical components (HSVM, VLM-based IG, memory-augmented verification) that are formalized in Section 3 and evaluated in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations (2D value map, empirical IG surrogate, open-vocabulary detection bottleneck) are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper's contributions are methodological and empirical. The Bayesian decomposition (Eq. 1) and the adaptive fusion weight (Eq. 3) are standard derivations rather than novel theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Datasets, simulator, foundation-model versions, hyperparameters, and hardware are listed in Section 4; prompts and thresholds are detailed in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and installation/evaluation instructions are available at `https://github.com/Pandakingxbc/InfoNav`. Third-party datasets and model weights require upstream access; online models require API access.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All components are zero-shot; no training is required. Evaluation protocols (HM3Dv1, HM3Dv2, SR/SPL/SSPL) follow prior work and are documented in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single-run evaluations and do not provide repeated-run standard deviations or significance tests. A fixed episode set does not establish low run-to-run variance; this limitation is acknowledged in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: A single RTX 4070 Ti with an Intel i9-13900K CPU is sufficient; remote DeepSeek-V3, Qwen3-VL, and GPT-5.2 API calls are used for reasoning. See Section 4 and Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No human subjects, no sensitive data, and only publicly available simulation scenes. Potential dual-use considerations are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader-impact discussion in Section 5 covers applications in household robotics and possible misuse scenarios.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: We release navigation code, but no new pretrained models or datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: HM3D [23], HM3D Semantics [29], Habitat [21], BLIP-2 [15], GroundingDINO [16], D-FINE [20], Mobile-SAM [37], DeepSeek-V3 [10], and Qwen3-VL [2] are cited and used under their respective licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The repository documents installation, configuration, and evaluation. Original code uses GPL-3.0; third-party components retain their own terms.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human-subject studies or crowdsourcing were conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subject research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs (DeepSeek-V3, GPT-5.2) and VLMs (Qwen3-VL, GPT-5.2, BLIP-2) are core components of InfoNav and are described in detail in Section 3. LLM/VLM usage for paper writing was limited to editing/grammar assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.