EGNAV: EVIDENCE-GUIDED MEMORY REPAIR FOR ZERO-SHOT GOAL-ORIENTED NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot goal-oriented navigation requires more than exploring new regions: a previously visited location may still contain a target that was misidentified or remained outside the agent's view. This ambiguity makes it difficult to determine whether to continue exploring or revisit a potential target location. We propose **EGNav**, a zero-shot navigation framework that treats goal search as sequential evidence management. At its core is a task-conditioned evidence memory that associates spatially grounded candidates with direct, contextual, and negative evidence. Source- and visibility-aware updates maintain evolving belief and uncertainty, distinguishing contextual support from direct target matches and reducing the influence of negative observations obtained under limited visibility. A unified memory-repair policy then ranks unexplored frontiers and memory candidates using accumulated belief, expected information gain, and execution costs. Within this policy, direct repair re-examines previously observed target-like instances, while contextual repair revisits locations that may contain an unseen target. Both modes update the same evidence state and participate in a shared decision process. Experiments across object-, image-, and text-goal navigation on HM3D and MP3D show that EGNav outperforms the compared universal zero-shot methods, with further gains over the compared training-free baselines on GOAT-Bench. Ablations support the complementarity of the two repair modes and the contributions of evidence updating and candidate ranking.

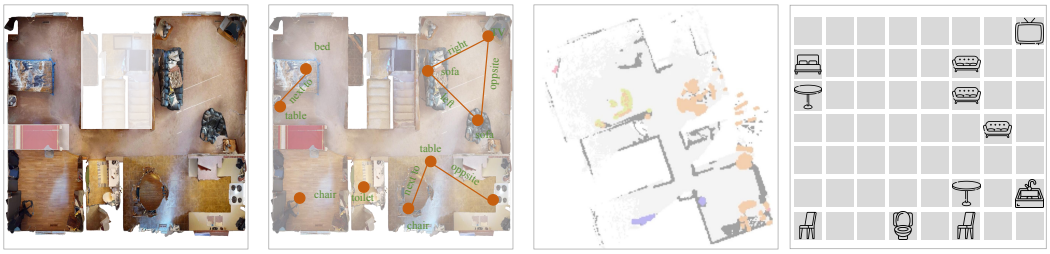
1 INTRODUCTION

Goal-oriented navigation requires an embodied agent to locate a specified target in an unknown environment. Depending on the goal specification, this includes object-goal navigation (ON) from an object category, instance-image navigation (IIN) from a reference image, and text navigation (TN) from a natural-language description. Recent advances in large language models (LLMs) (Achiam et al., 2023; Meta, 2024) and vision-language models (VLMs) (Bai et al., 2025a;b) have enabled zero-shot approaches that leverage semantic priors for goal reasoning and exploration without task-specific training (Zhou et al., 2023; Yin et al., 2025; Zhang et al., 2025b). However, effective search requires more than discovering new regions: an agent must also determine what previous observations imply about the target and whether an explored location warrants another visit.

Scene representations provide an important foundation for goal-directed search. Graph-based approaches (Yin et al., 2024; 2025) encode objects and their spatial relations, whereas map-based methods (Yu et al., 2023; Yokoyama et al., 2024; Kuang et al., 2024) associate semantic information with the spatial layout, as illustrated in Figure 1(b,c). While these representations organize scene information, recording an explored location does not by itself establish whether the goal has been reliably ruled out. An agent may observe the target but fail to match it under viewpoint or appearance changes, or miss it because of occlusion even when surrounding objects suggest its presence. Moreover, a missed detection under limited visibility provides weaker grounds for exclusion than one obtained from a clear, well-covered view. Treating different observations as equivalent failures can therefore discard useful evidence and overlook target locations that remain worth revisiting.

We therefore argue that navigation should distinguish between visiting a region and resolving whether it contains goal. Agent needs to retain task-relevant evidence, revise its confidence as observations accumulate, and compare the value of revisiting a known location with that of exploring a

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.



(a) Scene **(b) Scene graph** **(c) Semantic map** **(d) Evidence memory**

Figure 1: Scene representations and task-conditioned evidence memory for zero-shot goal-oriented navigation. (a) An example scene; (b) a scene graph; (c) a semantic occupancy map; and (d) the spatial organization of our evidence memory. Each memory candidate is associated with stored observations and an evolving evidence state, omitted here for clarity. Our memory augments the navigation backend rather than replacing its geometric map.

new one. To this end, we propose **EGNav**, a zero-shot navigation framework that treats goal search as sequential evidence management. Rather than replacing the geometric map or local planner, our framework augments the navigation backend with a task conditioned evidence memory and a unified memory repair policy, connecting accumulated observations to subsequent search decisions.

Specifically, the memory maintains spatially grounded candidate hypotheses, each associated with stored observations and evolving evidence about whether the goal can be found nearby. Figure 1(d) illustrates its spatial organization. We distinguish direct target evidence, contextual support, and negative observations, and maintain belief and uncertainty through source- and visibility-aware updates. This design gives limited or occluded observations less influence than clear views, while distinguishing contextual cues from direct target matches. A unified policy then ranks memory candidates and unexplored frontiers using accumulated belief, expected information gain, and execution costs. Within this policy, direct repair re-examines previously observed target-like instances, while contextual repair revisits locations that may contain a target not yet observed. Both modes update the same evidence state and participate in the same decision process, rather than operating as separate navigation algorithms. Repetition-aware control discourages unsuccessful repeated visits while allowing candidates to be reconsidered when new evidence becomes available. Experiments covering ON, IIN, and TN on HM3D (Ramakrishnan et al., 2021) and MP3D (Chang et al., 2017) demonstrate that EGNav achieves the best SR and SPL among the compared universal zero-shot navigation methods. Further evaluation on the open-vocabulary GOAT-Bench (Khanna et al., 2024) Val-Unseen split shows improvements over all compared training-free baselines. Ablation studies support the complementarity of direct and contextual repair, as well as the contributions of belief-aware ranking, visibility-weighted evidence updates, and information gain.

Our contributions are summarized as follows: (**i**) We introduce a task-conditioned evidence memory that associates spatially grounded goal hypotheses with direct, contextual, and negative evidence, using visibility-aware updates to maintain evolving belief and uncertainty. (**ii**) We develop a unified memory-repair policy that jointly ranks unexplored frontiers and memory candidates using belief, expected information gain, and execution costs, integrating direct and contextual repair within one decision process. (**iii**) Experiments on HM3D, MP3D, and GOAT-Bench demonstrate strong navigation performance without task-specific training, with ablations supporting the effectiveness of the proposed evidence updating and repair mechanisms.

2 Related Work

2.1 Zero-shot Navigation

Conventional navigation methods (Kwon et al., 2023; Ramakrishnan et al., 2022; Chaplot et al., 2020) require extensive training in simulated environments, which limits their generalization ability. Recently, with the emergence of LLMs (Liu et al., 2023a; Achiam et al., 2023; Bai et al., 2025b) demonstrating strong understanding and reasoning capabilities, some works (Chen et al.,

2023; Kuang et al., 2024; Zhang et al., 2024) have adopted zero-shot navigation approaches, achieving competitive performance without additional training. According to the goal type, zero-shot navigation can be broadly categorized into three classes: ON (Cai et al., 2024; Zhang et al., 2025a; Huang et al., 2026), IIN (Krantz et al., 2023) and TN (Long et al., 2024a). VLFM (Yokoyama et al., 2024) use frontier-based exploration (FBE) and use LLM for goal location reasoning. TopV-Nav (Zhong et al., 2024) directly reasons on the top-view map with sufficient spatial information. CogNav (Cao et al., 2025) uses cognitive process modeling with LLMs to search for the goal object. WMNav (Nie et al., 2025) integrates a world model with VLM to address ON tasks. For more challenging instance-level navigation such as IIN and TN, Mod-IIN (Krantz et al., 2023) leverages FBE with keypoint matching, while PSL (Sun et al., 2024) and GOAT (Chang et al., 2023) rely on additional task-specific training. UniGoal (Yin et al., 2025) and InstructNav (Long et al., 2024a) introduce goal-oriented and language-conditioned navigation tasks (Achiam et al., 2023; Long et al., 2024b; Chen et al., 2024), respectively. Building on these works, we incorporate spatial cognition to enable the agent to leverage memory of previously overlooked potential targets, which can benefit both language and vision-conditioned goal-oriented navigation tasks.

2.2 Scene exploration strategy

To enable more effective exploration and comprehensive scene understanding, various scene representation methods have been proposed, which can be broadly categorized into graph-based (Liu et al., 2023b; Rajvanshi et al., 2024; Wu et al., 2024) and map-based methods (Yokoyama et al., 2024; Zhang et al., 2025b; Yu et al., 2023). Graph-based representations explicitly model objects and their relationships, enabling structured reasoning and efficient relational inference for navigation. VoroNav (Wu et al., 2024) introduces the Reduced Voronoi Graph to extract exploratory paths and planning nodes for efficient navigation. UniGoal (Yin et al., 2025) is a representative graph-based approach, modeling both the goal graph and the scene graph with LLM, and exploring the environment via graph matching. TopoNav (Liu et al., 2025) leverages topological structures as a form of spatial memory that captures scene connectivity, adjacency, and semantic information. Map-based representations provide a dense spatial layout of the environment, allowing the agent to plan paths, track explored regions. L3mvn (Yu et al., 2023) first extracts the frontiers and computes their cost scores, which are then combined with LLM-based evaluations to select the optimal exploration direction. TriHelper (Zhang et al., 2024) extends L3mvn's framework with three distinct helper designs, enhancing the collision exploration and detection modules. ApexNav (Zhang et al., 2025b) builds on VLFM (Yokoyama et al., 2024) by incorporating a semantic scoring graph and a dynamic exploration strategy, achieving efficient and reliable scene coverage. Unlike these methods, we dynamically construct a sparse cognitive map during exploration, leveraging goal relocation and context-guided reasoning to assist target search.

3 Method

3.1 Problem formulation

We consider goal-oriented navigation in an unknown indoor environment. An episode is specified by a goal G and an initial agent pose. Depending on the task, G is an object category for object-goal navigation (ON), a reference image for instance-image navigation (IIN), or a natural-language description for text navigation (TN). The agent receives egocentric RGB-D observations and odometry, executes primitive controls $u_t \in \{\texttt{move_forward}, \texttt{turn_left}, \texttt{turn_right}, \texttt{stop}\}$, and must reach the target within a fixed action budget.

The key difficulty is that an unsuccessful observation is ambiguous. A target-like object may be visible but rejected because of viewpoint or appearance differences; the target may be hidden even though nearby context supports its presence; or the current view may not cover the relevant region well enough for a negative conclusion. We therefore formulate navigation as sequential evidence management. At time t, the agent maintains frontier candidates and task-relevant memory candidates. Its high-level decision h_t selects where to gather or verify evidence, while the navigation backend converts that decision into primitive controls.

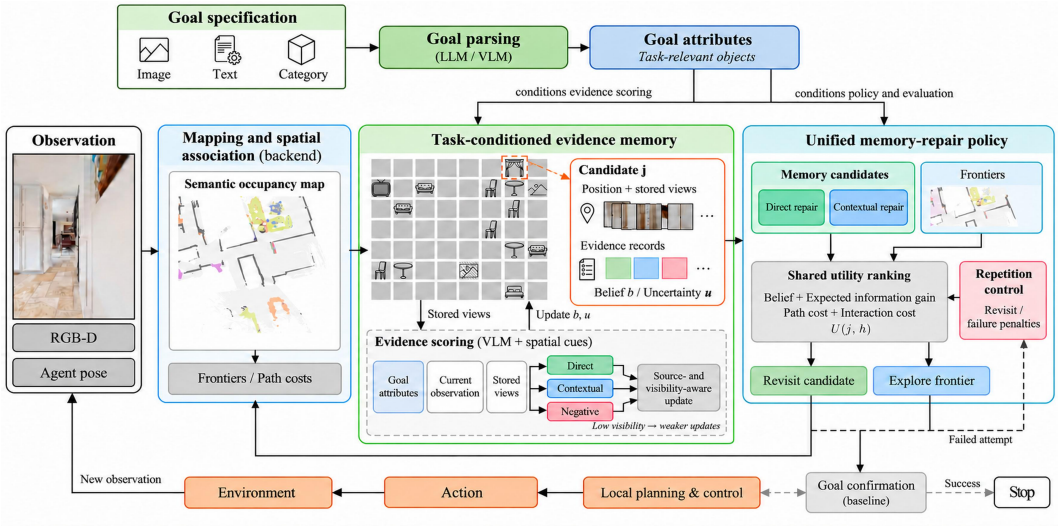


Figure 2: Overview of EGNav. Source- and visibility-aware updates maintain a task-conditioned evidence memory. A unified policy selects between memory revisitationand frontier exploration using belief, expected information gain, and execution costs. Execution produces new observations for subsequent memory updates.

3.2 Overview

EGNav leaves the occupancy map and local navigation backend unchanged and adds an evidence-management layer. The backend supplies collision-free paths and unexplored frontiers. The evidence layer parses the goal, associates observations with spatially grounded candidate locations, updates a task-conditioned memory, and scores whether a frontier or a previously visited location is the most useful next target.

The memory contains two complementary forms of support. A candidate with direct target evidence is evaluated using instance and appearance consistency, whereas a candidate with contextual evidence is evaluated using the likelihood that the target is nearby but outside the current view. Both modes are passed to the same repair policy. The policy can select frontier exploration, memory repair, controlled verification, or termination. A failed attempt changes the evidence state and revisit cost; it is not treated as an immediate proof that the candidate is absent.

3.3 Task-conditioned evidence memory

Task-conditioned hypotheses. The memory is a sparse spatial index over hypotheses that may help solve the current goal; it is not intended to reconstruct the complete environment. We write the goal representation inline as $G = \{c_G, a_G^{\text{int}}, a_G^{\text{ext}}\}$, where c_G is the goal category, a_G^{int} describes intrinsic appearance, and a_G^{ext} describes surrounding context. For ON, only c_G may be available. For IIN and TN, the VLM extracts the category together with intrinsic and extrinsic attributes from the reference image or description.

The detector vocabulary is task conditioned. We initialize a small set of common indoor landmarks and goal categories, and use the LLM or VLM to add categories likely to be informative for the current goal. Let $\mathcal{C}_{\text{cog}}$ denote the resulting vocabulary. Grounding-DINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) detect and segment instances from this vocabulary. RGB-D projection and odometry associate detections across views and provide their spatial support.

Memory state. Each memory item is a candidate hypothesis with state

$$m_j^t = \left(\mathbf{p}_j, c_j, \mathcal{I}_j, \alpha_j^t, \beta_j^t, q_{j,\text{vis}}^t, E_j^+, E_j^-, n_j^t, z_j\right), \tag{1}$$

where $\mathbf{p}_j$ is its spatial location, c_j is its category, and $\mathcal{I}_j$ is the ranked set of observations associated with the candidate. The pair (α_j^t, β_j^t) stores accumulated support and counter-evidence for the hypothesis that the goal is reachable near $\mathbf{p}_j$. The visibility score $q_{j,\text{vis}}^t$ describes how well the available views cover the relevant region. E_j^+ and E_j^- store the positive and negative evidence records,

n_j^t counts revisits, and $z_j \in \{\text{direct}, \text{context}\}$ records only how the candidate was initialized. In particular, z_j is not changed when a later observation provides another evidence type.

The spatial component of the memory is implemented with the same multi-frame fusion used by the original navigation backend. For a segmented detection, we project the mask into a 3D point cloud, apply voxel downsampling and DBSCAN filtering, and associate it with an existing item using point-cloud overlap. When an item is confirmed by repeated detections, we retain its position and the highest-quality full-frame observations. The fused object record is augmented with a task-conditioned evidence state rather than used only as a category-and-location list.

3.4 Evidence acquisition and belief update

Evidence channels and scores. At observation I_t, the VLM and the spatial backend produce three types of evidence. *Direct evidence* measures whether a target-like instance matches G under intrinsic and extrinsic attributes. *Contextual evidence* measures whether objects and room context around a memory item support the presence of the target outside the current field of view. *Negative evidence* is produced when a sufficiently visible view fails to support the target. The three channels are kept separate so that a weak contextual cue is not treated as equivalent to a confirmed instance match.

For reproducibility, the VLM is prompted to return normalized scores in $[0, 1]$: an intrinsic match score v_{int}, an extrinsic match score v_{ext}, and a context-support score v_{ctx}. For ON, $s_{j,t}^{\text{dir}}$ is given by category consistency together with the detector confidence; the intrinsic–extrinsic score is used for IIN and TN. For IIN and TN, we use $s_{j,t}^{\text{dir}} = (v_{\text{int}} + v_{\text{ext}})/2$ for a visible candidate, while $s_{j,t}^{\text{ctx}} = v_{\text{ctx}}$ is used for a contextual candidate. We set $s_{j,t}^{\text{match}} = s_{j,t}^{\text{dir}}$ when testing a visible target. These are confidence proxies used for fractional updates, rather than calibrated probabilities. The current evidence source is recorded separately as $r_{j,t} \in \{\text{direct}, \text{context}, \text{negative}\}$; a candidate initialized from context can therefore receive direct evidence after a later revisit.

Visibility-aware evidence. A missed detection is informative only when the current view provides adequate coverage. We compute a deterministic visibility proxy from normalized coverage, occlusion, image quality, and distance factors: $q_{j,\text{vis}}^t = \text{clip}(q_{\text{cov}}^t(1 - q_{\text{occ}}^t)q_{\text{img}}^t q_{\text{dist}}^t, 0, 1)$. A narrow or heavily occluded view consequently produces a small negative update, while a close, clear view produces a stronger one. The same score is used in the positive update to prevent low-quality observations from overwhelming the memory.

Belief and uncertainty. For each item, we maintain a Beta-shaped evidence accumulator

$$q_j^t \sim \text{Beta}(\alpha_j^t, \beta_j^t), \qquad b_j^t = \mathbb{E}[q_j^t] = \frac{\alpha_j^t}{\alpha_j^t + \beta_j^t}, \tag{2}$$

and use its variance as an uncertainty measure:

$$u_j^t = \frac{\alpha_j^t \beta_j^t}{(\alpha_j^t + \beta_j^t)^2(\alpha_j^t + \beta_j^t + 1)}. \tag{3}$$

This Beta parameterization is a lightweight, fractional evidence accumulator; it is not presented as a calibrated posterior over the physical world. We initialize it with a weak prior, so a single noisy VLM response cannot dominate the memory. Given a new observation, the update is $\alpha_j^{t+1} = \alpha_j^t + w_{j,t}^+$ and $\beta_j^{t+1} = \beta_j^t + w_{j,t}^-$, where the weights are determined by the evidence source and visibility:

$$w_{j,t}^+ = q_{j,\text{vis}}^t \begin{cases} \gamma_{\text{d}} s_{j,t}^{\text{dir}}, & r_{j,t} = \text{direct}, \\ \gamma_{\text{c}} s_{j,t}^{\text{ctx}}, & r_{j,t} = \text{context}, \\ 0, & r_{j,t} = \text{negative}, \end{cases} \qquad \gamma_{\text{d}} > \gamma_{\text{c}}. \tag{4}$$

Negative evidence is visibility weighted as $w_{j,t}^- = q_{j,\text{vis}}^t(1 - s_{j,t}^{\text{match}})$, with $w_{j,t}^- = q_{j,\text{vis}}^t$ when a visible view contains no target-compatible instance. Thus, "not observed" and "observed to be absent" lead to different updates.

3.5 UNIFIED MEMORY REPAIR POLICY

Candidate construction and information gain. At each decision stage, the policy considers the union $\mathcal{H}_t = \mathcal{F}_t \cup \mathcal{M}_t$ of frontier candidates $\mathcal{F}_t$ and non-permanently-failed memory candidates $\mathcal{M}_t$. Direct and contextual memory items therefore enter the same ranking process. Their source type changes the evidence likelihood, but not the policy that decides whether a revisit is worthwhile.

For a proposed high-level decision h, let $s_{j,t}(h)$ denote its expected target-relevance score: it is the direct or contextual VLM score for a memory item and the backend/VLM target-relevance prior for a frontier. We set $\pi_j^+(h) = q_{j,\mathrm{vis}}^t s_{j,t}(h)$ and $\pi_j^-(h) = q_{j,\mathrm{vis}}^t(1 - s_{j,t}(h))$. The remaining probability mass corresponds to an uninformative view. The expected uncertainty reduction is then

$$\begin{aligned}\mathrm{IG}(j,h) = u_j^t - \big[&\pi_j^+(h) u(\alpha_j^t + w_j^+, \beta_j^t) + \pi_j^-(h) u(\alpha_j^t, \beta_j^t + w_j^-) \\ &+ (1 - \pi_j^+(h) - \pi_j^-(h)) u_j^t\big],\end{aligned} \tag{5}$$

where $u(\alpha, \beta)$ is the Beta variance in Eq. 3. For this expectation, w_j^+ and w_j^- denote the same source- and visibility-weighted updates evaluated with the action-conditioned score $s_{j,t}(h)$. For a frontier, we use the backend's novelty estimate as $q_{j,\mathrm{vis}}^t$, its target-relevance prior as $s_{j,t}(h)$, and the neutral initialization $b_j^t = b_0$, $u_j^t = u_0$, $n_j^t = 0$, and $\Phi_j^t = 0$. This makes the same information-gain definition valid for both candidate types.

Action utility. We assign every candidate and high-level decision a utility

$$U(j,h) = \lambda_b b_j^t + \lambda_i \,\mathrm{IG}(j,h) - \lambda_d \widetilde{d}(\mathbf{p}_t, \mathbf{p}_j) - \lambda_v C_{\mathrm{VLM}}(h) - \lambda_r \log(1 + n_j^t) - \lambda_f \Phi_j^t, \tag{6}$$

where $\widetilde{d}$ is the normalized path cost, C_{VLM} is the expected interaction cost, and Φ_j^t is a failure penalty accumulated from unsuccessful verification. The coefficients are fixed on the validation split. The next high-level decision is

$$(j^*, h^*) = \arg \max_{j \in \mathcal{H}_t,\, h \in \mathcal{A}(j)} U(j,h). \tag{7}$$

The decision set contains frontier exploration, memory repair, controlled verification, and stopping. A high-belief but uncertain candidate favors verification; a contextual candidate with high expected information gain favors repair; a repeatedly visited low-value candidate is suppressed; and a novel frontier is selected when no memory item justifies its revisit cost.

Direct and contextual repair as one policy. Direct-evidence repair and context-evidence repair differ only in the query used to estimate the next evidence likelihood. For a direct candidate, the VLM compares the stored view with G and emphasizes instance and appearance consistency. For a contextual candidate, it evaluates whether the stored surrounding objects and room layout make an unseen target likely. Both queries update the same state in Eq. 1 and compete under Eq. 6. They are therefore two memory-repair modes under one policy, rather than separate navigation algorithms. Both modes share repetition-aware control: each repair attempt increments the revisit count n_j^t, while repeated verification failures increase the failure penalty Φ_j^t in Eq. 6. Repeatedly unsuccessful candidates may be temporarily suppressed and reconsidered when substantially different observations become available. A single low-visibility failure never causes permanent exclusion.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets: We evaluate EGNav in Habitat (Szot et al., 2021) on HM3D (Ramakrishnan et al., 2021) and MP3D (Chang et al., 2017). For IIN and TN, HM3D-Semantics-v0.2 contains 1,000 episodes across 36 scenes and six goal categories. For ON, HM3D-Semantics-v0.1 contains 2,000 episodes across 20 scenes and six goal categories. The MP3D ON evaluation contains 2,195 episodes across 11 scenes and 21 goal categories. We additionally evaluate on the open-vocabulary GOAT-Bench (Khanna et al., 2024) Val-Unseen split, which contains multimodal object-goal episodes requiring an agent to locate previously unseen objects from language and visual descriptions.

Table 1: Results of IIN, TN and ON on HM3D and MP3D.

Method	Zero-shot	Universal	InsINav		TextNav		ObjNav			
			HM3D		HM3D		MP3D		HM3D	
			SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
ZSON (Majumdar et al., 2022)	×	×	–	–	–	–	15.3	4.8	25.5	12.6
Cow (Gadre et al., 2023)	×	×	–	–	–	–	–	–	32.0	18.1
Krantz et al. (Krantz et al., 2022)	×	×	8.3	3.5	–	–	–	–	–	–
OVRL-v2-IIN (Yadav et al., 2023)	×	×	24.8	11.8	–	–	–	–	–	–
IEVE (Lei et al., 2024)	×	×	70.2	25.2	–	–	–	–	–	–
PSL (Sun et al., 2024)	×	✓	23.0	11.4	16.5	7.5	–	–	42.4	19.2
GOAT (Chang et al., 2023)	×	✓	37.4	16.1	17.0	8.8	–	–	50.6	24.1
ESC (Zhou et al., 2023)	✓	×	–	–	–	–	28.7	14.2	39.2	22.3
OpenFMNav (Kuang et al., 2024)	✓	×	–	–	–	–	37.2	15.7	52.5	24.1
L3MVN (Yu et al., 2023)	✓	×	–	–	–	–	–	–	50.4	23.1
VLFM (Yokoyama et al., 2024)	✓	×	–	–	–	–	36.2	15.9	52.4	30.3
SG-Nav (Yin et al., 2024)	✓	×	–	–	–	–	40.2	15.0	54.0	24.9
$\text{WMNav}_{\text{Gemini}}$ (Nie et al., 2025)	✓	×	–	–	–	–	**45.4**	17.2	58.1	31.2
Mod-IIN (Krantz et al., 2023)	✓	×	56.1	23.3	–	–	–	–	–	–
ApexNav (Zhang et al., 2025b)	✓	×	–	–	–	–	39.2	17.8	59.6	**33.0**
UniGoal (Yin et al., 2025)	✓	✓	60.2	23.7	20.2	11.4	41.0	16.4	54.5	25.1
EGNav (Ours)	✓	✓	**62.6**	**25.1**	**23.3**	**12.9**	43.1	**18.2**	**60.0**	32.9

Table 2: Results on the open-vocabulary GOAT-Bench Val-Unseen split.

Method	Training-free	SR↑	SPL↑
MTU3D (Zhu et al., 2025)	×	47.2	27.7
SenseAct-NN Skill Chain (Khanna et al., 2024)	×	29.5	11.3
VLMnav (Goetting et al., 2024)	✓	20.1	9.6
3D-Mem (Yang et al., 2025)	✓	28.8	15.8
TANGO (Ziliotto et al., 2025)	✓	32.1	16.5
EGNav (Ours)	✓	**39.1**	**23.7**

Evaluation metrics: We report Success Rate (SR) and Success weighted by Path Length (SPL). SR is the proportion of successful episodes. SPL weights each successful episode by the ratio of the shortest-path distance to the executed path length and is zero for failed episodes.

Implementation details. The agent moves $0.25\,\text{m}$ per step and rotates in increments of 30°. An onboard camera mounted at a height of $0.88\,\text{m}$ provides 640×480 RGB-D observations. For the IIN and TN tasks, the perception range is set to $[0, 30]\,\text{m}$, each episode is capped at 1000 steps, and the success threshold is set to $1.0\,\text{m}$. For the ON task, the perception range is restricted to $[0, 5]\,\text{m}$, each episode is capped at 500 steps, and the success threshold is set to $0.2\,\text{m}$. For the semantic occupancy map, we use a map size of 720×720 with a resolution of $0.05\,\text{m}$ per cell, covering an area of $36\,\text{m} \times 36\,\text{m}$. We initialize the evidence accumulators with $\alpha_0 = \beta_0 = 1$ and set the direct and contextual evidence weights to $\gamma_d = 1.0$ and $\gamma_c = 0.5$, respectively. The utility coefficients are set to $(\lambda_b, \lambda_i, \lambda_d, \lambda_v, \lambda_r, \lambda_f) = (1.0, 10.0, 0.3, 0.05, 0.1, 0.2)$. We use information gain without additional normalization and scale the path cost as $\tilde{d} = d_{\text{path}}/(10\,\text{m})$. The interaction cost C_{VLM} is measured by the expected number of VLM calls, and the failure penalty Φ_j^t increases by one for each recorded verification failure. Qwen3-VL-8B (Bai et al., 2025a) serves as both the LLM and VLM. All experiments are conducted using two NVIDIA GeForce RTX 4090 GPUs.

4.2 Main Results

As shown in Table 1, EGNav achieves the best SR and SPL among the compared universal navigation methods across all evaluated settings.

Instance-level navigation. On HM3D, EGNav achieves 62.6% SR and 25.1% SPL for IIN, and 23.3% SR and 12.9% SPL for TN, outperforming all compared zero-shot methods on both tasks. Relative to UniGoal, SR improves by 2.4 and 3.1 percentage points, respectively, alongside higher SPL, demonstrating consistent gains across image- and text-specified goals.

Table 3: Ablation of the two memory-repair modes on HM3D IIN. Direct and Context indicate the enabled repair modes. All variants share the baseline execution components.

Variant	Direct	Context	SR↑	SPL↑
Base navigator	–	–	58.8	22.5
Direct repair only	✓	–	61.4	23.6
Context repair only	–	✓	61.1	23.4
Direct + context repair	✓	✓	**62.6**	**25.1**

Table 4: Component ablations on HM3D IIN. The belief-term ablation removes its contribution to candidate ranking while retaining belief and uncertainty updates.

Variant	SR↑	SPL↑
w/o belief term ($\lambda_b = 0$)	61.2	23.7
w/o visibility weighting	61.5	23.8
w/o information gain ($\lambda_i = 0$)	61.1	23.2
EGNav (full)	**62.6**	**25.1**

Table 5: Recovery analysis of the two memory-repair modes on IIN. A trigger is a memory candidate selected for repair; a recovery is counted when the target is confirmed after the revisit.

Repair mode	Triggers	Recoveries	Rate (%)
Direct evidence	3.26	1.37	42.0
Context evidence	2.49	1.44	57.8

Object-goal navigation. On HM3D, EGNav achieves the highest SR of 60.0% among the compared methods, with SPL comparable to ApexNav (32.9% versus 33.0%). On MP3D, it achieves 43.1% SR and the highest SPL of 18.2%, improving upon ApexNav by 3.9 and 0.4 percentage points, respectively. Although WMNav attains higher SR on MP3D, EGNav achieves better SPL while supporting all three goal modalities.

Open-vocabulary navigation. As shown in Table 2, EGNav achieves 39.1% SR and 23.7% SPL on GOAT-Bench Val-Unseen, outperforming all compared training-free methods. Relative to TANGO, the strongest training-free baseline, our method improves SR and SPL by 7.0 and 7.2 percentage points, respectively. These results further support the effectiveness of our framework in open-vocabulary navigation without task-specific training.

4.3 Ablation Study

We conducted experiments on the HM3D dataset to validate the effectiveness of each proposed module and reported the final results on the IIN task.

Memory-repair modes. Table 3 evaluates the contributions of direct and contextual repair. Enabling direct repair improves SR and SPL over the baseline by 2.6 and 1.1 percentage points, respectively, while contextual repair provides gains of 2.3 and 0.9 points. These improvements are consistent with their complementary roles: direct repair re-examines previously observed target-like instances that were not reliably matched, whereas contextual repair revisits locations where surrounding evidence suggests a potentially unseen target. Combining both modes achieves 62.6% SR and 25.1% SPL, exceeding the baseline by 3.8 and 2.6 percentage points. It also improves over the stronger single-mode variant, direct repair, by 1.2 points in SR and 1.5 points in SPL. These results support integrating both evidence sources within a shared repair policy, rather than relying exclusively on observed target candidates or contextual cues.

Evidence updating and candidate ranking. Table 4 examines the components used to update evidence and prioritize navigation candidates. Removing the belief term ($\lambda_b = 0$) reduces SR and SPL by 1.4 percentage points each. Since belief and uncertainty updates remain active in this variant, the comparison specifically supports explicitly incorporating accumulated target support into candidate ranking, rather than the evidence-state representation as a whole. Removing visibility weighting decreases SR and SPL by 1.1 and 1.3 percentage points, respectively. This result is consistent with the motivation to account for view coverage and observation quality when accumulating positive and negative evidence, so that poorly observed regions do not receive the same update weights as clearly observed ones. Finally, removing information gain ($\lambda_i = 0$) yields 61.1% SR and 23.2% SPL, corresponding to decreases of 1.5 and 1.9 percentage points. This suggests that considering expected uncertainty reduction complements current belief when selecting between exploration and revisitation. Overall, the full model outperforms all three variants on both metrics, supporting the combined

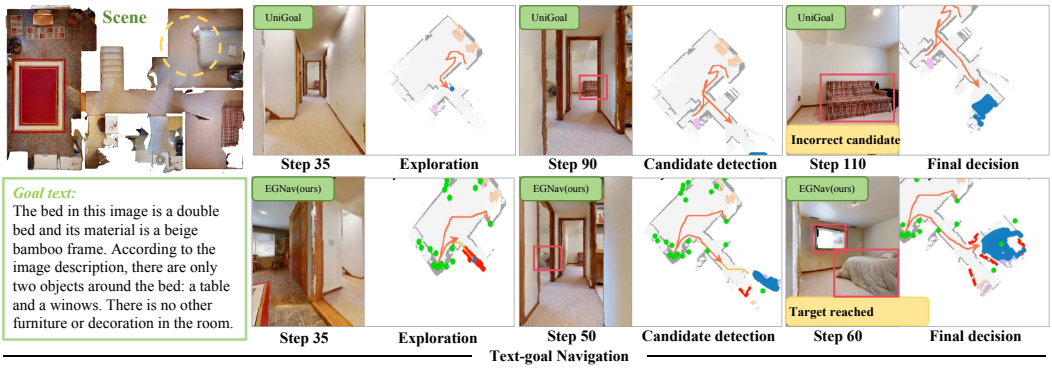


Figure 3: Qualitative comparison with UniGoal on a text-goal navigation episode. The left panel shows the scene, the highlighted target location, and an abbreviated goal description. UniGoal (top) terminates at an incorrect candidate, whereas EGNav (bottom) reaches the specified target.

Table 6: Sensitivity to the information-gain weight λ_i on HM3D IIN.

λ_i	SR↑	SPL↑
5.0	62.0	23.9
10.0	62.6	25.1
20.0	62.3	24.5

Table 7: Sensitivity to the contextual evidence weight γ_c on HM3D IIN.

γ_c	SR↑	SPL↑
0.25	62.1	24.7
0.50	62.6	25.1
0.75	61.9	24.7

use of visibility-aware evidence updates, belief-aware ranking, and information gain under shared baseline execution settings.

Repair-mode recovery analysis. Table 5 compares the two memory-repair modes. Direct-evidence repair is triggered more frequently (3.26 versus 2.49), while contextual repair has the higher recovery rate (57.8% versus 42.0%). This pattern is consistent with the role of contextual evidence: it is less frequent, but when selected it provides a more reliable cue for a previously unseen target.

Information-gain weight. Table 6 shows that the default $\lambda_i = 10$ achieves the best SR and SPL among the tested values. Both smaller and larger weights lead to modest performance decreases, with SPL varying more than SR. Overall, performance remains relatively stable across the tested range, suggesting limited sensitivity to the exact information-gain weight.

Contextual evidence weight. As shown in Table 7, the default $\gamma_c = 0.5$ achieves the best results with γ_d fixed. Both weaker and stronger contextual updates produce only small changes in SR and SPL. These results support the default setting while indicating limited sensitivity to the relative strength of contextual evidence within the tested range.

4.4 Qualitative Results

Figure 3 compares EGNav with UniGoal on a TN episode specifying both object appearance and surrounding context. UniGoal stops at an incorrect candidate, while EGNav reaches the specified bed. This example illustrates the full system's instance-level goal selection, complementing the quantitative results.

5 Conclusion

We presented EGNav, a zero-shot navigation framework that treats goal search as sequential evidence management. A task-conditioned evidence memory maintains belief and uncertainty through source- and visibility-aware updates, while a unified policy selects between frontier exploration and memory revisitation. Direct and contextual repair operate within the same decision process. Experiments on HM3D, MP3D, and GOAT-Bench demonstrate strong navigation performance, with ablations supporting the complementarity of the repair modes and the contributions of evidence updating and candidate ranking.

AI USE STATEMENT

In this work, we used generative AI tools to assist with English language polishing, sentence rewriting, and code generation. The research ideas, methodology, and experimental design were developed by the authors. Generative AI was not used to generate experimental results; all experiments were conducted by the authors. AI-assisted code was manually reviewed, verified, and tested before use. All AI-assisted manuscript content was also carefully reviewed and revised by the authors to ensure technical correctness and consistency with the actual methodology and experimental results. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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 *ICRA*, 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 *ICCV*, 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.

Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*, 2023.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *NeurIPS*, 2020.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *ACL*, 2024.

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. *RSS*, 2023.

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 *CVPR*, 2023.

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

Xun Huang, Shijia Zhao, Yunxiang Wang, Xin Lu, Wanfa Zhang, Rongsheng Qu, Weixin Li, Yunhong Wang, and Chenglu Wen. Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation. In *CVPR*, 2026.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *CVPR*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-specific image goal navigation: Training embodied agents to find object instances. *arXiv preprint arXiv:2211.15876*, 2022.

Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to objects specified by images. In *ICCV*, 2023.

Yuxuan Kuang, Hai Lin, and Meng Jiang. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. *arXiv preprint arXiv:2402.10670*, 2024.

Obin Kwon, Jeongho Park, and Songhwai Oh. Renderable neural radiance map for visual navigation. In *CVPR*, 2023.

Xiaohan Lei, Min Wang, Wengang Zhou, Li Li, and Houqiang Li. Instance-aware exploration-verification-exploitation for instance imagegoal navigation. In *CVPR*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023a.

Peiran Liu, Qiang Zhang, Daojie Peng, Lingfeng Zhang, Yihao Qin, Hang Zhou, Jun Ma, Renjing Xu, and Yiding Ji. Toponav: Topological graphs as a key enabler for advanced object navigation. *arXiv preprint arXiv:2509.01364*, 2025.

Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023b.

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 *ECCV*, 2024.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *CoRL*, 2024a.

Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. In *ICRA*, 2024b.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *NeurIPS*, 2022.

AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 2024.

Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. *IROS*, 2025.

Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. In *Proceedings of the International Conference on Automated Planning and Scheduling*, 2024.

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 *CVPR*, 2022.

Xinyu Sun, Lizhao Liu, Hongyan Zhi, Ronghe Qiu, and Junwei Liang. Prioritized semantic learning for zero-shot instance navigation. In *ECCV*, 2024.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021.

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 *ICML*, 2024.

Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. *arXiv preprint arXiv:2303.07798*, 2023.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *CVPR*, 2025.

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. *NeurIPS*, 2024.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *CVPR*, 2025.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *ICRA*, 2024.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *IROS*, 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 *IROS*, 2024.

Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. In *ICRA*, 2025a.

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 RA-L*, 2025b.

Linqing Zhong, Chen Gao, Zihan Ding, Yue Liao, Huimin Ma, Shifeng Zhang, Xu Zhou, and Si Liu. Topv-nav: Unlocking the top-view spatial reasoning potential of mllm for zero-shot object navigation. *arXiv preprint arXiv:2411.16425*, 2024.

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 *ICML*, 2023.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *ICCV*, 2025.

Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. Tango: Training-free embodied ai agents for open-world tasks. In *CVPR*, 2025.

A Baseline and Execution Details

Navigation backend and baseline. We build upon ApexNav (Zhang et al., 2025b) for ON and UniGoal (Yin et al., 2025) for IIN and TN. The backend handles RGB-D integration, semantic mapping, frontier extraction, local path planning, and primitive control execution. Our ablation baseline retains the navigation backend and task-specific goal confirmation, without memory-guided repair. These execution components are shared across the repair-mode variants, while EGNav adds evidence memory and unified selection between memory candidates and frontiers.

Goal confirmation and repetition control. For ON, goal confirmation checks category agreement and spatial consistency. For IIN and TN, the VLM compares the current candidate observation with the goal's intrinsic and extrinsic attributes. Goal confirmation is a shared execution component rather than a separate repair policy. When memory repair is enabled, each attempt increments n_j^t, and each recorded verification failure increases Φ_j^t by one. Repeatedly unsuccessful candidates may be temporarily suppressed and reconsidered when substantially different observations become available. A single low-visibility failure is insufficient for permanent exclusion.

B Prompt Templates

We organize the main model queries into goal parsing, direct evidence scoring, contextual evidence scoring, and goal confirmation. Direct scoring compares visible candidates with the goal, whereas contextual scoring evaluates support for a target that may lie outside the observed view. The returned scores are confidence proxies for evidence updates, rather than calibrated probabilities. Belief, uncertainty, and candidate utilities are computed by the navigation system, not generated directly by the VLM.

Goal parsing. The goal query extracts the target attributes and identifies object categories that may provide useful search context.

```
Task: Parse a goal for indoor navigation.

Input:
- Task type: {ON / IIN / TN}
- Goal: {category / reference image / text description}

Identify:
1. The target category.
2. Intrinsic attributes, such as appearance, material,
   shape, and size, when available.
3. Extrinsic attributes describing nearby objects and
   the surrounding context, when available.
4. Object categories that may help locate the target.

Separate attributes supported by the goal specification
from inferred contextual associations. Do not describe
inferred associations as observed scene facts.

Return the target category, intrinsic attributes,
extrinsic attributes, and informative object categories.
```

Direct evidence scoring. For IIN and TN, the query evaluates a visible candidate against the goal's intrinsic and extrinsic attributes. The returned scores are combined as $s_{j,t}^{\mathrm{dir}} = (v_{\mathrm{int}} + v_{\mathrm{ext}})/2$. For ON, direct evidence uses category consistency and detector confidence, as described in the main text.

```
Task: Evaluate a visible candidate against the goal.

Input:
- Goal category and attributes: {goal}
- Candidate to evaluate: {candidate}
```

```
- Supplied observation(s): {images}

Compare the candidate's appearance with the intrinsic
goal attributes. Compare its surroundings with the
extrinsic goal attributes.

Distinguish category agreement from instance agreement.
Base the assessment on the supplied observations.

Return:
- v_int: intrinsic match score in [0, 1]
- v_ext: extrinsic match score in [0, 1]
```

Contextual evidence scoring. The contextual query evaluates whether the surroundings of a memory candidate support revisiting the location to search for an unseen target. Its output is used as $s_{j,t}^{\mathrm{ctx}} = v_{\mathrm{ctx}}$.

```
Task: Evaluate contextual support for revisiting a location.

Input:
- Goal category and attributes: {goal}
- Memory candidate: {candidate}
- Stored surrounding observations: {images}

Evaluate whether the surrounding objects and room layout
support the possibility that the target is nearby but
outside the observed view.

Do not treat contextual support as a direct target match.
A target outside the view is not necessarily absent
from the surrounding region.

Return:
- v_ctx: contextual support score in [0, 1]
```

Goal confirmation. For IIN and TN, final confirmation compares the current candidate observation with the goal specification. Spatial acceptance conditions and any memory-dependent uncertainty condition are checked separately by the system.

```
Task: Confirm whether the current candidate is the target.

Input:
- Goal specification and attributes: {goal}
- Current candidate observation(s): {images}

Compare the visible candidate with the target's intrinsic
attributes and the specified surrounding context.

Determine whether the available visual evidence supports
an instance-level match. Category agreement alone is
insufficient for an instance-specific goal.

Return the final instance-level match decision.
```