HAVE YOU SEEN IT CLEARLY? UNCERTAINTY-GUIDED BELIEF REFINEMENT FOR ZERO-SHOT OBJECT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-Shot Object Navigation (ZSON) requires an agent to locate target objects in unfamiliar environments without additional training for the test categories or environments. Recent approaches leverage large language models and vision-language models to construct spatial representations and guide exploration. However, whether the available evidence sufficiently supports a navigation choice, and how new evidence changes the search value of previously visited locations, remain insufficiently addressed. We propose RefBeliefMap, an uncertainty-guided mapping framework that distinguishes target-location belief from spatial and semantic evidence uncertainty. By jointly modeling intra-observation uncertainty and inter-observation disagreement over a sliding memory window, RefBeliefMap assesses the reliability of both evidence channels. When competing frontiers remain insufficiently resolved, these uncertainty estimates guide active observation and semantic reasoning to refine the evidence supporting target-location beliefs before navigation commitment. Building on the refined map, we introduce RefBeliefNav, which jointly evaluates unexplored frontiers and historical waypoints according to their target-finding potential, expected information gain, and navigation cost. This unified objective allows historical locations to be reconsidered as new evidence changes their search value, coordinating exploration and revisitation. Experiments on HM3D and MP3D demonstrate state-of-the-art performance among the evaluated zero-shot baselines, with ablation studies examining the contributions of modules. Real-world experiments on a Unitree Go2 quadruped robot further demonstrate the sim-to-real transferability of our method.

1 INTRODUCTION

Zero-shot object navigation (ZSON) requires an embodied agent to locate and reach a target specified by a category name or natural-language description in an unfamiliar environment, without additional training for the test categories or environments. This capability supports applications such as household assistance and autonomous inspection Yokoyama et al. (2024a); Kuang et al. (2024); Cai et al. (2024); Ginting et al. (2024), while open-vocabulary benchmarks extend navigation beyond predefined target categories Yokoyama et al. (2024b). By transferring prior knowledge to new goals and environments, ZSON reduces the need for deployment-specific data collection and model adaptation, offering a scalable approach to navigation across diverse settings.

Recent advances in large language models (LLMs) and vision-language models (VLMs) have enabled ZSON methods to leverage transferable semantic knowledge for spatial search. Many modular approaches comprise two recurring stages: map construction and navigation decision-making. For map construction, VLFM, OpenFMNav, and GAMap encode semantic relevance and affordance cues in two-dimensional maps Yokoyama et al. (2024a); Kuang et al. (2024); Cai et al. (2024); Yuan et al. (2024), while BeliefMapNav integrates semantic priors and observations into a three-dimensional voxel-based belief map Zhou et al. (2025). Complementary representations include open-vocabulary and hierarchical scene graphs Gu et al. (2024); Werby et al. (2024); Yin et al. (2024; 2025) and multi-scale Gaussian-language maps Zhang et al. (2026a). For decision-making, strategies range from greedy frontier selection to topology-aware and probabilistic planning, coarse-to-fine localization, and adaptive exploration Wu et al. (2024); Ginting et al. (2024); Zhang et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2025b; 2024); Cao et al. (2025); Zhang et al. (2025a); Nie et al. (2025). Recent methods further incorporate benefit-aware reasoning and retrieval of past navigation experience Chen et al. (2026); Wang et al. (2026). Despite these advances, two limitations remain insufficiently addressed. First, most existing methods directly rank candidate navigation goals using currently available observations and semantic cues, but rarely assess whether the available evidence is itself sufficient to reliably distinguish among competing choices. Second, although recent methods increasingly preserve or retrieve historical navigation information, they generally do not explicitly re-evaluate the search value of previously visited locations as new evidence changes the agent's belief about the environment. As a result, locations that appeared uninformative when first encountered may remain neglected even after subsequent observations make them relevant to the target search. These limitations motivate explicit assessment of evidence sufficiency during exploration and evidence-driven re-evaluation of historical locations.

We introduce RefBeliefMap, an uncertainty-guided mapping framework that separately represents target-location belief and spatial and semantic evidence uncertainty. A sliding observation window supports temporal evidence assessment, while spatial and semantic uncertainty guide active observation and semantic reasoning, respectively. These complementary mechanisms improve observational coverage and resolve ambiguous interpretations to refine the belief map. Building on this representation, RefBeliefNav jointly evaluates unexplored frontiers and historical waypoints according to their target-finding potential, expected information gain, and navigation cost. By updating candidate values as new evidence becomes available, the planner coordinates exploration and revisitation under a unified objective.

Our contributions are threefold: 1) We introduce RefBeliefMap, which distinguishes target-location belief from spatial and semantic evidence uncertainty, enabling uncertainty-guided active observation and semantic reasoning to refine evidence for navigation decisions. 2) We propose RefBeliefNav, which unifies exploration and revisitation through belief-dependent evaluation of unexplored frontiers and historical waypoints. 3) We validate RefBeliefNav through simulation and real-world experiments, achieving 69.2% SR and 35.3% SPL on HM3D and 49.1% SR and 18.5% SPL on MP3D, outperforming the evaluated zero-shot baselines on both metrics. Deployment on a Unitree Go2 quadruped robot further demonstrates its sim-to-real transferability.

2 RELATED WORK

2.1 OBJECT NAVIGATION

Object navigation aims to guide a robot to locate a specified target object in an unknown environment. Existing approaches broadly fall into two groups. Training-based methods, including reinforcement learning and imitation learning, typically require extensive training on task-specific datasets Ramakrishnan et al. (2022); Ramrakhya et al. (2023); Fukushima et al. (2022); Maksymets et al. (2021); Ye & Yang (2021). Zero-shot methods instead exploit knowledge from pre-trained vision-language models (VLMs) Li et al. (2023); Radford et al. (2021) and large language models (LLMs) to support navigation in unseen environments Kuang et al. (2024); Cao et al. (2025). Following this paradigm, our framework uses LLM-derived semantic priors to initialize target-location beliefs at the start of a task and invokes online reasoning when semantic uncertainty is high to reassess ambiguous cues and inform belief refinement.

2.2 UNCERTAINTY-AWARE OBJECT NAVIGATION

Recent object navigation methods increasingly account for the reliability of semantic maps and accumulated observations. CARe estimates uncertainty through prediction confidence and multi-view consistency to revise decisions based on pre-explored semantic maps Ko et al. (2025). OneMap uses probabilistic semantic updates to mitigate feature-extraction errors and guide exploration, while BeliefMapNav integrates semantic priors and observations into voxel-level target beliefs Busch et al. (2025); Zhou et al. (2025). Extending this perspective to changing environments, IGV-RRT combines uncertainty-aware scene priors with confidence-weighted online semantic observations for object search Zhang et al. (2026b). Recent evidence-driven approaches further connect observation reliability with navigation: AECNav consolidates target and negative evidence into cluster-level log-odds beliefs and actively acquires informative observations, whereas AeroBelief separates con-

textual plausibility from qualified target-specific evidence for aerial ObjectNav Liu et al. (2026); Xiao et al. (2026). These advances establish the importance of evidence accumulation, reliability assessment, and active search. Although these methods assess evidence reliability, they do not fully characterize uncertainty in continuous robot navigation. At each step, uncertainty may arise from insufficient spatial observations or ambiguous semantic cues, while across steps, inconsistencies between current and previous observations introduce an additional source of uncertainty.

2.3 Exploration Strategies for Object Navigation

Exploration strategies for object navigation use geometric and semantic information to prioritize search locations. PONI learns complementary potential functions to guide search, while Semantic Frontiers and ESC incorporate semantic context and commonsense constraints into frontier selection Ramakrishnan et al. (2022); Chen et al. (2023); Zhou et al. (2023). VoroNav uses topological paths to support waypoint reasoning, and SEEK formulates probabilistic search using relational semantic knowledge Wu et al. (2024); Ginting et al. (2024). More recent approaches adapt the search process to available evidence. ApexNav switches between semantic and geometric exploration according to the distribution of semantic cues, while BeliefMapNav optimizes frontier visitation sequences using observation beliefs and travel costs Zhang et al. (2025a); Zhou et al. (2025). Historical observations also support further search: SSMG-Nav maintains persistent semantic memory and revisits mapped locations to recover potential missed detections after exploration is complete Niu et al. (2026).However, frontier sequencing and staged revisiting do not directly establish when revisiting is preferable to further exploration. RefBeliefNav evaluates historical waypoints alongside unexplored frontiers using updated beliefs, expected information gain, and navigation cost, dynamically determining which waypoint to revisit.

3 Method

3.1 Task Definition

We formulate the ZSON task as follows. The robot is required to locate a specified target object in an unfamiliar environment that has not been encountered previously. The target object c is specified by a free-form natural-language description. At each discrete time step t, the agent observes the environment through an RGB-D sensor while estimating its relative state $s_t = (x_t, r_t)$ using on-board odometry, where $x_t \in \mathbb{R}^3$ denotes the agent position and $r_t \in SO(3)$ represents its orientation. Based on the current observation and state estimate, the agent selects a discrete action a_t from the action space: $\mathcal{A}$: = {`MOVE_FORWARD`, `TURN_LEFT`, `TURN_RIGHT`, `LOOK_UP`, `LOOK_DOWN`, `STOP`}. Motion is executed using fixed action increments, $0.25\,\mathrm{m}$ for forward translation and 30° for rotational actions. An episode is considered successful if the agent issues the `STOP` action within a distance threshold of $0.1\,\mathrm{m}$ from the target object before reaching the maximum interaction budget of 500 steps.

3.2 System overview

RefBeliefNav is an uncertainty-guided zero-shot object navigation framework that refines the evidence supporting target-location beliefs before navigation commitment. As illustrated in Fig. 1, the framework consists of three modules: (1) **RefBeliefMap**, which represents both where the target is likely to be and how reliably the available spatial and semantic evidence supports these beliefs, providing a shared basis for refinement and planning, details see in Sec. 3.3; (2) **Evidence Refinement**, which identifies when the current evidence is insufficient for a navigation decision and selectively resolves spatial or semantic uncertainty, helping the agent distinguish promising search regions before committing to a goal, details see in Sec. 3.4; (3) **Unified Search Planning**, which determines where to search next by jointly considering unexplored frontiers and previously visited locations, coordinating exploration and revisitation to balance target discovery, uncertainty reduction, and navigation cost, details see in Sec. 3.5.

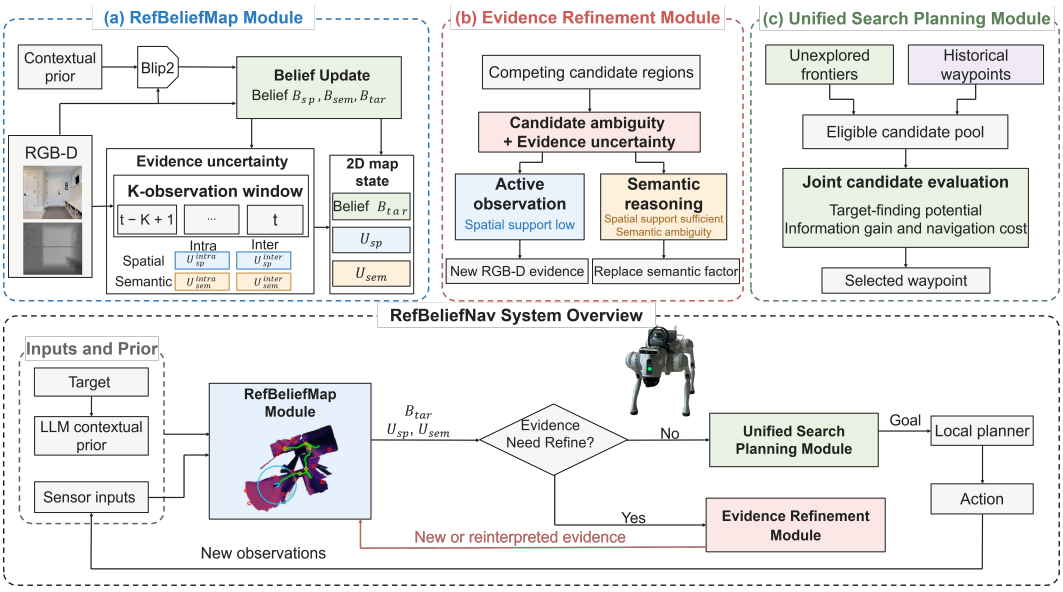

Figure 1: Overview of RefBeliefNav. RefBeliefNav comprises three components: RefBeliefMap models target relevance and evidence uncertainty; evidence refinement improves insufficient support through active observation and semantic reasoning; unified search planning selects among frontiers and historical waypoints.

3.3 REFBELIEFMAP MODULE

RefBeliefMap is a 2D grid map in which each cell maintains three belief quantities, $(B_{\text{tar}}, B_{\text{sem}}, B_{\text{sp}})$, and two uncertainty measures, $(U_{\text{sp}}, U_{\text{sem}})$. The target-location belief B_{tar} is jointly determined by the semantic belief B_{sem} and the spatial belief B_{sp}. Their definitions and update rules are presented in Sec.3.3.1. The uncertainty measures U_{sp} and U_{sem} characterize the reliability of the evidence supporting the spatial and semantic beliefs, respectively. Both account for current and historical observations, as detailed in Sec.3.3.2.

3.3.1 BELIEF REPRESENTATION

For a target category c, RefBeliefMap represents observational evidence through a spatial belief B_{sp} and a semantic belief B_{sem}, which jointly update the target-location belief B_{tar}. Below, we describe a single spatially grounded RGB-D observation region; its spatial support and semantic scores refer to the same region.

Spatial Belief. We model spatial support on the 2D map using a visibility-gated, camera-aligned Gaussian:

$$B_{\text{sp},t}(\mathbf{q}) = \chi_t(\mathbf{q}) \exp\left[-\frac{(r_t(\mathbf{q}) - \mu_{r,t})^2}{2\sigma_{r,t}^2} - \frac{(a_t(\mathbf{q}) - \mu_{a,t})^2}{2\sigma_{a,t}^2} \right]. \tag{1}$$

Here, $r_t(\mathbf{q})$ and $a_t(\mathbf{q})$ are the signed lateral and longitudinal coordinates relative to the camera. The means are determined by the support center obtained from depth and camera pose, while $\sigma_{r,t}, \sigma_{a,t} > 0$ control the support extents. The ray-based visibility gate $\chi_t(\mathbf{q}) \in \{0, 1\}$ excludes invalid or occluded observation associations. Thus, $B_{\text{sp},t}(\mathbf{q}) \in [0, 1]$ measures spatial support rather than target-presence probability. Coordinate definitions, visibility gating, and peak normalization are detailed in Appendix A.1.

Semantic Belief. Given c, an LLM generates a set of target-conditioned prompts $\mathcal{P}_c = \{p_m(c)\}_{m=1}^{M}$. These prompts retain the same target category while varying its descriptions and associated-object cues. Let f_I and f_T denote the Blip-2 Li et al. (2023) image and text encoders.

For the observed RGB region I_t^{rgb}, the semantic relevance under prompt $p_m(c)$ is

$$s_{t,m} = \frac{f_I(I_t^{\mathrm{rgb}})^\top f_T(p_m(c))}{\|f_I(I_t^{\mathrm{rgb}})\|_2 \|f_T(p_m(c))\|_2}. \tag{2}$$

We approximate the distribution of prompt-conditioned scores with a Gaussian whose parameters are $\mu_{\mathrm{sem},t} = \frac{1}{M}\sum_{m=1}^{M} s_{t,m}, \sigma^2_{\mathrm{sem},t} = \frac{1}{M}\sum_{m=1}^{M}(s_{t,m} - \mu_{\mathrm{sem},t})^2 + \varepsilon^2$, where $\varepsilon > 0$ provides a positive variance floor. For cells associated with this observation region, the single-step semantic belief is $B_{\mathrm{sem},t}(\mathbf{q}) = \mathcal{N}\big(\cdot\,; \mu_{\mathrm{sem},t}, \sigma^2_{\mathrm{sem},t}\big)$. The mean summarizes target relevance, whereas the variance captures sensitivity to the prompt set.

Target-Belief Update. Let $b(\mathbf{q})$ denote the latent target-relevance score at location $\mathbf{q}$. We represent its accumulated belief as $B_{\mathrm{tar},t}(\mathbf{q}) = \mathcal{N}\big(\cdot\,; \mu_{\mathrm{tar},t}(\mathbf{q}), \sigma^2_{\mathrm{tar},t}(\mathbf{q})\big)$, with an initial prior specified by $\mu_{\mathrm{tar},0}(\mathbf{q})$ and $\sigma^2_{\mathrm{tar},0}(\mathbf{q}) > 0$. To couple spatial support with semantic evidence, we adopt a spatially weighted Gaussian observation model: $\mu_{\mathrm{sem},t} \mid b(\mathbf{q}) \sim \mathcal{N}\big(b(\mathbf{q}), R_t(\mathbf{q})\big), R_t(\mathbf{q}) = \frac{\sigma^2_{\mathrm{sem},t}}{B_{\mathrm{sp},t}(\mathbf{q})}$,

for $B_{\mathrm{sp},t}(\mathbf{q}) > 0$. Stronger spatial support increases the observation precision. Under a conditionally independent observation approximation, Gaussian fusion gives

$$\begin{aligned}
\sigma^2_{\mathrm{tar},t}(\mathbf{q}) &= \left[\frac{1}{\sigma^2_{\mathrm{tar},t-1}(\mathbf{q})} + \frac{B_{\mathrm{sp},t}(\mathbf{q})}{\sigma^2_{\mathrm{sem},t}}\right]^{-1}, \\
\mu_{\mathrm{tar},t}(\mathbf{q}) &= \sigma^2_{\mathrm{tar},t}(\mathbf{q}) \left[\frac{\mu_{\mathrm{tar},t-1}(\mathbf{q})}{\sigma^2_{\mathrm{tar},t-1}(\mathbf{q})} + \frac{B_{\mathrm{sp},t}(\mathbf{q})\mu_{\mathrm{sem},t}}{\sigma^2_{\mathrm{sem},t}}\right].
\end{aligned} \tag{3}$$

When $B_{\mathrm{sp},t}(\mathbf{q}) = 0$, no observation is incorporated and the previous target belief is retained. The spatial and semantic evidence beliefs are computed from each incoming observation, while B_{tar} accumulates their contributions over time.

3.3.2 Evidence Uncertainty

We assess spatial uncertainty as insufficient visible support and semantic uncertainty as disagreement intra and inter observations. These quantities characterize evidence reliability.

Intra Observation Evidence Uncertainty. Since $B_{\mathrm{sp},t}$ incorporates both Gaussian support and ray-based visibility, we define $U^{\mathrm{intra}}_{\mathrm{sp},t}(\mathbf{q}) = 1 - B_{\mathrm{sp},t}(\mathbf{q})$. Semantic uncertainty measures the dispersion of similarity scores across the M target-conditioned prompts: $U^{\mathrm{intra}}_{\mathrm{sem},t}(\mathbf{q}) = \sigma^2_{\mathrm{sem},t} = \frac{1}{M}\sum_{m=1}^{M}(s_{t,m} - \mu_{\mathrm{sem},t})^2 + \varepsilon^2$. These statistics are associated with $\mathbf{q}$ only when $\chi_t(\mathbf{q}) = 1$.

Inter Observation Evidence Uncertainty. We maintain the latest K observations and identify those providing valid evidence for each location: $\mathcal{W}_t = \{o_i \mid \max(1, t-K+1) \le i \le t\}$, $\mathcal{I}_t(\mathbf{q}) = \{i \mid o_i \in \mathcal{W}_t,\ \chi_i(\mathbf{q}) = 1\}$, $n_t(\mathbf{q}) = |\mathcal{I}_t(\mathbf{q})|$.

(1) Spatial Evidence Uncertainty. Historical uncertainty measures remaining support deficit rather than disagreement:

$$U^{\mathrm{inter}}_{\mathrm{sp},t}(\mathbf{q}) = 1 - \max_{\max(1,t-K+1)\le i<t} B_{\mathrm{sp},i}(\mathbf{q}), \tag{4}$$

where an empty maximum is defined as zero. Combining current and historical support gives

$$U_{\mathrm{sp},t}(\mathbf{q}) = \min\big\{U^{\mathrm{intra}}_{\mathrm{sp},t}(\mathbf{q}), U^{\mathrm{inter}}_{\mathrm{sp},t}(\mathbf{q})\big\} = 1 - \max_{\max(1,t-K+1)\le i\le t} B_{\mathrm{sp},i}(\mathbf{q}). \tag{5}$$

(2) Semantic Evidence Uncertainty. Only observations associated with the same local support and target category are compared. For $n_t(\mathbf{q}) > 0$, define

$$\bar{\mu}_{\mathrm{sem},t}(\mathbf{q}) = \frac{1}{n_t(\mathbf{q})}\sum_{i\in\mathcal{I}_t(\mathbf{q})} \mu_{\mathrm{sem},i}, \quad U^{\mathrm{inter}}_{\mathrm{sem},t}(\mathbf{q}) = \frac{1}{n_t(\mathbf{q})}\sum_{i\in\mathcal{I}_t(\mathbf{q})} \big(\mu_{\mathrm{sem},i} - \bar{\mu}_{\mathrm{sem},t}(\mathbf{q})\big)^2. \tag{6}$$

The total semantic uncertainty is $U_{\mathrm{sem},t}(\mathbf{q}) = \frac{1}{n_t(\mathbf{q})}\sum_{i\in\mathcal{I}_t(\mathbf{q})} U^{\mathrm{intra}}_{\mathrm{sem},i}(\mathbf{q}) + U^{\mathrm{inter}}_{\mathrm{sem},t}(\mathbf{q})$.

3.4 Evidence Refinement Module

3.4.1 Trigger mechanism

We trigger evidence refinement when a promising local candidate lacks sufficient support or competes with another uncertain candidate. Let $\mathcal{C}_t$ contain reachable waypoints within radius R of the agent with valid evidence in RefBeliefMap. Spatially redundant candidates are merged before comparison. For each waypoint w, let $\mathbf{q}_w$ denote its associated evidence location. We define

$$S_t(w) = \mu_{\mathrm{tar},t}(\mathbf{q}_w), H_t(w) = \mathbb{I}[U_{\mathrm{sp},t}(\mathbf{q}_w) > \tau_{\mathrm{sp}} \ \vee \ U_{\mathrm{sem},t}(\mathbf{q}_w) > \tau_{\mathrm{sem}}], \tag{7}$$

where S_t measures target relevance and H_t indicates insufficient evidence. Unavailable semantic uncertainty is treated as insufficient evidence rather than zero uncertainty. Let $w_{(1)}$ and $w_{(2)}$ be the highest-scoring candidates, with $\Delta_t = S_t(w_{(1)}) - S_t(w_{(2)})$. Refinement is triggered when

$$T_t = \mathbb{I}\big[S_t(w_{(1)}) \geq \tau_{\mathrm{rel}}\big] \cdot \mathbb{I}\big[H_t(w_{(1)}) = 1 \ \vee \ \big(\Delta_t \leq \delta \wedge H_t(w_{(2)}) = 1\big)\big]. \tag{8}$$

3.4.2 Active Observation

When spatial support is insufficient, we select a viewpoint that jointly observes the local candidates under a bounded motion cost. Let $\mathcal{Q}_t = \{\mathbf{q}_w \mid w \in \mathcal{C}_t\}$ denote their associated evidence locations. This set is fixed during each viewpoint-selection step, so candidate views are evaluated against the same observation targets. A candidate view is specified by $v = (\mathbf{p}_v, \theta_v)$, comprising a planar camera position and heading. Using the current occupancy map, we predict visibility as

$$\widehat{\chi}_t(\mathbf{q}; v) = \mathbb{I}[d_{\min} \leq \|\mathbf{q} - \mathbf{p}_v\|_2 \leq d_{\max}] \cdot \mathbb{I}[|\alpha_v(\mathbf{q})| \leq \phi_{\mathrm{obs}}/2] \prod_{\mathbf{u} \in \mathcal{R}_v(\mathbf{q})} \mathbb{I}[\mathbf{u} \in \mathcal{F}_t], \tag{9}$$

where $\alpha_v(\mathbf{q})$ is the wrapped bearing relative to θ_v, $[d_{\min}, d_{\max}]$ is the observation distance range, and ϕ_{obs} does not exceed the camera's horizontal field of view. The ray cells $\mathcal{R}_v(\mathbf{q})$ exclude the camera and endpoint cells, and $\mathcal{F}_t$ contains known-free cells. We assign each evidence location a positive weight $\omega_t(\mathbf{q}) = \frac{1+U_{\mathrm{sp},t}(\mathbf{q})}{\sum_{\mathbf{q}' \in \mathcal{Q}_t}[1+U_{\mathrm{sp},t}(\mathbf{q}')]}$. The constant term encourages joint coverage, while spatial uncertainty prioritizes insufficiently supported locations. The selected view solves

$$v_t^\star \in \arg\max_{v \in \mathcal{P}_t} \quad \sum_{\mathbf{q} \in \mathcal{Q}_t} \omega_t(\mathbf{q}) \widehat{\chi}_t(\mathbf{q}; v) \text{s.t.} \quad c_t(v) \leq C_{\max}, \tag{10}$$

where $\mathcal{P}_t$ is a finite set of collision-free, reachable camera poses, and $c_t(v)$ accounts for translation and rotation from the current pose. Ties are resolved by the smaller motion cost. The selected viewpoints are subsequently converted into simulator-executable actions by an action converter, as detailed in the Appendix A.2.

3.4.3 Semantic Reasoning

Semantic reasoning reassesses visible regions with sufficient spatial support but high semantic uncertainty within the triggered neighborhood, while keeping the robot pose fixed. A multimodal LLM evaluates current crops and their BLIP-2 scores (intra-observation) and checks consistency with comparable historical observations of the same region (inter-observation). The prompts are provided in Appendix A.3.

For each region j, the two stages produce reliability scores $r_{t,j}^{\mathrm{intra}}, r_{t,j}^{\mathrm{inter}} \in \{1, \ldots, 10\}$, with higher scores indicating stronger support. We set $\gamma_{t,j} = \min(r_{t,j}^{\mathrm{intra}}, r_{t,j}^{\mathrm{inter}})/10$, or $\gamma_{t,j} = r_{t,j}^{\mathrm{intra}}/10$ when comparable history is unavailable. This heuristic reliability weight tempers the semantic belief for each associated cell $\mathbf{q}$:

$$\widetilde{B}_{\mathrm{sem},t}(\mathbf{q}) = \mathcal{N}\big(\cdot; \mu_{\mathrm{sem},t}, \tfrac{\sigma_{\mathrm{sem},t}^2}{\gamma_{t,j}}\big). \tag{11}$$

This downweights unreliable evidence without changing its mean. The adjusted variance is used for target-belief fusion (Sec. 3.3.1).

3.5 Unified Search Planning Module

Using the refined RefBeliefMap, we jointly evaluate unexplored frontiers and historical waypoints selected for their potential revisitation value. Frontier generation follows VLFM Yokoyama et al. (2024a), while the filtering criteria for historical waypoints are detailed in Appendix A.4.

$\mathcal{V}_t^{\mathrm{fr}}$ and $\mathcal{V}_t^{\mathrm{hist}}$ denote the corresponding reachable candidate sets, and let $\mathcal{V}_t = \mathcal{V}_t^{\mathrm{fr}} \cup \mathcal{V}_t^{\mathrm{hist}}$. The next waypoint is selected by

$$w_t^\star \in \arg\max_{w \in \mathcal{V}_t} \left[J_{\mathrm{tar},t}(w) + \lambda_{\mathrm{unc}} J_{\mathrm{unc},t}(w) - \lambda_{\mathrm{cost}} c_t(w) \right], \tag{12}$$

where the three terms represent target relevance score, potential information benefit, and navigation cost, respectively, and the weights are nonnegative.

Ray-based Target-Relevance Score. For a candidate w at $\mathbf{p}_w$, we model a local 2π yaw scan using uniformly spaced 2D rays within sensing range, terminating at and including the first occupied or unknown cell. Let $\mathcal{E}_t$ denote cells with accumulated target evidence, and let $\mathcal{O}_t(w)$ contain the unique ray-reached cells in $\mathcal{E}_t$ together with first unknown endpoints. Using the distance weight $\kappa_w(\mathbf{q}) = \exp(-\|\mathbf{q} - \mathbf{p}_w\|_2/\ell)$ with $\ell > 0$, we compute $J_{\mathrm{tar},t}(w) = \sum_{\mathbf{q} \in \mathcal{O}_t(w) \cap \mathcal{E}_t} \kappa_w(\mathbf{q}) \mu_{\mathrm{tar},t}(\mathbf{q})$.

Uncertainty-driven Information Benefit. We use the same ray footprint to identify locations where additional observations may improve the evidence. Since $U_{\mathrm{sp},t} \in [0,1]$ while $U_{\mathrm{sem},t}$ is a score variance, we first normalize semantic uncertainty: $\bar{U}_{\mathrm{sem},t}(\mathbf{q}) = U_{\mathrm{sem},t}(\mathbf{q}) / [U_{\mathrm{sem},t}(\mathbf{q}) + \tau_{\mathrm{sem}}]$ using the positive semantic threshold as a scale. The information-benefit score is

$$J_{\mathrm{unc},t}(w) = \sum_{\mathbf{q} \in \mathcal{O}_t(w)} \kappa_w(\mathbf{q}) \Big[U_{\mathrm{sp},t}(\mathbf{q}) + \big(1 - U_{\mathrm{sp},t}(\mathbf{q})\big) \bar{U}_{\mathrm{sem},t}(\mathbf{q}) \Big]. \tag{13}$$

The first term rewards opportunities to improve insufficient spatial support. The second emphasizes semantic ambiguity where spatial evidence is already supported.

Navigation Cost. We use A*Hart et al. (1968) to compute a collision-free path from agent position $\mathbf{p}_t$ to the candidate position $\mathbf{p}_w$ on the map, and define its length as the navigation cost $c_t(w)$.

4 Experimental Results

4.1 Benchmarks and Implementation details

Datasets. We evaluate our method on HM3D Ramakrishnan et al. (2021) and MP3D Chang et al. (2017) using the standard Object Goal Navigation (OGN) benchmark. The HM3D validation split comprises 2,000 episodes across 20 scenes and six object categories, while the MP3D validation split includes 2,195 episodes across 11 scenes and 21 categories.

Evaluation Metrics. We report two standard metrics: Success Rate (SR) and Success weighted by inverse Path Length (SPL). SR measures the proportion of episodes completed successfully, whereas SPL evaluates navigation efficiency by weighting successful episodes according to the ratio between the shortest-path distance and the executed trajectory length.

Implementation Details. We set $K = 5$, $M = 8$, $\sigma_{r,t} = 0.50\,\mathrm{m}$, $\sigma_{a,t} = 1.00\,\mathrm{m}$, and $\varepsilon = 0.02$. The initial target belief uses $\mu_{\mathrm{tar},0}(\mathbf{q}) = 0$ and $\sigma^2_{\mathrm{tar},0}(\mathbf{q}) = 1$. For refinement triggering, we use $R = 3.0\,\mathrm{m}$, $\delta = 0.03$, $\tau_{\mathrm{rel}} = 0.25$, $\tau_{\mathrm{sp}} = 0.50$, and $\tau_{\mathrm{sem}} = 0.01$. For active observation, we choose $d_{\min} = 0.50\,\mathrm{m}$, $d_{\max} = 3.0\,\mathrm{m}$, $\phi_{\mathrm{obs}} = \phi_{\mathrm{cam}}$, and $C_{\max} = 8$, where ϕ_{cam} denotes the fixed horizontal camera field of view. The viewpoint cost $c_t(v)$ counts discrete translation and rotation actions, so $C_{\max}$ is expressed in actions. For unified planning, we set $\ell = 3.0\,\mathrm{m}$, $\lambda_{\mathrm{unc}} = 0.25$, and $\lambda_{\mathrm{cost}} = 5.0\,\mathrm{m}^{-1}$, while $c_t(w)$ remains the A* path length in meters. The reliability factor $\gamma_{t,j}$ remains observation-dependent and is not assigned a fixed default. On the selection of General Models, We use Grounding-DINO Liu et al. (2024) for open-vocabulary object detection and MobileSAM Zhang et al. (2023) for object segmentation. we employ BLIP-2 Li et al. (2023) for semantic similarity representations and GPT-4o for high-level reasoning. All experiments are conducted online on a workstation equipped with a single NVIDIA RTX 4090 GPU.

Table 1: Comparison of RefBeliefNav and zero-shot methods on the HM3D and MP3D datasets.

Method	Venue	HM3D		MP3D	
		SR↑	SPL↑	SR↑	SPL↑
ZSON Majumdar et al. (2022)	NeurIPS'22	25.5	12.6	15.3	4.8
ESC Zhou et al. (2023)	ICML'23	39.2	22.3	28.7	14.2
L3MVN Yu et al. (2023)	IROS'23	50.4	23.1	34.9	14.5
PSL Sun et al. (2024)	ECCV'24	42.4	19.2	18.9	6.9
VLFM Yokoyama et al. (2024a)	ICRA'24	52.5	30.4	36.4	17.5
OpenFMNav Kuang et al. (2024)	NAACL'24	52.5	24.1	37.2	15.7
SG-NAV Yin et al. (2024)	NeurIPS'24	54.0	24.9	40.2	16.0
BeliefMapNav Zhou et al. (2025)	NeurIPS'25	61.4	30.6	37.3	17.6
ASCENT Gong et al. (2026)	RAL'26	65.4	33.5	44.5	15.5
RefBeliefNav	**ours**	**69.2**	**35.3**	**49.1**	**18.5**

4.2 Method Performance

Table 1 compares our RefBeliefNav with representative zero-shot object navigation methods on HM3D and MP3D, descriptions of these baselines are provided in the Appendix A.5. Our RefBeliefNav achieves the highest SR and SPL among the compared methods on both benchmarks. On HM3D, it obtains an SR of 69.2% and an SPL of 35.3%, exceeding ASCENTGong et al. (2026), the strongest baseline on both metrics, by 3.8 and 1.8 percentage points. On MP3D, RefBeliefNav achieves an SR of 49.1%, surpassing ASCENTGong et al. (2026) by 4.6 percentage points. Overall, these results demonstrate consistent gains in SR and SPL efficiency across both datasets.

4.3 Ablative Study

In order to efficiently evaluate the contribution of each module in our method and the sensitivity of the parameters, we randomly sampled and tested 1000 navigation segments from HM3D, using a fixed random seed. We present ablation results for the key modules in the main text, with additional ablation studies and parameter sensitivity analyses provided in the Appendix A.6.

Table 2: Impact of the Evidence Refinement Module.

Method	SR↑	SPL↑
w/o semantic evidence refinement	65.2	33.9
w/o spatial evidence refinement	64.5	33.7
w/o evidence refinement	61.4	31.3
RefBeliefNav	**69.5**	**35.5**

Table 3: Ablation of Unified Search Planning.

Method	SR↑	SPL↑
Frontier-only candidates	64.6	33.9
w/o information benefit term	58.2	33.8
w/o navigation cost	66.7	30.4
w/o target-relevance score	56.1	31.2
RefBeliefNav	**69.5**	**35.5**

Effectiveness of Evidence Refinement Module. Table 2 shows that removing either semantic or spatial refinement lowers SR and SPL. Disabling both yields the largest drops of 8.1 and 4.2 percentage points, respectively, supporting the joint use of both mechanisms.

Effectiveness of Unified Search Planning. Table 3 shows that removing target relevance or information benefit reduces SR to 56.1% or 58.2%, respectively, highlighting their complementary roles in balancing exploration and exploitation. Removing navigation cost lowers SPL from 35.5% to 30.4%, while frontier-only candidates reduce SR to 64.6%, supporting cost-aware planning and the use of historical waypoints. The full planner achieves the highest SR and SPL.

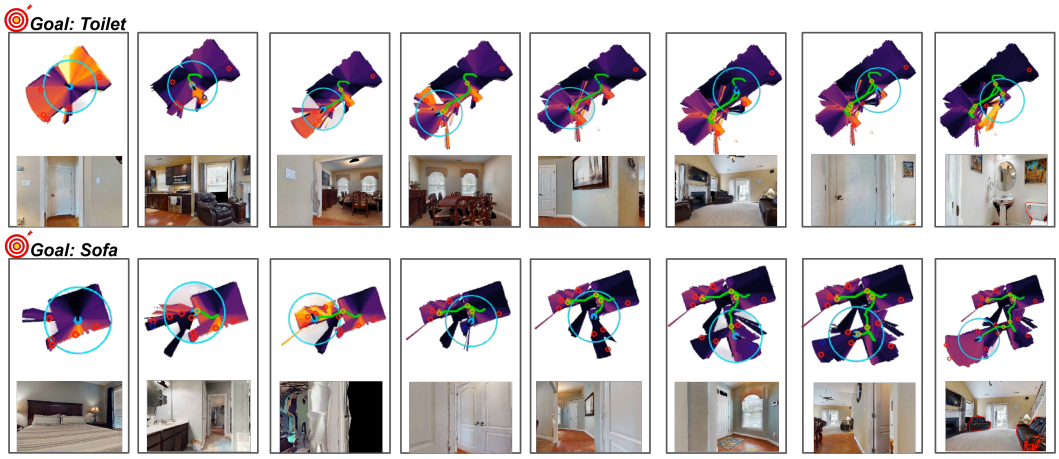

Figure 2: Qualitative examples of RefBeliefNav searching for a toilet (top) and a sofa (bottom) in simulation. Each panel shows the evolving RefBeliefMap and the corresponding egocentric RGB observation. Orange, red, and dark blue hollow circles denote historical waypoints, frontiers, and the selected waypoint, respectively. The light blue circle marks the radius-R neighborhood centered on the robot used to evaluate refinement triggers (Sec. 3.4.1).

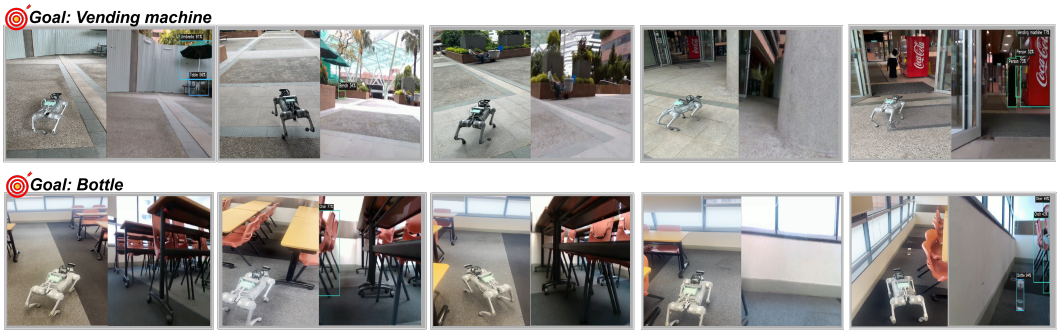

Figure 3: Real-world searches for a vending machine in an open public area (top) and a bottle in a classroom (bottom). The sequences progress from left to right, with each panel pairing an external view of the robot with its egocentric RGB observation. Bounding boxes indicate object detections.

4.4 Qualitative Analysis

In simulation (Fig. 2), RefBeliefNav jointly considers target beliefs and spatial and semantic evidence uncertainty to guide exploration and revisitation. Real-world deployment on the Unitree Go2 (Fig. 3) further illustrates how position and viewpoint adjustments reveal initially unseen targets under clutter and limited visibility. Additional examples and complete navigation sequences are provided in the supplementary video.

5 Conclusion

We presented RefBeliefNav, an uncertainty-guided framework for zero-shot object navigation. Its RefBeliefMap distinguishes target beliefs from spatial and semantic evidence uncertainty, guiding active observation and semantic reasoning to refine the evidence supporting navigation decisions. Joint evaluation of frontiers and historical waypoints unifies exploration and revisitation. Experiments on HM3D and MP3D demonstrate higher SR and SPL than the evaluated baselines, while deployment on a Unitree Go2 demonstrates real-world applicability.

AI USE STRATEGYMENT

In this work, we used generative AI tools (e.g., large language models) for manuscript language editing, grammar refinement, and code debugging. We have not used generative AI tools for generating core research ideas, designing methodologies, or creating experimental data. We have thoroughly reviewed all AI-assisted text and code to ensure their accuracy, correctness, and originality. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

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)*, pp. 14835–14842, 2025. doi: 10.1109/ICRA55743.2025.11128393.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5228–5234, 2024.

Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. Cognav: Cognitive process modeling for object goal navigation with llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9550–9560, 2025.

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. *arXiv preprint arXiv:1709.06158*, 2017.

Hanrui Chen, Liqi Yan, Qifan Wang, Jianhui Zhang, Fangli Guan, and Pan Li. Chain-of-Search: Parameter-efficient reasoning for zero-shot object navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1721–1729, 2026.

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 *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Rui Fukushima, Kei Ota, Asako Kanezaki, Yoko Sasaki, and Yusuke Yoshiyasu. Object memory transformer for object goal navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11288–11294, 2022.

Muhammad Fadhil Ginting, Sung-Kyun Kim, David D. Fan, Matteo Palieri, Mykel J. Kochenderfer, and Ali akbar Agha-Mohammadi. SEEK: Semantic reasoning for object goal navigation in real world inspection tasks. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Guoyang Zhao, Yiyi Ding, and Junwei Liang. Stairway to success: An online floor-aware zero-shot object-goal navigation framework via llm-driven coarse-to-fine exploration. *IEEE Robotics and Automation Letters (RAL)*, 2026.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024.

Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Po-Chen Ko, Hung-Ting Su, CY Chen, Jia-Fong Yeh, Min Sun, and Winston H. Hsu. Context-aware replanning with pre-explored semantic map for object navigation. In *Proceedings of the 8th Conference on Robot Learning*, volume 270, pp. 4253–4267, 2025.

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*, pp. 338–351, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pp. 12888–12900, 2022.

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 *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

Guanlin Liu, Shaobin Ling, Renyuan Liu, Zeying Gong, and Junjie Hu. AECNav: Active evidence consolidation for efficient zero-shot open-vocabulary object navigation. *arXiv preprint arXiv:2608.10817*, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:32340–32352, 2022.

Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. Thda: Treasure hunt data augmentation for semantic navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15354–15363, 2021.

Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2392–2399, 2025.

Haochen Niu, Lantao Zhang, Xingwu Ji, Rendong Ying, Peilin Liu, and Fei Wen. SSMG-Nav: Enhancing lifelong object navigation with semantic skeleton memory graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18868–18878, 2022.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17896–17906, 2023.

Xinyu Sun, Lizhao Liu, Hongyan Zhi, Ronghe Qiu, and Junwei Liang. Prioritized semantic learning for zero-shot instance navigation. In *European Conference on Computer Vision (ECCV)*, pp. 161–178, 2024.

Yiyao Wang, Sixian Zhang, Keming Zhang, Xinhang Song, Songjie Du, and Shuqiang Jiang. Trajrag: Retrieving geometric-semantic experience for zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15166–15176, 2026.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

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 International Conference on Machine Learning (ICML)*, pp. 53737–53775, 2024.

Jianqiang Xiao, Xiang Deng, Yuexuan Sun, Yanjin Wu, Wenbiao Yan, and Liqiang Nie. Dual-layer semantic-spatial belief mapping for aerial object goal navigation. *arXiv preprint arXiv:2609.08164*, 2026.

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.

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. volume 37, pp. 5285–5307, 2024.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19057–19066, 2025.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48, 2024a.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5543–5550, 2024b.

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)*, pp. 3554–3560, 2023.

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)*, volume 37, pp. 39386–39408, 2024.

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.

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)*, pp. 10035–10042, 2024.

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 (RAL)*, 10(11):11530–11537, 2025a.

Sixian Zhang, Xinyao Yu, Xinhang Song, Yiyao Wang, and Shuqiang Jiang. Function-centric bayesian network for zero-shot object goal navigation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19535–19545, 2025b.

Sixian Zhang, Yiyao Wang, Xinhang Song, Keming Zhang, Zijian Xu, and Shuqiang Jiang. Multi-scale gaussian-language map for zero-shot embodied navigation and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37086–37097, 2026a.

Wei Zhang, Ping Gong, Yujie Wang, Leilei Yao, Minghui Bai, Rongfeng Ye, Yinchuan Wang, Yachao Wang, Chen Sun, and Chaoqun Wang. Igv-rrt: Prior-real-time observation fusion for active object search in changing environments. *arXiv preprint arXiv:2603.21887*, 2026b.

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 *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 42829–42842, 2023.

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)*, volume 38, pp. 83008–83036, 2025.

A Appendix

A.1 Spatial Support Construction

Camera-Aligned Coordinates. Let $\mathbf{p}_t$ denote the camera position on the 2D map. Let $\mathbf{e}_{r,t}$ and $\mathbf{e}_{a,t}$ be orthogonal unit vectors perpendicular and parallel to the camera heading, respectively. For a map location $\mathbf{q}$, we define

$$\begin{aligned} r_t(\mathbf{q}) &= \mathbf{e}_{r,t}^\top(\mathbf{q} - \mathbf{p}_t), \\ a_t(\mathbf{q}) &= \mathbf{e}_{a,t}^\top(\mathbf{q} - \mathbf{p}_t). \end{aligned} \tag{14}$$

Given the support center $\mathbf{q}_t^0$ obtained from depth and camera pose, the two Gaussian components are

$$\begin{aligned} G_{r,t}(r) &= \mathcal{N}(r; \mu_{r,t}, \sigma_{r,t}^2), \\ G_{a,t}(a) &= \mathcal{N}(a; \mu_{a,t}, \sigma_{a,t}^2), \end{aligned} \tag{15}$$

where $\mu_{r,t} = r_t(\mathbf{q}_t^0)$ and $\mu_{a,t} = a_t(\mathbf{q}_t^0)$. The positive bandwidths $\sigma_{r,t}$ and $\sigma_{a,t}$ determine the lateral and longitudinal support extents.

Ray-Based Visibility Gating. To prevent support from extending behind occluders, we cast a ray from $\mathbf{p}_t$ to $\mathbf{q}$ on the occupancy map. Let $\mathcal{A}_t$ contain cells within the effective field of view and sensing range that admit a valid association with the current RGB-D observation. Let $\mathcal{F}_t$ contain known-free cells after incorporating the current depth observation. The set $\mathcal{R}_t(\mathbf{q})$ contains the intermediate ray cells, excluding the camera cell and the endpoint cell. The visibility gate is

$$\chi_t(\mathbf{q}) = \mathbb{I}[\mathbf{q} \in \mathcal{A}_t] \prod_{\mathbf{u} \in \mathcal{R}_t(\mathbf{q})} \mathbb{I}[\mathbf{u} \in \mathcal{F}_t], \tag{16}$$

where $\mathbb{I}[\cdot]$ is the indicator function and an empty product equals one. An occupied or unknown intermediate cell blocks the association, whereas an occupied endpoint can still receive support. This test is a planar approximation under the 2D occupancy representation.

Peak-Normalized Spatial Support. We combine the Gaussian components through their peak-normalized product:

$$B_{\mathrm{sp},t}(\mathbf{q}) = \chi_t(\mathbf{q}) \frac{G_{r,t}(r_t(\mathbf{q}))G_{a,t}(a_t(\mathbf{q}))}{G_{r,t}(\mu_{r,t})G_{a,t}(\mu_{a,t})}. \tag{17}$$

Expanding the Gaussian ratio gives Eq. 1. Peak normalization bounds the support score in $[0, 1]$, while visibility gating suppresses invalid or occluded associations.

A.2 Action Transitions

Path Following and View Alignment. The action converter translates a selected waypoint or camera view into discrete simulator actions. We first compute an A* path through reachable, collision-free space on the current occupancy map. For the next local path target $\mathbf{g}$, the heading error is $e_\theta = \mathrm{wrap}(\mathrm{atan2}(g_y - p_y, g_x - p_x) - \theta)$, where $(\mathbf{p}, \theta)$ is the current planar pose and wrap maps angles to $(-\pi, \pi]$. The controller issues `TURN_LEFT` or `TURN_RIGHT` when $|e_\theta| > 15°$; otherwise, it issues `MOVE_FORWARD` if the predicted motion is collision-free. Each forward action advances $0.25\,\mathrm{m}$, and each turn changes the heading by $30°$. The pose and local target are updated after each action, and an invalidated path is replanned.

For an active-observation view $v = (\mathbf{p}_v, \theta_v)$, the controller first reaches $\mathbf{p}_v$ within its positional tolerance and then performs in-place turns to the nearest executable heading to θ_v. Candidate visibility is evaluated at the corresponding discretized terminal pose, rather than an unattainable continuous pose. Although the task action space includes `LOOK_UP` and `LOOK_DOWN`, this planar converter controls translation and yaw while leaving camera pitch unchanged.

Action Accounting and Feedback. Let $\boldsymbol{a}_t(v) = (a_1, \ldots, a_{N_v})$ denote the planned action sequence for view v. The active-observation cost is $c_t(v) = N_v$, including forward actions, intermediate turns, and final heading alignment. Only views satisfying $N_v \le C_{\max}$ are eligible, with $C_{\max} = 8$ by default. All executed actions count toward both the refinement budget and the episode's interaction budget; replanning does not reset the remaining refinement budget.

A.3 LLMs PROMPTS

Target-Conditioned description generation. The semantic belief uses M descriptions of the same target category. The descriptions vary visual attributes and associated-object cues without changing the target. The resulting prompt set is kept fixed across observations of that category. For the prompt-count analysis, smaller sets are prefixes of the same ordered description list.

> **Input:** Target category `TARGET`; required number of descriptions `M`.
>
> **Instruction:** Generate `M` concise visual descriptions for image–text matching of `TARGET`. Each description must explicitly refer to the same target category. Vary its visual description and plausible associated-object context, while keeping the target central. Do not replace the target with a related object or treat contextual cues as proof that the target is present. Do not assume any particular scene or target location. Return only an ordered list of descriptions.

Inputs for Semantic Reassessment. Within a triggered neighborhood, currently visible cells with sufficient spatial support and high semantic uncertainty are mapped back to RGB pixels using depth and camera pose. Connected pixel regions form padded crops, with the selected evidence highlighted. Each crop is supplied with the target category and its associated BLIP-2 prompt–score pairs. Historical inputs are restricted to observations associated with the same map region and target category.

Shared Reliability Instruction. Both assessment stages evaluate the reliability of the existing prompt-conditioned semantic assessment, not the probability that the target is present.

> **Instruction:** Assess whether the supplied visual evidence supports the current BLIP-2 target-relevance assessment. The similarity scores are not calibrated probabilities. Distinguish direct target evidence, contextual cues, and visually similar non-target objects. Use an integer reliability score from 1 to 10: 1 indicates that the assessment is unsupported or contradicted, and 10 indicates that it is clearly supported. A high score does not necessarily imply target presence. For example, a clearly non-target region can receive high reliability when its existing assessment also indicates low target relevance. Provide a brief evidence summary rather than a replacement similarity score.

Intra-Observation Assessment.

> **Input:** The target category, current highlighted crop, and associated BLIP-2 prompt–score pairs.
>
> **Instruction:** Evaluate whether the current semantic assessment is justified by the highlighted region. Use the surrounding crop only as context. Identify visible target-specific features, related-object cues, and plausible confounders. Check whether the score pattern is supported by the image or is driven mainly by contextual associations. Account for occlusion, truncation, and insufficient detail when judging reliability. Return `r_intra`, an integer from 1 to 10, and a one-sentence evidence summary.

Inter-Observation Assessment.

> **Input:** The current crop, prompt–score pairs, and intra-observation assessment, together with available historical crops and their associated scores for the same map region and target category.
>
> **Instruction:** First determine whether the historical observations provide comparable evidence for the highlighted region. Then assess whether they support or contradict the current semantic assessment. Distinguish genuine contradictions from changes caused by viewpoint, occlusion, or incomplete image coverage. An object not being visible in another view is not, by itself, evidence of contradiction. Return `comparable`, `r_inter`, and a one-sentence evidence summary. If no comparable historical evidence is available, set `comparable` to false and `r_inter` to null. Otherwise, return an integer `r_inter` from 1 to 10 using the shared reliability criterion.

Use of the Outputs. For region j, the scores determine $\gamma_{t,j} = \min(r^{\text{intra}}_{t,j}, r^{\text{inter}}_{t,j})/10$; without comparable history, $\gamma_{t,j} = r^{\text{intra}}_{t,j}/10$. The factor changes the semantic variance to $\sigma^2_{\text{sem},t}/\gamma_{t,j}$ while

leaving $\mu_{\text{sem},t}$ unchanged. The original prompt statistics remain available for uncertainty assessment. Reassessment replaces the semantic contribution of the corresponding observation rather than adding a new, independent measurement.

A.4 Rules for Generating Historical Waypoints

After each movement step, the agent evaluates whether to store its current position $\mathbf{p}_t$ as a historical waypoint, considering both observation coverage and the step interval since the last waypoint insertion.

Let $\mathcal{D}_t$ denote the map cells within the current camera field of view and sensing range before occlusion filtering. Applying the visibility gate χ_t to the full current RGB-D observation, we measure spatial coverage as

$$C_{\text{view},t} = \frac{\sum_{\mathbf{q}\in\mathcal{D}_t} \chi_t(\mathbf{q})}{\max(1, |\mathcal{D}_t|)}. \tag{18}$$

This score measures the fraction of the nominal viewing footprint with valid, unoccluded observation support.

Let Δ_t be the number of navigation steps since the most recent historical waypoint was stored. We define the admission score as

$$S_{\text{hist},t} = C_{\text{view},t} \frac{\Delta_t}{\Delta_t + T_{\text{hist}}}, \qquad T_{\text{hist}} > 0, \tag{19}$$

where T_{hist} controls how quickly the step-dependent weight increases. The current position is stored when $S_{\text{hist},t} \geq \tau_{\text{hist}}$, with $\tau_{\text{hist}} \in (0, 1)$. Thus, broader observation coverage favors admission, while a longer step interval increases the admission score for the same coverage.

The step counter resets after each insertion; before the first insertion, it counts from the episode start. Stored positions form the waypoint memory, whose currently reachable entries constitute $\mathcal{V}_t^{\text{hist}}$ for joint evaluation with frontier waypoints.

A.5 Baseline Description

We compare RefBeliefNav with nine zero-shot ObjectNav baselines. ZSON (Majumdar et al., 2022) transfers image-goal policies to text-specified goals. ESC (Zhou et al., 2023) guides exploration through commonsense-based soft constraints, while L3MVN (Yu et al., 2023) uses language-model priors for frontier selection. PSL (Sun et al., 2024) improves semantic generalization through prioritized image-goal training. VLFM (Yokoyama et al., 2024a) guides frontier exploration using a vision-language value map. OpenFMNav (Kuang et al., 2024) supports free-form instructions through semantic score mapping. SG-Nav (Yin et al., 2024) combines hierarchical scene graphs with LLM reasoning. BeliefMapNav (Zhou et al., 2025) integrates semantic and spatial evidence into a 3D target-belief map. ASCENT (Gong et al., 2026) addresses multi-floor navigation through floor-aware mapping and coarse-to-fine reasoning.

A.6 Additional Ablation Studies.

Effectiveness of LLMs and VLMs. As shown in Table 4, GPT-4o achieves the highest SR (69.5%) and SPL (35.5%), followed by Gemini with 69.1% SR and 35.3% SPL, and Qwen2.5-VL with 68.9% SR and 34.9% SPL. The relatively small differences suggest that RefBeliefNav maintains comparable performance across the evaluated LLM backbones. Table 5 compares different VLM backbones. BLIP-2 Li et al. (2023) achieves the highest SR (69.5%) and SPL (35.5%), slightly outperforming CLIP Radford et al. (2021) at 69.3% SR and 35.2% SPL. Both outperform BLIP, which obtains 68.2% SR and 35.1% SPL.

Effectiveness of the Trigger Mechanism. As shown in Table 6, adaptive triggering achieves the highest SR (69.5%) and SPL (35.5%) with only 8.2 LLM calls. In comparison, triggering every 10 steps yields 66.7% SR and 33.5% SPL with 37.2 calls, while triggering every 30 steps achieves 66.1% SR and 33.2% SPL with 13.2 calls. These results suggest that refinement timing matters more than frequency alone: selectively reassessing ambiguous evidence can improve navigation success and efficiency while reducing LLM usage.

Table 4: Impact of the LLMs.

Model	SR↑	SPL↑
Gemini	69.1	35.3
Qwen2.5-VL	68.9	34.9
GPT-4o	**69.5**	**35.5**

Table 5: Impact of the VLMs.

Model	SR↑	SPL↑
BlipLi et al. (2022)	68.2	35.1
ClipRadford et al. (2021)	69.3	35.2
Blip-2 Li et al. (2023)	**69.5**	**35.5**

Table 6: Impact of refinement-triggering strategy.

Trigger Strategy	LLM Calls↓	SR↑	SPL↑
Fixed steps (per 10)	37.2	66.7	33.5
Fixed steps (per 30)	13.2	66.1	33.2
Adaptive	**8.2**	**69.5**	**35.5**

Observation Window and Semantic Statistics. Table 7 varies the observation-window length K, prompt count M, and semantic standard-deviation floor ε. Changing K affects recent-evidence uncertainty estimation, but does not reset the recursively accumulated target belief or delete archived observations. The prompt-count comparison uses nested prompt sets so that larger sets retain the prompts from smaller sets. The variance-floor comparison examines the balance between excessive observation precision and overly weak semantic updates.

Spatial Gaussian Bandwidths. Table 8 varies the radial and axial support bandwidths independently. These parameters control the spatial extent of evidence association. Smaller bandwidths concentrate support around the projected center, whereas larger bandwidths spread the same observation over a broader map region.

Trigger Neighborhood and Belief Thresholds. Table 9 varies the local radius R, belief-gap threshold δ, and relevance threshold τ_{rel}. The radius determines which nearby waypoints participate in the trigger decision. The gap threshold controls when competing waypoint beliefs are considered similar, while the relevance threshold prevents refinement of candidates with insufficient target relevance. These comparisons examine the balance between missing potentially useful refinement opportunities and activating refinement too broadly.

Active-Observation Geometry and Action Budget. Table 10 examines the admissible observation distances, angular coverage, and action budget. The distance bounds are restricted to the valid sensing range. The angular comparison changes the visibility criterion through $\phi_{\mathrm{obs}}/\phi_{\mathrm{cam}}$ without altering the actual camera field of view. The action-budget comparison varies the maximum number of simulator actions available to one active-observation operation, while retaining the same overall navigation budget.

Planning Weights and Distance Decay. Table 11 varies the uncertainty weight λ_{unc}, navigation-cost weight λ_{cost}, and distance-decay scale ℓ. The zero-weight configurations disable the corresponding planning terms examined in Table 3. These comparisons assess the contributions of uncertainty-driven exploration and navigation-cost regularization to search success and path efficiency.

Spatial and Semantic Uncertainty Thresholds. Table 12 varies τ_{sp} and τ_{sem}. Lower thresholds make the corresponding refinement branch eligible more frequently, whereas higher thresholds restrict refinement to stronger evidence deficiencies. Because τ_{sem} also appears in semantic uncertainty normalization for search planning, its sweep changes both refinement triggering and planning scores.

Joint Parameter Sensitivity. Table 13 supplements the one-factor comparisons with two local parameter grids. The first varies λ_{unc} jointly with λ_{cost} to examine the interaction between uncertainty-driven exploration and travel efficiency. The second varies δ jointly with τ_{sp} to examine the interac-

Table 7: Sensitivity analysis of the observation window and semantic statistics on HM3D.

Parameter	Value	SR↑	SPL↑
K	1	66.3	33.4
	3	68.9	34.7
	5	**69.5**	**35.5**
	10	68.9	34.5
	20	69.1	34.9
M	2	66.3	33.6
	4	68.3	34.8
	8	**69.5**	**35.5**
	16	69.4	35.2
ε	0.002	68.1	34.4
	0.006	68.8	35.1
	0.020	**69.5**	**35.5**
	0.060	68.7	34.9
	0.200	65.3	32.6

Table 8: Sensitivity analysis of spatial support bandwidths on HM3D.

Parameter	Value	SR↑	SPL↑
$\sigma_{r,t}$	0.250	66.8	34.0
	0.375	68.5	35.0
	0.500	**69.5**	**35.5**
	0.750	69.3	35.1
	1.000	68.6	34.4
$\sigma_{a,t}$	0.50	66.5	33.8
	0.75	68.6	35.0
	1.00	**69.5**	**35.5**
	1.50	69.3	35.0
	2.00	68.1	34.1

Table 9: Sensitivity analysis of trigger neighborhood and belief thresholds.

Parameter	Value	SR↑	SPL↑
R (m)	1.50	66.7	34.0
	2.25	68.7	35.1
	3.00	**69.5**	**35.5**
	4.50	69.4	34.8
	6.00	68.6	33.6
δ	0.01	67.2	34.3
	0.02	68.6	35.1
	0.03	**69.5**	**35.5**
	0.05	69.3	35.0
	0.08	68.5	34.1
τ_{rel}	0.15	68.8	34.1
	0.20	69.3	35.0
	0.25	**69.5**	**35.5**
	0.30	68.2	34.9
	0.35	66.4	33.8

Table 10: Sensitivity analysis of active-observation geometry and action budget.

Parameter	Value	SR↑	SPL↑
$d_{\min}$ (m)	0.25	68.4	34.6
	0.50	**69.5**	**35.5**
	0.75	68.8	35.0
	1.00	67.5	34.2
$d_{\max}$ (m)	2.0	67.2	34.3
	3.0	**69.5**	**35.5**
	4.0	69.3	35.0
	5.0	68.8	34.5
$\phi_{\mathrm{obs}}/\phi_{\mathrm{cam}}$	0.50	66.9	33.7
	0.75	68.6	34.8
	1.00	**69.5**	**35.5**
$C_{\max}$ (actions)	4	66.8	34.2
	6	68.7	35.1
	8	**69.5**	**35.5**
	12	69.4	34.7
	16	69.3	33.6

tion between waypoint-belief ambiguity and spatial evidence deficiency. Entries along the reference row or column match the corresponding one-factor tables.

Table 11: Sensitivity analysis of planning weights and distance decay.

Parameter	Value	SR↑	SPL↑
λ_{unc}	0.000	58.2	33.8
	0.125	68.4	35.0
	0.250	**69.5**	**35.5**
	0.500	68.8	34.2
	1.000	66.7	32.6
λ_{cost} (m^{-1})	0.0	66.7	30.4
	2.5	69.3	34.8
	5.0	**69.5**	**35.5**
	10.0	67.8	35.1
	20.0	64.8	33.9
ℓ (m)	1.0	66.5	34.2
	2.0	68.6	35.1
	3.0	**69.5**	**35.5**
	4.5	69.0	34.8
	6.0	68.3	34.1

Table 12: Sensitivity analysis of uncertainty thresholds.

Parameter	Value	SR↑	SPL↑
τ_{sp}	0.30	69.4	34.1
	0.40	69.3	35.0
	0.50	**69.5**	**35.5**
	0.60	68.4	34.9
	0.70	66.8	34.0
τ_{sem}	0.0025	68.6	33.9
	0.0050	69.1	34.9
	0.0100	**69.5**	**35.5**
	0.0200	68.2	34.5
	0.0400	66.5	33.5

Table 13: Joint parameter sensitivity: (a) planning weights and (b) trigger thresholds. λ_{cost} is expressed in m^{-1}.

(a) Planning weights

λ_{unc}	λ_{cost}	SR↑	SPL↑
0.125	2.5	68.7	34.7
0.125	5.0	68.4	35.0
0.125	10.0	67.0	34.5
0.250	2.5	69.3	34.8
0.250	**5.0**	**69.5**	**35.5**
0.250	10.0	67.8	35.1
0.500	2.5	69.1	33.8
0.500	5.0	68.8	34.2
0.500	10.0	68.0	34.5

(b) Trigger thresholds

δ	τ_{sp}	SR↑	SPL↑
0.02	0.40	68.9	34.9
0.02	0.50	68.6	35.1
0.02	0.60	67.9	34.7
0.03	0.40	69.3	35.0
0.03	**0.50**	**69.5**	**35.5**
0.03	0.60	68.4	34.9
0.05	0.40	69.4	34.6
0.05	0.50	69.3	35.0
0.05	0.60	68.7	34.8