KNOWING WHERE TO LOOK NEXT: ADAPTIVE REGIONAL INDEXING AND EVIDENCE-GUIDED SEARCH FOR MULTIMODAL LIFELONG NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal lifelong navigation requires robots to find successive targets specified by categories, language descriptions, or reference images in the same initially unknown environment while reusing past observations. Three bottlenecks limit effective search. First, object-indexed retrieval relies on detections made before future goals are known, carrying early omissions and mislabels into later attempts to retrieve target evidence and search cues. Second, semantic scoring ranks frontier candidates generated by geometry-only clustering of explored–unexplored boundaries, but cannot add cue-indicated observation points absent from that set. Third, scheduling must prevent new nearby geometric frontiers from repeatedly postponing pending inspections and identify worthwhile revisits when candidates are exhausted but the target remains unfound. We present TreeNav, a framework without task-specific model training that uses visual evidence not only to recognize targets, but also to decide where to search and where to look again. Its Scene Tree organizes observations into a content-adaptive regional hierarchy and retrieves raw images through regional context for goal-specific target verification and cue extraction. Regional budgets limit redundant checks while preserving verification opportunities across relevant regions. Grounded Semantic Frontiers (GSFs) turn identified search cues—goal-related objects and relevant regional information in historical and current images—into reachable observation candidates. These candidates extend historical reuse beyond direct target sightings and add key inspection locations missed by geometric frontiers during online exploration. Evidence-first wave scheduling balances discovery benefits against navigation costs, limits repeated displacement of pending inspections by new geometric frontiers, and promptly incorporates new GSFs to respond to new visual cues. After all executable candidates are exhausted, targeted recovery reuses the Scene Tree's regional index to select goal-relevant, insufficiently observed mapped locations for reinspection. In system-level comparisons on full GOAT-Bench val-unseen, TreeNav achieves 73.14% success rate and 50.14% success weighted by path length (SPL), exceeding the best reported result for each metric under the same protocol by 6.62 and 5.44 percentage points, respectively.

1 INTRODUCTION

Robots operating over long periods need to find different targets in the same environment without restarting exploration for every request. Multimodal lifelong navigation captures this requirement: goals arrive sequentially as object categories, language descriptions, or reference images, and the robot retains observations across tasks (Chang et al., 2024; Khanna et al., 2024). Efficient search requires retrieving useful historical evidence, using current visual information to select exploration locations, and allocating a limited navigation budget to inspections and revisits.

Three bottlenecks limit these capabilities.

(1) Historical retrieval depends on early recognition and underuses visual context. 3D-Mem and MSGNav retrieve historical images through object indices built before later goals are known (Yang et al., 2025; Huang et al., 2026). When early omissions or mislabels block this retrieval route,

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.

the images cannot be reexamined for the new goal. Even when images are retrieved, checking only for target presence leaves search cues from related objects and spatial layouts unused. Room-label grouping provides a regional entry point, but visually distinct areas sharing one label remain in a common retrieval group (Ge et al., 2026). Retrieval therefore needs to organize evidence by visual context and cover distinct goal-relevant regions within a limited verification budget.

(2) Visual cues are confined to frontier scoring rather than observation-location generation. Methods such as VLFM and SG-Nav use semantic information to rank geometric frontiers (Yokoyama et al., 2024; Yin et al., 2024). Candidate locations are determined by geometric clustering of explored–unexplored boundaries, not by objects or regional entrances in the images that warrant further inspection. When a cue-indicated observation point is absent from the resulting set, even accurate scores can only choose among the existing geometric candidates. Better frontier ranking is therefore insufficient: visual cues must also participate in generating observation locations, making key inspection points omitted by boundary clustering available for search.

(3) New geometric frontiers preempt pending inspections, while observation gaps in goal-relevant regions lack targeted reinspection. Prior work combines target relevance and travel cost to plan visits and reduce backtracking (Liu et al., 2025; Niu et al., 2026). In online search, however, newly generated nearby geometric frontiers can repeatedly rank ahead of pending cue-guided inspections. Candidate exhaustion raises a further challenge: where to inspect again. Visiting an area does not establish adequate observation, particularly under occlusion or from distant viewpoints. Further search must target observation gaps in goal-relevant mapped regions, rather than simply repeat earlier routes.

We present **TreeNav**, a visual evidence-guided navigation framework that requires no task-specific model training (Figure 1). Our guiding principle is to organize lifelong navigation as a continuous cycle of using and acquiring goal-relevant visual evidence: visual evidence should drive target search, not just target recognition.

Scene Tree uses adaptive regions to organize evidence retrieval and verification budgets. Scene Tree groups incoming observations by visual context and creates subregions for visually distinct local spaces, forming an adaptive hierarchy that matches the observed environment. For each goal, it allocates verification budgets across relevant regions and selects images that cover different locations, limiting repeated checks of similar views. It then reexamines these raw images to recognize the target and extract search cues, without requiring correct target identification at storage time.

GSFs turn visual search cues into observation locations, not merely frontier scores. We define a Grounded Semantic Frontier (GSF) as a reachable candidate observation location guided by goal-related visual evidence. These cues are visible objects or spatial structures that indicate where to search next. For example, when searching for robes, a clothes rail provides a reason to inspect the surrounding area without requiring the robes to appear in the image. TreeNav constructs GSFs from identified cues in historical and current images, adding inspection locations omitted by geometric frontiers. Cue-bearing historical images thus guide search without directly showing the target, while current visual understanding influences where exploration candidates are generated rather than only how existing frontiers are ranked.

Evidence-first search limits geometric-frontier preemption and selects revisits by relevance and observation insufficiency. Evidence-first wave scheduling balances discovery benefits against navigation costs, protects pending inspections from repeated geometric-frontier preemption, and promptly admits new GSFs to respond to visual cues. After all executable candidates are exhausted, targeted recovery reuses the Scene Tree's regional index, combining regional relevance with observation insufficiency. The former identifies where the goal warrants further search; the latter identifies where additional observation is needed. Reinspection thus targets relevant locations that remain poorly observed, rather than simply returning to nearby historical locations.

In system-level comparisons on the full GOAT-Bench val-unseen protocol, TreeNav improves success rate (SR) and success weighted by path length (SPL) over the best previously reported result for each metric under the same protocol by **6.62 and 5.44 percentage points**, respectively. These comparisons involve different backbones; controlled ablations under matched model and execution settings further support the contributions of retrieval, GSFs, scheduling, and recovery.

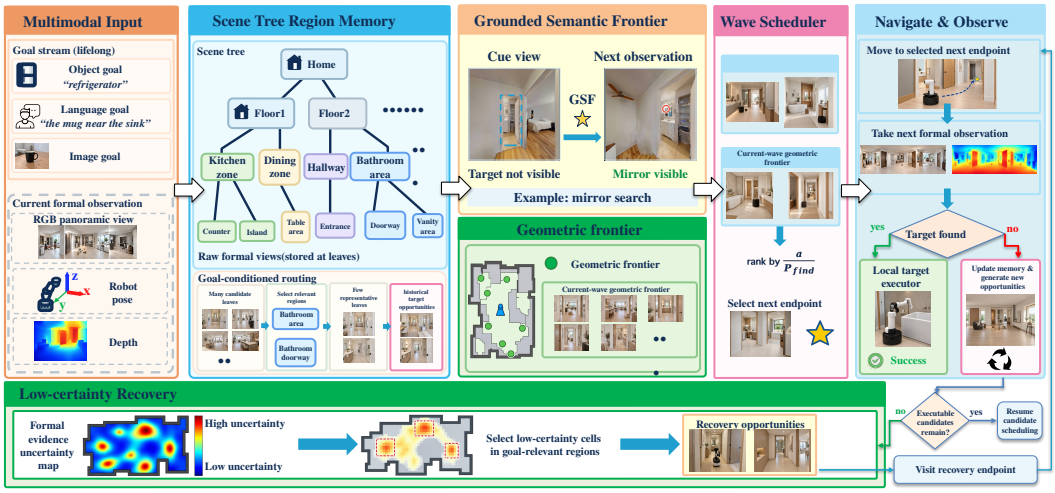

Figure 1: The TreeNav framework. The Scene Tree retrieves raw observations. Historical cues generate GSFs for seeking unseen targets; current cues generate GSFs for online exploration. Historical target candidates are verified first; GSFs and geometric frontiers in the current wave are then ranked by discovery benefits and navigation costs. Execution updates visual memory and the coverage map. Recovery inspects underobserved goal-relevant regions after candidates are exhausted.

2 METHOD

TreeNav connects three decisions: which historical images to reexamine (Section 2.2), where visual cues warrant new observations (Section 2.3), and which inspections or revisits to prioritize (Section 2.4). We first define the shared observations and coverage quantities.

2.1 Preliminaries

Task and observations. The agent receives sequential goals $g_1, \ldots, g_K$ in the same unknown environment, specified by object categories, language descriptions, or reference images (Khanna et al., 2024). It uses RGB-D observations and an incremental navigation map. Each search or verification location provides a formal four-view panorama $\mathcal{O}_t = (I_t, D_t, \mathbf{q}_t, \boldsymbol{\kappa}_t)$: images, aligned depth, acquisition pose, and camera parameters. The Scene Tree and observation map persist across subtasks.

Shared observation sufficiency. To distinguish mapped space from adequately observed space, let $C_t(c)$ measure observation sufficiency at cell c, initialized to one at the robot footprint and zero elsewhere. A formal panorama updates only cells $\mathcal{V}_t$ covered by valid depth:

$$C_t(c) = \max\left\{C_{t-1}(c), 1 - \frac{d(c, \mathbf{q}_t)}{d_{\max}}\right\}, \qquad c \in \mathcal{V}_t. \tag{1}$$

Here $d_{\max}$ is the valid depth range; uncovered cells retain their values. Closer views contribute more, and the maximum preserves the best coverage across observations. The complement $U_t(c) = 1 - C_t(c)$ measures the remaining insufficiency of geometric observation coverage.

For any observation subset $\mathcal{S}$, let $\gamma_s(c) = \max\{0, 1 - d(c, \mathbf{q}_s)/d_{\max}\}$ on cells covered by valid depth in observation s, and zero elsewhere. Its sufficiency at cell c is $C_{\mathcal{S}}(c) = \max_{s\in\mathcal{S}} \gamma_s(c)$; summing over cells gives the subset's total coverage, $\sum_c C_{\mathcal{S}}(c)$. The cellwise maximum retains the best observation at each location and prevents overlapping views from inflating coverage.

2.2 Scene Tree: Content-Adaptive Regional Indexing

Scene Tree retrieves raw observations through regional context, without requiring correct target recognition at storage time. It builds an adaptive regional hierarchy from visual content, then selects representative images to verify the target and extract search cues when a goal arrives. Regional budgets and complementary coverage limit redundant checks.

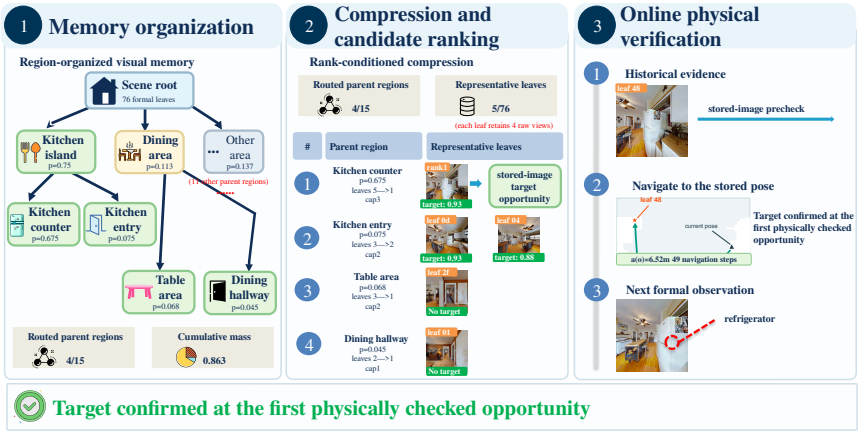

Figure 2: A Scene Tree retrieval example. Routing selects 5 representative leaves from 76 historical leaves; a new formal observation verifies the first historical target candidate visited. Node scores are regional routing probabilities.

Online regional organization. The variable-depth tree $\mathcal{T} = (\mathcal{V}_{\text{reg}} \cup \mathcal{V}_{\text{leaf}}, \mathcal{E})$ has a scene root and containment edges. Each regional node v stores a summary, spatial center, and child set $\mathcal{C}_v$; terminal regions contain leaves storing raw $\mathcal{O}_s$, deduplicated by acquisition location. A vision-language model (VLM) compares each new panorama with nearby candidate regions, joining a visually consistent region or creating a new region or subregion. As observations accumulate, tree depth follows visual differences rather than a fixed room hierarchy. This goal-independent construction gives distinct contexts separate indexing units, including within the same room. These units subsequently receive goal-dependent relevance scores and verification budgets. Growth operations and summary updates are detailed in Appendix C.

Goal-conditioned regional routing. Given goal g, the VLM scores sibling regions from their labels and summaries, assigning nonnegative $z_v(u; g)$ to child $u \in \mathcal{C}_v$. Normalization and path aggregation give

$$\pi(u \mid v, g) = \frac{z_v(u; g)}{\sum_{w \in \mathcal{C}_v} z_v(w; g)}, \qquad p_{\mathcal{T}}(u \mid g) = \prod_{(v \to w) \in \text{path}(u)} \pi(w \mid v, g). \tag{2}$$

Sort terminal regions $R_1, \ldots, R_n$ by descending $p_{\mathcal{T}}$. A cumulative-probability rule selects a goal-dependent prefix:

$$m_\rho = \min \left\{ m : \sum_{r=1}^{m} p_{\mathcal{T}}(R_r \mid g) \geq \rho \right\}. \tag{3}$$

Region-count bounds give the retained set $\widehat{\mathcal{R}}_g$. Concentrated routing retains fewer regions; dispersed routing broadens inspection. These scores guide verification toward goal-relevant regions.

Region-balanced original-image verification. Each retained nonempty region is a unit of verification allocation: at least one slot preserves access to its visual context, while a rank-dependent cap limits repeated checks within the region. For region R_r, spatially deduplicate its leaves into $\mathcal{L}_r$ and set $b_r = \min\{|\mathcal{L}_r|, \max(1, b_{\max} - \lfloor r/2 \rfloor)\}$. Using the cellwise sufficiency $C_{\mathcal{S}}(c)$ from Section 2.1, select the representative subset with the greatest total coverage:

$$\mathcal{S}_r^\star = \arg\max_{\substack{\mathcal{S} \subseteq \mathcal{L}_r \\ 1 \leq |\mathcal{S}| \leq b_r}} \sum_c C_{\mathcal{S}}(c), \qquad \mathcal{Q}_g = \bigcup_{R_r \in \widehat{\mathcal{R}}_g} \mathcal{S}_r^\star. \tag{4}$$

The VLM reexamines $\mathcal{Q}_g$ for two outputs: direct target matches create on-site verification candidates, while identified goal-related cues support GSF generation (Section 2.3), including when the target is absent. Figure 2 illustrates retrieval; Appendix D gives routing and budget settings.

2.3 GSFs: Observation Locations from Visual Search Cues

A Grounded Semantic Frontier (GSF) is a reachable candidate observation location linked to an

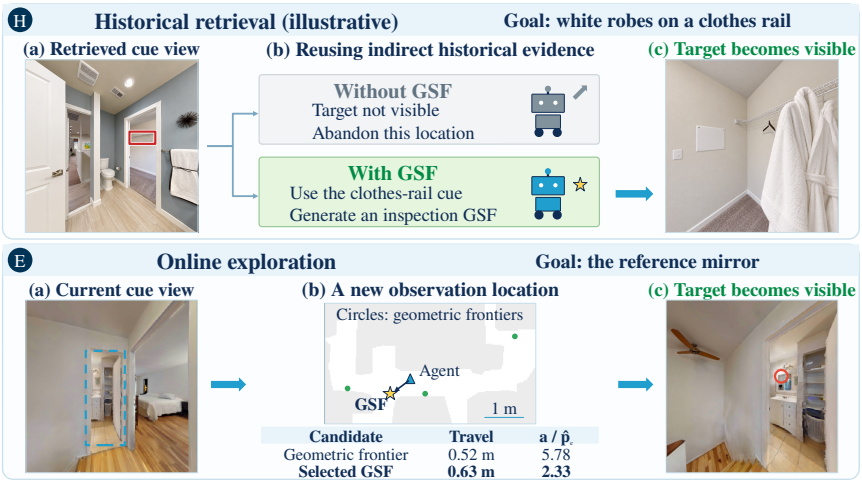

Figure 3: GSFs from historical and current cues. Top: the mechanism illustration contrasts abandoning this location after a negative target check with using the clothes-rail cue to generate a GSF and later find the robes. Bottom: current doorway-and-vanity cues guide a closer mirror observation; the map compares the GSF with geometric frontiers.

identified visual cue for where to search. Cues are goal-related visible objects or spatial structures: a clothes rail, for example, supports searching nearby for robes without showing the robes themselves (Figure 3). The design lets visual evidence determine inspection locations instead of only scoring geometric frontiers. Geometric frontiers remain available when the images provide no such cue.

Grounding visual cues. Historical retrieval and current perception supply the same cue-to-location mechanism. For a localizable cue c, the VLM identifies its source observation s, view j_c, and pixel $\mathbf{u}_c$. Source depth and acquisition geometry locate the cue:

$$\mathbf{z}_c = \mathbf{T}(\mathbf{q}_s, \boldsymbol{\kappa}_s^{(j_c)})\Pi^{-1}\left(\mathbf{u}_c, D_s^{(j_c)}(\mathbf{u}_c)\right). \tag{5}$$

Here Π^{-1} back-projects the depth pixel and $\mathbf{T}$ transforms source-camera coordinates to world coordinates. The navigation module selects reachable observation endpoints around $\mathbf{z}_c$. The grounded cue provides a spatial anchor for observation candidates, including locations outside the existing geometric-frontier set.

For a cue specifying only a direction or region, reachable frontiers aligned with its source view supply endpoints:

$$\mathcal{X}_t(c) = \{\mathbf{x}_f : f \in \mathcal{F}_t^{\text{reach}},\ \text{view}_s(\mathbf{x}_f) = j_c\}. \tag{6}$$

Point cues can add locations beyond the geometric set, while directional cues guide existing frontiers. Duplicate object or regional cues are merged. GSF generation details are given in Appendix J.

2.4 Evidence-First Scheduling and Targeted Recovery

We introduce evidence-first wave scheduling and targeted recovery: wave-based admission limits repeated displacement of pending inspections by new geometric frontiers while promptly incorporating new GSFs to respond to visual cues; within each wave, GSFs and geometric frontiers are ranked jointly by comparable target-discovery benefits and navigation costs; after candidate exhaustion, recovery directs reinspection to goal-relevant, insufficiently observed regions.

Waves protect pending inspections while admitting new visual cues. Historical target candidates first undergo on-site verification in order of current shortest navigable distance. Once they are exhausted and the target remains unfound, the system enters wave-based search over GSFs and geometric frontiers. To limit repeated displacement of pending inspections by newly generated nearby geometric frontiers, each wave w admits the currently executable geometric frontiers into a fixed membership set $\mathcal{W}^w$. Let $\mathcal{W}_t^w$ contain its valid, reachable, uninspected members and $\mathcal{O}_t^{\text{GSF}}$ denote executable GSFs. The current active set is

$$\mathcal{A}_t = \mathcal{W}_t^w \cup \mathcal{O}_t^{\text{GSF}}. \tag{7}$$

New geometric frontiers wait until the current geometric members are exhausted and then form the next wave; new GSFs promptly join the active set to respond to newly observed visual cues. Waves control candidate admission, while the unified rule below determines visit order within the active set. Only geometric membership is fixed; distances and validity continue to update during execution.

Unified ranking by discovery benefit and navigation cost. We compare GSFs and geometric frontiers through target visibility in the next formal observation $\mathcal{O}_o^+$ after execution. For their respective candidate types $e_o \in \{\text{context}, \text{geometric}\}$, define

$$Y_o = \mathbb{I}[\mathcal{O}_o^+ \text{ clearly shows a target matching } g],$$
$$p_{\text{find}}(o) = \widehat{p}_{e_o}, \qquad \widehat{p}_e = \frac{1}{|\mathcal{D}_e|} \sum_{o \in \mathcal{D}_e} Y_o. \tag{8}$$

The development set $\mathcal{D}_e$ contains type-e candidates whose execution produces a formal observation. Their empirical discovery rate is fixed before evaluation as the candidate-type prior p_{find}. A shared discovery event makes the types comparable without direct VLM scoring of each candidate (Appendix E.1).

For endpoint $\mathbf{x}_t(o)$, let $a_t(o) = d_{\text{nav}}(\mathbf{q}_t, \mathbf{x}_t(o))$ be the current navigable distance. The scheduler selects from active set $\mathcal{A}_t$:

$$o_t^* = \arg\min_{o \in \mathcal{A}_t} J_t(o), \qquad J_t(o) = \frac{a_t(o)}{p_{\text{find}}(o)}. \tag{9}$$

The ratio favors higher discovery benefits and shorter paths. Costs and validity update after each movement; Appendix E.2 gives the pairwise rationale under independent outcomes and fixed costs.

Recovery selects relevant, insufficiently observed locations. Recovery starts only when no historical target candidate, GSF, or active or pending geometric frontier remains executable. Recovery reuses the Scene Tree's regional index to associate goal relevance with mapped locations, while U_t identifies where accumulated observations remain insufficient. Scene Tree routing selects regions $\mathcal{R}_g^{\text{rec}}$. Reachable mapped cells associated with these regions and satisfying $C_t(c) < \tau$ are scored by

$$u(c \mid g) = \bar{p}_{\mathcal{T}}(R_g(c) \mid g)\, U_t(c), \quad \bar{p}_{\mathcal{T}}(R \mid g) = \frac{p_{\mathcal{T}}(R \mid g)}{\sum_{R' \in \mathcal{R}_g^{\text{rec}}} p_{\mathcal{T}}(R' \mid g)}. \tag{10}$$

Here $R_g(c)$ is the associated region. The score selects spatially dispersed endpoints, then navigation cost determines their visit order. New observations update memory and coverage; newly generated GSFs or geometric frontiers return the system to candidate scheduling (Appendix E.3). Any formal observation revealing the target initiates local approach and stopping verification (Appendix F).

3 Experiments

3.1 Evaluation Protocol

Datasets and metrics. The main evaluation uses the full GOAT-Bench val-unseen split (Khanna et al., 2024), with 36 scenes, 360 episodes, and 2,669 sequential subtasks. Following the reduced evaluation protocol commonly adopted in prior work (Yang et al., 2025; Zhang et al., 2025; Li et al., 2026a), controlled component comparisons use the first episode (episode 0) of each scene, yielding a fixed diagnostic subset of 36 episodes and 278 subtasks. TreeNav uses a single agent and a 1 m success threshold throughout; we report SR and SPL (Anderson et al., 2018). Appendix G.2 describes data collection and evaluation for retrieval comparisons; Appendices G and H give modality breakdowns and HM3Dv2 ObjectNav results.

Implementation and measurement. TreeNav and all ablations use Qwen3-VL-8B; each ablation changes only the component under study (Appendix G.5). Discovery priors are estimated on val-seen and fixed during val-unseen evaluation (Appendix E.1). Ground-truth positions and instance annotations are used only offline. SR and SPL include all subtasks; distance statistics use those successfully completed by each policy.

3.2 Overall Navigation Performance

TreeNav completes 1,952 of 2,669 subtasks, achieving 73.14% SR and 50.14% SPL (Table 1). These

Table 1: Single-agent results on the full GOAT-Bench val-unseen protocol (360 episodes, 2,669 subtasks, and a 1 m threshold). Baselines are taken from their original papers; MSGNav uses its threshold-sweep result. Different backbones make this a system-level comparison.

Method	Backbone / Policy	Training	SR ↑	SPL ↑
SenseAct-NN Monolithic (Khanna et al., 2024)	End-to-end policy	RL	12.3	6.8
Modular CLIP on Wheels (Khanna et al., 2024)	Modular pipeline	None	16.1	10.4
Modular GOAT (Khanna et al., 2024)	Modular pipeline	None	24.9	17.2
SenseAct-NN Skill Chain (Khanna et al., 2024)	Skill-specific policies	RL	29.5	11.3
EvoMemNav (Ge et al., 2026)	Qwen3-VL-8B	None	59.6	38.9
3D-Mem (Yang et al., 2025)	GPT-4o	None	62.9	44.7
STEGNav (Chen et al., 2026b)	GPT-5.4-mini	None	66.3	39.7
MSGNav (1 m) (Huang et al., 2026)	GPT-4o	None	66.52	–
TreeNav (Ours)	Qwen3-VL-8B	None	**73.14**	**50.14**

Table 2: Core ablations on the fixed subset of 36 episodes and 278 subtasks. ΔSPL is relative to the full method.

Variant	Controlled Change	SR ↑	SPL ↑	ΔSPL
Full method	None	81.65	58.94	–
M-flat	Replace the Scene Tree with budget-matched flat image memory	77.70	55.34	−3.60
M-none	Disable historical memory	58.63	17.88	−41.06
C-none	Disable GSF	77.34	53.02	−5.92
S-no-wave	Disable waves; retain a/p_{find}	77.70	51.88	−7.06
S-equal-prior	Use one prior for GSF and geometric frontiers	79.50	55.67	−3.27
S-VLM-score	Replace p_{find} with direct VLM scores	78.78	55.02	−3.92
C/S-none	Disable both GSF and waves	76.26	50.10	−8.84
R-no-lowcert	Disable low-certainty recovery	76.98	54.97	−3.97

exceed MSGNav's reported SR by 6.62 percentage points and 3D-Mem's SPL by 5.44 percentage points. Because backbones differ, this is a system-level comparison. TreeNav achieves 81.65% SR and 58.94% SPL on the diagnostic subset; Appendix G provides published reduced-protocol references and modality results.

3.3 How Does Regional Memory Improve Retrieval and Navigation?

Replacing the Scene Tree with budget-matched flat memory reduces SPL by 3.60 points (Table 2); removing memory entirely causes a much larger loss. Navigation thus benefits from both retaining observations and organizing them by region.

Following SG-Nav (Yin et al., 2024), replacing the Scene Tree with an object-level 3D scene graph reduces SR/SPL from 76.98%/54.97% (R-no-lowcert in Table 2) to 70.14%/48.36%, a 6.84/6.61-percentage-point drop. Both variants disable recovery because it relies on Scene Tree regions. Three factors help explain the gap: (i) **Regional indexing** helps filter irrelevant same-category lookalikes and reduces reliance on early object recognition; (ii) **Indirect cues** generate GSFs for unseen targets, whereas the baseline reuses only direct target sightings; (iii) **Raw-image verification** reexamines appearance and context for the current goal. Using the same object-level scene-graph baseline, we further test retrieval under early recognition errors with identical historical panorama pools and a $B = 12$ cap. Queries require a target-containing historical image; the natural-error subset comprises queries whose targets were previously missed or mislabeled in the scene graph's records. Hit@12 is the fraction of queries retrieving at least one target-containing image. The object-level scene-graph baseline and Scene Tree achieve Hit@12 of 70.65% and 80.60%, respectively, over all evaluated queries, and 46.67% and 75.00% on the natural-error subset. The advantage widens from 9.95 points overall to 28.33 on natural-error queries, supporting regional retrieval of raw images for goal-specific reexamination after early object-recognition errors.

Table 3: Historical retrieval (%). $B = 12$; exhaustive retrieval inspects all views.

Method	Target Hit@B ↑	Large-room Hit@B ↑	Spatial Redundancy ↓
Flat retrieval	68.66	58.82	57.21
Wall partition	71.14	62.75	12.48
Tree w/o region cap	73.13	66.67	14.69
Scene Tree	80.60	76.47	13.00
Exhaustive (all)	100.00	100.00	77.50

Hit rates use queries with historical target evidence; large rooms have at least $45\,\mathrm{m}^2$ of navigable area.
Redundancy uses 1 m same-floor spatial clusters.

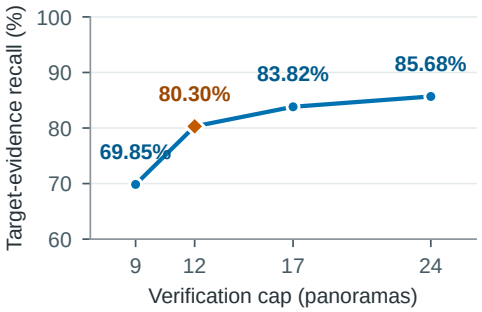

Figure 4: Full-evaluation recall. The diamond marks the default ($b_{\max} = 3$); the x-axis gives verification caps, not mean counts.

Table 4: Mechanism diagnostics. (a) Probabilities and Brier scores use a 0–1 scale. (b) Only successful subtasks are included. (c) Distance is averaged only over subtasks with successful recovery, while recovery SR includes all subtasks that triggered recovery. Distances are in meters and SR is a percentage.

(a) Discovery probability

Candidate	Prior	Test rate	Brier ↓
GSF	0.27	0.25	0.1879
Geometric	0.09	0.08	0.0737

(b) First-discovery distance

Scheduler	Median ↓	P90 ↓
Dynamic ranking	19.37	54.84
Wave scheduling	13.06	34.29

(c) Observation-sufficiency recovery

Recovery Policy	Mean Extra Distance (Successes) ↓	Recovery SR ↑
No recovery	–	0
Nearest reachable revisit	–	0
Observation sufficiency only	30.86	7.14
Scene-tree-conditioned sufficiency	22.81	21.43

To examine retrieval quality directly, retrievers process the same goals and historical panorama pools, with a cap of $B = 12$ panoramas except for exhaustive retrieval. Flat retrieval ranks panorama descriptions by goal relevance. Wall partitioning builds regions from known free-space geometry and uses the Scene Tree's regional filtering and representative-leaf rules. The no-region-cap variant retains the tree, summaries, and total budget, removing only the per-region limit. Appendix G.2 details these implementations.

Scene Tree achieves 80.60% target hit rate, 11.94 points above flat retrieval, with 13.00% spatial redundancy (Table 3). Its large-room hit rate is 76.47%, exceeding wall partitioning by 13.72 points and supporting subdivision by visual content. Removing regional caps reduces overall and large-room hit rates by 7.47 and 9.8 points, respectively, supporting verification allocation across distinct visual contexts.

Across all 360 episodes, the Scene Tree inspects 5149 of 29776 candidate historical leaves, reducing verification by 82.7% while retaining 80.30% target-evidence recall. Increasing the verification cap to 17 and 24 yields recall of 83.82% and 85.68%, respectively, indicating diminishing returns from additional verification (Figure 4; Appendix G.2).

3.4 How Do GSFs Support Historical Reuse and Online Exploration?

Removing GSFs while retaining historical target verification reduces SR/SPL by 4.31/5.92 percentage points (Table 2). The source ablation (Table 8) shows a 3.95/3.91-point loss from disabling historical GSFs, versus 1.07/2.08 for current-view GSFs. On this subset, using historical cues to seek targets absent from stored images provides the larger benefit; visual cues from the current viewpoint add observation locations missed by geometric frontiers, improving online exploration efficiency. Both sources perform best together.

Table 5: Count-matched GSF candidate comparisons. Discovery rate and SR/SPL are percentages; distance is the mean over successful subtasks, in meters.

Candidate Set	Next-Observation Discovery ↑	First-Discovery Distance ↓	SR ↑	SPL ↑
Geometric only	9.37	30.04	76.62	52.79
$+K$ dense geometric	14.22	26.25	78.78	54.44
$+K$ reachable points	13.66	27.03	78.42	53.88
$+K$ GSF	28.49	21.82	81.65	58.94

To isolate semantic cues from candidate count, groups share initial states and retain the original geometric frontiers. We compare no additions with K frontier-boundary samples, K reachable free-space samples independent of goal semantics, or K valid GSFs, where K is the state's valid GSF count. All three expansions match candidate counts, subsequent search, and execution budgets.

GSFs achieve 28.49% discovery versus 14.22% for dense frontiers. Against reachable points, they gain 14.83 points and reduce successful subtasks' mean discovery distance by 19.27%, supporting cue-guided placement beyond geometric densification (Table 5; Appendix G.3).

3.5 How Do Waves and Recovery Change Search?

Removing only waves reduces SPL by 7.06 points, the largest individual non-memory loss (Table 2). Under this model and evaluation setting, fixed type-specific discovery priors outperform equal priors and direct VLM candidate scores. Test discovery rates also remain close to development priors for both types (Table 4(a)).

Under matched states, candidate generation, and budgets, waves reduce successful subtasks' 90th-percentile (P90) first-discovery distance by 37.47%. The GSF–wave interaction is 4.14 SPL points: GSF gains rise from 1.78 without waves to 5.92 with waves, supporting cue-driven generation paired with protected inspection scheduling (Appendix G.4).

Removing recovery leaves 13 additional goals unfinished, reducing SR by 4.68 percentage points (Table 2). To separate revisit-location selection from extra search budget, four policies resume independently from the same first-exhaustion states: no recovery, nearest reachable historical location, observation sufficiency alone, and sufficiency conditioned on Scene Tree relevance. Active policies share the subsequent budget, endpoint spacing, and local execution rules. The latter two inspect underobserved known free space without requiring new frontiers. Recovery SR uses all triggered subtasks, counting each once; mean extra distance uses successful recoveries and includes resumed candidate search.

Nearest-location revisiting recovers no targets, while observation sufficiency alone achieves 7.14% recovery SR. Adding regional relevance raises it to 21.43% and reduces successful recoveries' mean extra distance from 30.86 to 22.81 m (Table 4(c)). This supports reusing regional context to direct additional observation toward relevant coverage gaps. Appendix G.4 gives policy definitions, full statistics, and the success-conditioned distance protocol.

4 Conclusion

TreeNav connects historical retrieval, cue-driven exploration, and inspection scheduling for multimodal lifelong navigation. The Scene Tree uses content-adaptive regional indexing to retrieve raw images for goal-specific reexamination; GSFs use identified search cues to propose observation locations before the target is seen; evidence-first waves limit geometric-frontier preemption, while recovery selects relevant, underobserved locations for reinspection. GOAT-Bench results and controlled ablations support these complementary designs. Together, they show the value of using historical and current visual evidence not only to recognize targets, but also to determine where to search and where to look again.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and editing, translation, and code organization. The authors reviewed the AI-assisted materials and take responsibility for the final manuscript, code, and claims.

REPRODUCIBILITY STATEMENT

We provide selected source-code excerpts for Scene Tree, GSF construction, wave scheduling, and targeted recovery as supplementary material. The excerpts do not include the model-serving, mapping, or benchmark-execution infrastructure and are not a standalone implementation. Appendices C–E specify the algorithms, and Appendix G documents the evaluation protocols and configurations. The complete codebase will be released publicly upon acceptance.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. URL https://arxiv.org/abs/1807.06757.

Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, Roozbeh Mottaghi, Jitendra Malik, and Devendra Singh Chaplot. GOAT: GO to any thing. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.073. URL https://www.roboticsproceedings.org/rss20/p073.html.

Peixin Chen, Guoxi Zhang, Jianwei Ma, and Qing Li. Hypothesis graph refinement: Hypothesis-driven exploration with cascade error correction for embodied navigation. *arXiv preprint arXiv:2604.04108*, 2026a. URL https://arxiv.org/abs/2604.04108.

Yang Chen, Zhenyu Huang, Wenbo Fu, Danyang Peng, Shi-Yu Tian, Kun-Yang Yu, and Lan-Zhe Guo. STEGNav: Spatio-temporal event graph reasoning for multimodal lifelong object navigation. *arXiv preprint arXiv:2608.28279*, 2026b. URL https://arxiv.org/abs/2608.28279.

Zuhao Ge, Xiaosong Jia, Chao Wu, Yuchen Zhou, Zuxuan Wu, and Yu-Gang Jiang. EvoMemNav: Efficient self-evolving fine-grained memory for zero-shot embodied navigation. *arXiv preprint arXiv:2606.03509*, 2026. URL https://arxiv.org/abs/2606.03509.

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 *IEEE International Conference on Robotics and Automation*, pp. 5021–5028, 2024. URL https://arxiv.org/abs/2309.16650.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation*, 2023. URL https://arxiv.org/abs/2210.05714.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2511.10376.

Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. In *Proceedings of Robotics: Science and Systems*, 2022. doi: 10.15607/RSS.2022.XVIII.050. URL https://www.roboticsproceedings.org/rss18/p050.html.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Ganesh Iyer, Soroush Saryazdi, Tao Chen, Alaa Maalouf, Shuang Li, Nikhil Varma Keetha, Ayush Tewari, Joshua Tenenbaum, Celso de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set multimodal 3D mapping. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.066. URL https://www.roboticsproceedings.org/rss19/p066.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16373–16383, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Khanna_GOAT-Bench_A_Benchmark_for_Multi-Modal_Lifelong_Navigation_CVPR_2024_paper.html.

Ji Li, Bo Wang, Jing Xia, Mingyi Li, and Shiyan Hu. HIMM: Human-inspired long-term memory modeling for embodied exploration and question answering. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2026a. URL https://arxiv.org/abs/2602.15513.

Linfeng Li, Jian Zhao, Yuan Xie, Xin Tan, and Xuelong Li. CompassNav: Steering from path imitation to decision understanding in navigation. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=eqcDckWHik.

Xueying Li, Feng Lyu, Hao Wu, Mingliu Liu, Jia-Nan Liu, and Guozi Liu. Stop wandering: Efficient vision-language navigation via metacognitive reasoning. *arXiv preprint arXiv:2604.02318*, 2026c. URL https://arxiv.org/abs/2604.02318.

Ruimeng Liu, Xinhang Xu, Shenghai Yuan, and Lihua Xie. Handle object navigation as weighted traveling repairman problem. *arXiv preprint arXiv:2503.06937*, 2025. URL https://arxiv.org/abs/2503.06937.

Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3D scene graphs. *IEEE Robotics and Automation Letters*, 9(10):8921–8928, 2024. doi: 10.1109/LRA.2024.3451395. URL https://arxiv.org/abs/2404.13696.

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*, 2026. URL https://arxiv.org/abs/2603.01813.

Santhosh K. Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for ObjectGoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ramakrishnan_PONI_Potential_Functions_for_ObjectGoal_Navigation_With_Interaction-Free_Learning_CVPR_2022_paper.html.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1904.01201.

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*, 2024. doi: 10.15607/RSS.2024.XX.077. URL https://www.roboticsproceedings.org/rss20/p077.html.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-Matterport 3D semantics dataset. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. URL https://arxiv.org/abs/2210.05633.

Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 146–151, 1997. doi: 10.1109/CIRA.1997.613851. URL https://doi.org/10.1109/CIRA.1997.613851.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17294–17303, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_3D-Mem_3D_Scene_Memory_for_Embodied_Exploration_and_Reasoning_CVPR_2025_paper.html.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. SG-Nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/098491b37deebbe6c007e69815729e09-Abstract-Conference.html.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. UniGoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19057–19066, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_UniGoal_Towards_Universal_Zero-shot_Goal-oriented_Navigation_CVPR_2025_paper.html.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation*, 2024. URL https://arxiv.org/abs/2312.03275.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: Leveraging large language models for visual target navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3554–3560, 2023. doi: 10.1109/IROS55552.2023.10342512. URL https://arxiv.org/abs/2304.05501.

Gengyuan Zhang, Mingcong Ding, Jingpei Wu, Ruotong Liao, and Volker Tresp. ReEXplore: Improving MLLMs for embodied exploration with contextualized retrospective experience replay. *arXiv preprint arXiv:2511.19033*, 2025. URL https://arxiv.org/abs/2511.19033.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42829–42842. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhou23r.html.

A Related Work

Spatial representations and lifelong visual memory. Spatial memories differ in how they encode observations and expose them to subsequent goals. Open-vocabulary feature maps support language-conditioned spatial queries (Huang et al., 2023; Jatavallabhula et al., 2023), whereas scene graphs organize object relations for structured reasoning (Gu et al., 2024; Yin et al., 2024). Hierarchical organization (Hughes et al., 2022; Werby et al., 2024) and task-driven abstraction (Maggio et al., 2024) further control the granularity of the representation. Multimodal goals and long-term reuse, as studied in GOAT and GOAT-Bench (Chang et al., 2024; Khanna et al., 2024), have also motivated goal–scene graph matching, event histories, and episodic–semantic memories (Yin et al., 2025; Chen et al., 2026b; Li et al., 2026a). Image-backed approaches retain appearance and background context for later reexamination. Their retrieval units differ: 3D-Mem filters co-visibility snapshots by detected categories (Yang et al., 2025), MSGNav retrieves image-bearing subgraphs through relevant object nodes (Huang et al., 2026), and EvoMemNav shortlists views by room-type

grouping (Ge et al., 2026). These approaches make image retrieval tractable, while differing in their dependence on earlier detections and their ability to distinguish local contexts within a room. TreeNav investigates content-adaptive regional indexing for goal-specific reexamination of raw observations.

Semantic exploration and observation selection. Frontier exploration generates destinations at the boundary between known free space and unexplored space (Yamauchi, 1997). Semantic methods prioritize these destinations using learned exploration potentials (Ramakrishnan et al., 2022), object–room commonsense (Zhou et al., 2023), or image–goal relevance and scene-graph reasoning (Yokoyama et al., 2024; Yin et al., 2024). Contextual experience replay and revisable semantic hypotheses extend this line by updating frontier decisions as evidence accumulates (Zhang et al., 2025; Chen et al., 2026a). Their main emphasis is the utility assigned to candidate destinations. In contrast, MSGNav's visibility-based viewpoint selection addresses final approach after the target has been localized (Huang et al., 2026). Cue-driven search addresses the complementary question of where to acquire observations before target localization.

Cost-aware search and revisitation. Search planning also considers the order and travel cost of visits. WTRP-Searcher minimizes semantically weighted waiting times over viewpoints (Liu et al., 2025), while SSMG-Nav plans over semantic candidates anchored to topological keypoints (Niu et al., 2026). Both move beyond independent local destination scores to organize search globally. Complementary mechanisms address incorrect commitments or stalled progress: SG-Nav re-perceives suspected targets to reject false detections (Yin et al., 2024), whereas EvoMemNav temporarily forces frontier exploration after repeated selections or cooldown-induced deadlock (Ge et al., 2026). These mechanisms target route efficiency, verification errors, and search deadlock; our work additionally considers candidate admission and goal-related observation gaps in mapped space.

B Limitations and Future Work

Regional context. Scene Tree benefits from discriminative regional context. Its retrieval advantages may diminish in environments with repetitive layouts or weak associations between regional functions and target locations.

Dynamic environments. Our evaluation is limited to static indoor simulations. When objects move while regional functions remain stable, regional context may provide more persistent search guidance than recorded object locations. This potential requires dynamic-scene validation and mechanisms for refreshing historical observations.

Adaptive discovery estimates. Candidate-type discovery priors remain fixed during testing. A future extension could update discovery estimates from online target-verification feedback, using development priors for smoothing to adapt to changes in scene and goal distributions.

C Online Growth of the Scene Tree

Algorithm 1 details the online growth procedure in Section 2.2. Observations are deduplicated by acquisition location: an existing leaf is reused and its parent becomes the active region; otherwise, a new leaf stores the four-view RGB-D observation, pose, and camera parameters. Candidate parents comprise the root, the parents of the nearest $K = 5$ historical leaves, and their ancestors, with $\mathrm{Anc}(v)$ including v itself. The VLM compares the new panorama with these regions and returns a growth operation α_t, parent p_t, and new regional label s_t, without using the current navigation goal.

Regional summaries and spatial updates. Regional summaries describe types and functions such as kitchens and hallways. MERGE adds an observation to the same region and reuses its summary; a new region receives a summary whose information is propagated along the ancestor chain by REFRESHREGIONALSUMMARIES. After each leaf insertion, REFRESHSPATIALCENTERS updates affected ancestors' spatial centers, recording new observation locations while keeping regional summaries compact.

Parent and child node types. LEAFCOMPATIBLE and INTERNALCOMPATIBLE insert grouping nodes as needed to prevent a parent from directly mixing regional nodes with observation leaves. These groups inherit existing regional information; nodes directly connected to observation leaves form the terminal regions used for retrieval and recovery.

Algorithm 1 Online growth of the Scene Tree from a formal observation.

Require: Scene Tree $\mathcal{T}$ with root r; formal observation $\mathcal{O}_t = (I_t, D_t, \mathbf{q}_t, \boldsymbol{\kappa}_t)$; $K = 5$
Ensure: Updated Scene Tree $\mathcal{T}$

$\ell^\star \leftarrow$ FINDLEAFAT$(\mathcal{T}, \mathbf{q}_t)$
if $\ell^\star \neq \emptyset$ **then**
  $\mathcal{T}$.activeParent $\leftarrow \mathrm{parent}(\ell^\star)$
  return $\mathcal{T}$
end if
$\ell_t \leftarrow$ MAKELEAF$(I_t, D_t, \mathbf{q}_t, \boldsymbol{\kappa}_t)$
$\mathcal{N}_t \leftarrow$ KNEARESTLEAVES$(\mathcal{T}, \mathbf{q}_t, K)$
$\mathcal{P}_t \leftarrow \mathrm{Unique}\left(\{r\} \cup \bigcup_{\ell \in \mathcal{N}_t} \mathrm{Anc}(\mathrm{parent}(\ell))\right)$
$(\alpha_t, p_t, s_t) \leftarrow$ VLM-GROW$(I_t, \mathbf{q}_t, \mathcal{N}_t, \mathcal{P}_t)$
  $\alpha_t \in \{$MERGE, SUBFORK, ROOTFORK$\}$
if $\alpha_t =$ MERGE **then**
  $p_t \leftarrow$ RESOLVELOCALPARENT$(p_t, \mathcal{P}_t, \mathcal{T}.\mathrm{activeParent})$
  $p_t \leftarrow$ LEAFCOMPATIBLE$(\mathcal{T}, p_t)$
  ATTACHLEAF$(\mathcal{T}, p_t, \ell_t)$
else if $\alpha_t =$ SUBFORK **then**
  $p_t \leftarrow$ RESOLVELOCALPARENT$(p_t, \mathcal{P}_t, \mathcal{T}.\mathrm{activeParent})$
  $p_t \leftarrow$ INTERNALCOMPATIBLE$(\mathcal{T}, p_t)$
  $u_t \leftarrow$ ADDREGION$(\mathcal{T}, p_t, s_t)$
  ATTACHLEAF$(\mathcal{T}, u_t, \ell_t)$
else
  $u_t \leftarrow$ ADDREGION$(\mathcal{T}, r, s_t)$
  ATTACHLEAF$(\mathcal{T}, u_t, \ell_t)$
end if
$p_t \leftarrow \mathrm{parent}(\ell_t)$
if p_t has no regional summary **then**
  $h_{p_t} \leftarrow$ VLM-REGIONSUMMARY(I_t, s_{p_t}) ▷ Regional identity and function
  REFRESHREGIONALSUMMARIES$(\mathcal{T}, p_t)$
end if
REFRESHSPATIALCENTERS$(\mathcal{T}, p_t)$
$\mathcal{T}$.activeParent $\leftarrow p_t$
VALIDATETREE$(\mathcal{T})$
return $\mathcal{T}$

D GOAL-CONDITIONED ROUTING AND REGION SELECTION

Branch scoring. Routing compares sibling regions for goal relevance at each level. Given goal g, the VLM jointly reads the labels s_u and summaries h_u of regional children $u \in \mathcal{C}_v$ under node v, assigning nonnegative scores $z_v(u; g)$. Normalizing these scores gives branch probabilities, whose product along a path gives the regional probability:

$$\pi(u \mid v, g) = \frac{z_v(u; g)}{\sum_{w \in \mathcal{C}_v} z_v(w; g)}, \qquad p_{\mathcal{T}}(u \mid g) = \prod_{(a \to b) \in \mathrm{path}(u)} \pi(b \mid a, g). \tag{11}$$

If all valid scores are zero, branches receive equal probability. Algorithm 2 uses a stack to visit higher-scoring branches first and compute terminal-region path probabilities.

Cumulative-probability selection. Terminal regions are retained in descending path-probability order until their cumulative probability reaches $\rho = 0.85$, subject to count bounds $R_{\min} = 2$ and $R_{\max} = 8$. All regions are retained if fewer than two exist; no more are added after reaching the upper bound. Original-image verification draws observations only from the retained regions.

Observation sufficiency of representative subsets. Within each retained region, historical leaves are first spatially deduplicated by floor and acquisition location using a 2 m radius. This filters verification candidates while preserving stored observations. Stored depth and poses quantify how well each view subset observes the surrounding space. Let $\mathcal{L}_r$ be the remaining historical leaves of region R_r, $\mathcal{V}_\ell$ the cells covered by valid depth in leaf ℓ, and $\mathbf{q}_\ell$ its acquisition pose. Per-leaf contribution and subset sufficiency are

$$\gamma_\ell(c) = \begin{cases} \max\{0, 1 - d(c, \mathbf{q}_\ell)/d_{\max}\}, & c \in \mathcal{V}_\ell, \\ 0, & c \notin \mathcal{V}_\ell, \end{cases} \qquad C_{\mathcal{S}}(c) = \max_{\ell \in \mathcal{S}} \gamma_\ell(c), \qquad U_{\mathcal{S}}(c) = 1 - C_{\mathcal{S}}(c). \tag{12}$$

The coverage resolution and maximum valid depth $d_{\max} = 5\,\mathrm{m}$ are inherited from the observation map. Each cell takes the largest contribution from selected views, or zero if uncovered. Equation (4) sums $C_{\mathcal{S}}(c)$ over a common map grid without geometric region boundaries; maximizing this sum is equivalent to minimizing total residual uncertainty. Whereas $C_{\mathcal{S}}$ measures only the selected historical views, C_t accumulates online observations. Maximum sufficiency may be attained with fewer leaves than the budget allows.

Regional budgets. Verification slots follow regional relevance rank, with per-region leaf cap

$$b_r = \min\left(|\mathcal{L}_r|, \max(1, b_{\max} - \lfloor r/2 \rfloor)\right), \qquad r = 1, 2, \ldots. \tag{13}$$

With default $b_{\max} = 3$, the first eight caps are $(3, 2, 2, 1, 1, 1, 1, 1)$, totaling at most 12 four-view panoramas. Fewer retained regions or available leaves, or a smaller subset attaining maximum sufficiency, reduce the actual verification count. Selected views $\mathcal{S}_r^\star$ support target judgment and cue extraction. In Figure 2, cap is the rank-based limit and need not be filled; Figure 4 compares different budgets.

Deterministic subset optimization. Equation (4) is solved by exhaustive enumeration within each retained region. After spatial deduplication, all nonempty view subsets of size at most b_r are evaluated, and the subset with the largest total coverage is selected. Equal-coverage ties favor fewer views, then the lexicographically smaller tuple of candidate indices. Indices follow the fixed order of the deduplicated candidate list. The default cap is at most three panoramas per region.

E Discovery Priors, Candidate Ordering, and Recovery

E.1 Offline Annotation of Target-Discovery Events

Event collection. The first formal panorama after candidate execution defines an observation event, labeled by whether it reveals the target. Development data use the first 3 episodes of each val-seen scene in numeric identifier order, collected with the reference policy in Appendix G.2 and random seed 0. For executed GSFs and geometric frontiers, we save the candidate identifier, type, goal, poses before and after execution, and this RGB-D panorama, recording at most one event per execution. Navigation failures without a formal observation are logged separately and excluded from the denominator. Test events follow the same rules on reference trajectories from the val-unseen diagnostic subset. Both event lists are fixed before analysis.

Target-visibility annotation. Valid targets are determined by the task: allowed same-category instances for category goals, and the specified instance for language and image goals. Aligned RGB images and instance masks are resized to the diagnostic model's input dimensions, preserving the RGB aspect ratio and using nearest-neighbor interpolation for masks. A panorama is positive if a valid target in any view occupies at least 0.1% of that view and at least 64 pixels. Two annotators review all positives and a fixed-seed sample of 10% of negatives, correct instance-matching or rendering errors, and resolve disagreements jointly. All methods use this criterion:

$$Y_o = \mathbb{I}\left[\begin{array}{c}\text{the next formal panorama after candidate execution} \\ \text{meets the target-visibility criterion}\end{array}\right]. \tag{14}$$

Panoramas showing only related objects or room cues are labeled zero. Historical panoramas use the same annotation rules, and retrieval hit rate is measured on queries with target evidence in history.

Algorithm 2 Scene Tree routing with spatial deduplication and image budgets.

Require: Scene Tree $\mathcal{T}$ with root r; goal g; historical coverage γ_ℓ
Parameters: $\rho = 0.85$; $R_{\min} = 2$; $R_{\max} = 8$; $b_{\max} = 3$
Ensure: Representative observation leaves $\mathcal{Q}$

$\mathcal{F} \leftarrow [(r, 1)]$; $\mathcal{R} \leftarrow \emptyset$; $\mathcal{Q} \leftarrow \emptyset$
while $\mathcal{F} \neq \emptyset$ **do**
  $(v, p_v) \leftarrow \textsc{Pop}(\mathcal{F})$
  if v directly contains observation leaves **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{(v, p_v)\}$; **continue**
  end if
  if $\mathcal{C}_v = \emptyset$ **then**
    continue
  end if
  $\pi_v \leftarrow \textsc{SiblingProbabilities}(g, \{(s_u, h_u) : u \in \mathcal{C}_v\})$
  for all $u \in \textsc{AscendingOrder}(\mathcal{C}_v, \pi_v)$ **do**
    $\textsc{Push}(\mathcal{F}, (u, p_v \pi_v(u)))$
  end for
end while
$\mathcal{R} \leftarrow \textsc{DescendingOrder}(\mathcal{R}, p_v)$
$\widehat{\mathcal{R}} \leftarrow [\,]$; $M \leftarrow 0$
for all $(v, p_v) \in \mathcal{R}$ in ranked order **do**
  if $|\widehat{\mathcal{R}}| \geq R_{\max}$ **or** $(|\widehat{\mathcal{R}}| \geq R_{\min}$ **and** $M \geq \rho)$ **then**
    break ▷ Prune the remaining retrieval regions
  end if
  $\textsc{Append}(\widehat{\mathcal{R}}, (v, p_v))$; $M \leftarrow M + p_v$
end for
for all (v, p_v) at rank r_v in $\widehat{\mathcal{R}}$ **do**
  $\mathcal{L}_v \leftarrow \textsc{SpatialDeduplicate}(\mathcal{C}_v, 2\,\mathrm{m})$
  $b_{r_v} \leftarrow \min(|\mathcal{L}_v|, \max(1, b_{\max} - \lfloor r_v/2 \rfloor))$
  $\mathcal{S}_v^\star \leftarrow \textsc{MaxSufficiencySubset}(\mathcal{L}_v, b_{r_v})$
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{S}_v^\star$
end for
return $\mathcal{Q}$

Priors and test statistics. Let $\mathcal{D}_e^{\mathrm{dev}}$ denote valid observation events of type e in the development set. The empirical prior is

$$\widehat{p}_e = \frac{1}{|\mathcal{D}_e^{\mathrm{dev}}|} \sum_{o \in \mathcal{D}_e^{\mathrm{dev}}} Y_o. \tag{15}$$

Priors are fixed before navigation evaluation; Table 4(a) reports their values. Let $\mathcal{D}_e^{\mathrm{test}}$ denote test events of the same type. Their discovery rate and Brier score are

$$\widehat{r}_e = \frac{1}{|\mathcal{D}_e^{\mathrm{test}}|} \sum_{o \in \mathcal{D}_e^{\mathrm{test}}} Y_o, \qquad \mathrm{Brier}_e = \frac{1}{|\mathcal{D}_e^{\mathrm{test}}|} \sum_{o \in \mathcal{D}_e^{\mathrm{test}}} (\widehat{p}_e - Y_o)^2. \tag{16}$$

Because candidates of each type share a fixed prior, the Brier score equals $\widehat{r}_e(1-\widehat{p}_e)^2 + (1-\widehat{r}_e)\widehat{p}_e^2$. Development and test statistics are conditioned on obtaining a formal observation; instance masks, target identifiers, and complete reference maps are used for offline annotation.

E.2 Rationale for Candidate Ordering

Consider candidates i, j with independent outcomes, each executed at most once. Their discovery probabilities are $p_i, p_j > 0$, and costs a_i, a_j are fixed and independent of visit order. Search stops at discovery, giving expected costs

$$E_{i \to j} = a_i + (1 - p_i)a_j, \qquad E_{j \to i} = a_j + (1 - p_j)a_i. \tag{17}$$

Therefore,

$$E_{i \to j} \leq E_{j \to i} \iff p_j a_i \leq p_i a_j \iff \frac{a_i}{p_i} \leq \frac{a_j}{p_j}. \tag{18}$$

This comparison motivates ranking by cost divided by discovery probability. Online, movement changes navigation distances and map updates generate new candidates, so each decision recomputes costs and checks candidate validity. Wave rules control admission separately: new geometric candidates are delayed to avoid repeatedly postponing existing candidates, while GSFs from new cues join promptly.

E.3 Selecting Recovery Regions and Observation Locations

Observation map initialization. As defined in Section 2.1, sufficiency starts at $C_0(c) = 0$, except at the robot footprint, where it is one. Formal panoramas update only cells covered by valid depth. The map retains the best observation sufficiency at each location across subtasks.

Goal-relevant regions. Recovery begins once historical target candidates, GSFs, and both current and pending geometric candidates are exhausted. The Scene Tree selects up to $R_{\mathrm{rec}} = 6$ terminal regions with the highest $p_{\mathcal{T}}(R \mid g)$ to form $\mathcal{R}_g^{\mathrm{rec}}$. Their historical observation locations associate nearby cells with regions for further inspection.

Spatial association of cells and regions. For a reachable free cell c in the current map with $C_t(c) < \tau$, let $L_{\mathrm{floor}(c)}(R)$ contain region R's historical leaves on the same floor. The distance to the nearest historical observation in this region is

$$\delta(c, R) = \min_{\ell \in L_{\mathrm{floor}(c)}(R)} \|\mathbf{x}_c - \mathbf{x}_\ell\|_{XZ}. \tag{19}$$

Here, XZ is the horizontal plane; the distance is $+\infty$ if the same-floor leaf set is empty. Regions within distance r form the association set:

$$\mathcal{B}_g(c) = \{R \in \mathcal{R}_g^{\mathrm{rec}} : \delta(c, R) \le r\}. \tag{20}$$

The radius r equals the maximum valid depth, 5 m. Only cells with nonempty $\mathcal{B}_g(c)$ are retained as recovery candidates.

Resolving overlapping associations. If several regions are associated with a cell, we choose the one with the highest routing score, breaking ties by the distance to its nearest historical observation:

$$R_g(c) = \operatorname*{arg\,max}_{R \in \mathcal{B}_g(c)} {}_{\mathrm{lex}} \big(p_{\mathcal{T}}(R \mid g), -\delta(c, R)\big), \tag{21}$$

where lexicographic ordering compares relevance first, then distance.

Candidate generation and execution. We first normalize routing probabilities over the selected regions:

$$\bar{p}_{\mathcal{T}}(R \mid g) = \frac{p_{\mathcal{T}}(R \mid g)}{\sum_{R' \in \mathcal{R}_g^{\mathrm{rec}}} p_{\mathcal{T}}(R' \mid g)}, \tag{22}$$

and then score candidate cells by

$$u(c \mid g) = \bar{p}_{\mathcal{T}}(R_g(c) \mid g)\,[1 - C_t(c)]. \tag{23}$$

The score combines goal relevance and observation uncertainty to select spatially dispersed endpoints; navigation cost determines their visit order. New observations update memory and coverage. Detecting the target initiates local verification, while new GSFs or geometric frontiers return the agent to candidate scheduling.

F Local Execution after Target Recognition

When a formal observation clearly shows the target, TreeNav suspends candidate scheduling and localizes it using the image, depth, camera parameters, and robot pose. It selects reachable viewpoints near the target, approaches incrementally, and acquires a formal observation after each movement for verification. The subtask ends when the pose satisfies the stopping condition; scheduling resumes if observations no longer support further approach. These local viewpoints serve to approach a recognized target and do not enter wave ranking as GSFs or geometric frontiers.

G Supplementary Results and Controlled Evaluation Protocols

G.1 Published Results and Performance by Goal Modality

Table 6 compares published GOAT-Bench results with TreeNav on the fixed 278-subtask diagnostic subset. Published methods retain their reported backbones and evaluation protocols as system-level references.

Table 6: Published GOAT-Bench results and TreeNav's results on 278 subtasks. SR and SPL are percentages; published rows retain their reported backbones, protocols, and numerical precision.

Method	Vision-Language Model	SR ↑	SPL ↑
3D-Mem (Yang et al., 2025)	LLaVA-7B	49.6	29.4
3D-Mem (Yang et al., 2025)	GPT-4o	69.1	48.9
ReEXplore (Zhang et al., 2025)	Qwen2.5-VL-7B	53.2	32.6
ReEXplore (Zhang et al., 2025)	GPT-4o	59.8	42.5
HIMM (Li et al., 2026a)	Qwen2.5-VL-7B	53.9	31.7
HIMM (Li et al., 2026a)	GPT-4o	72.8	56.1
MetaNav (Li et al., 2026c)	GPT-4o	71.4	51.8
TreeNav (Ours)	Qwen3-VL-8B	81.65	58.94

Table 7 reports subtask counts, SR, and SPL for object-category, language-description, and reference-image goals. Overall metrics are weighted by subtask count using unrounded per-modality statistics. Appendix G.3 analyzes the contributions of current and retrieved views to GSFs.

Table 7: Goal-modality breakdown on the full val-unseen protocol. SR and SPL are percentages.

Goal Modality	Subtasks	SR ↑	SPL ↑
Object category	991	81.43	58.28
Language description	856	62.85	42.61
Reference image	822	73.84	48.18
Overall	2,669	73.14	50.14

G.2 Scene Tree Retrieval: Query Construction and Metrics

Constructing shared queries. All retrievers process the same saved goals and historical observations offline. On episode 0 of each of the 36 scenes, we run the full TreeNav once as the reference collection policy with random seed 0, retaining memory, GSFs, waves, and recovery. After receiving each subtask goal, but before any new action or observation insertion, we save the map, Scene Tree, and historical panorama pool H_q. This pool contains panoramas from earlier subtasks in the same episode. Each nonempty pool and its goal form a query in $\mathcal{Q}$, accompanied by scene, episode, subtask, timestamp, and panorama identifiers.

Panorama descriptions and flat retrieval. All comparisons use Qwen3-VL-8B, greedy decoding, and the navigation experiments' image preprocessing, with four 640×640 RGB views per panorama. The model first generates and caches panorama descriptions without a goal. Flat retrieval then scores each description independently for goal relevance on a 0–100 scale and selects the top $B = 12$, breaking ties by panorama identifier.

Wall partitioning. Geometric partitioning uses the known layered free-space map at query time. Obstacle distances are computed on 0.05 m cells, free space is eroded by 0.60 m, and the remaining connected components become regional seeds, excluding those smaller than $1.0\,\mathrm{m}^2$. A free-space component without a seed uses its cell farthest from obstacles. Regions then expand along shortest paths in the original free space. Panoramas are assigned by acquisition location, and their cached descriptions generate regional summaries. This partition follows geometric separation, leaving open spaces in the same region.

Region selection and budget controls. Wall partitioning and the Scene Tree score regional summaries using the same $\rho = 0.85$, $R_{\min} = 2$, $R_{\max} = 8$, and maximum-sufficiency subset rule from Appendix D. Each retained region first receives one slot; further slots follow relevance rank up to $b(r) = \max(1, 3 - \lfloor r/2 \rfloor)$ and the available leaf count. The no-region-cap variant raises the per-region limit during this additional allocation to B while retaining total cap B, testing the role of cross-region allocation. Unused slots are not filled. Exhaustive retrieval inspects all panoramas in H_q as an unbudgeted reference.

Positive samples and target hit rate. Historical pools are annotated using the visibility rules in Appendix E.1. Let $P_q \subseteq H_q$ contain panoramas showing a valid instance of the current target, S_q be the retrieval result, and $\mathcal{Q}^+ = \{q : P_q \neq \varnothing\}$ be queries with positive samples. Target hit rate is

$$\text{Hit@}\, B = \frac{100}{|\mathcal{Q}^+|} \sum_{q \in \mathcal{Q}^+} \mathbb{I}[S_q \cap P_q \neq \varnothing]. \tag{24}$$

A result containing at least one positive sample is a hit; an empty result is a miss. Since the denominator includes only queries with historical target evidence, exhaustive retrieval achieves 100%. Queries without such evidence still contribute to spatial redundancy. Hit rate measures recall of direct target evidence; source ablations separately assess the navigation contribution of contextual cues that generate GSFs without showing the target.

Verification volume under the full protocol. Across the full 360 episodes, retrievals consider 29776 candidate historical leaves in total and verify 5149, reducing verification by $100\big(1 - 5149/29776\big) \approx 82.7\%$. Equation (24) gives 80.30% target-evidence recall on positive queries from the full run. The 80.60% in Table 3 uses the same metric but covers only the fixed diagnostic subset.

Verification-budget comparison. Figure 4 compares $b_{\max} \in \{2, 3, 4, 5\}$ on the full evaluation using the budget function in Appendix D. With $R_{\max} = 8$, total cap $B = \sum_{r=1}^{8} \max(1, b_{\max} - \lfloor r/2 \rfloor)$ is 9, 12, 17, or 24 four-view panoramas, with recall of 69.85%, 80.30%, 83.82%, and 85.68%, respectively. The horizontal axis gives the configured cap; actual verification depends on retained regions, available leaves, and the selected subsets.

Large-room target hit rate. This metric evaluates retrieval of target evidence inside large rooms. Before retrieval, two annotators delineate and review room boundaries using the complete scene's walls, door openings, and layered traversability map. Open spaces without doorway separation form one room; corridors and stairs are excluded. Navigable room area is $A(R) = N_{\text{free}}(R)\, s^2$, with $s = 0.05$ m; rooms enter the large-room group if $A(R) \geq 45\,\text{m}^2$. Let P_q^L contain panoramas acquired inside a large room that show a valid target in that room, and $\mathcal{Q}^L = \{q : P_q^L \neq \varnothing\}$. Substituting P_q^L and $\mathcal{Q}^L$ for P_q and $\mathcal{Q}^+$ in Eq. (24) gives large-room hit rate. Room annotations are used only for offline statistics.

Spatial redundancy. Spatial redundancy measures repeated coverage of the same local area in retrieval results. Historical panoramas are clustered in acquisition-time and identifier order: a panorama joins the earliest-created cluster whose members are all within $r = 1.0$ m in same-floor navigation distance, or starts a new cluster if none qualifies. Distances use the known map at query time and are infinite across floors or without a traversable path. With $n_q = |S_q|$ and C_q denoting the number of clusters represented by selected panoramas,

$$\text{Redundancy} = \frac{100}{|\mathcal{Q}_{\text{out}}|} \sum_{q \in \mathcal{Q}_{\text{out}}} \left(1 - \frac{C_q}{n_q}\right), \quad \mathcal{Q}_{\text{out}} = \{q \in \mathcal{Q} : n_q > 0\}. \tag{25}$$

All retrievers share these clusters, defined only by locations and navigation distances, for consistent measurement.

G.3 GSFs: Candidate Comparisons and Evidence Sources

Starting states and candidate generation. Groups start from the same goal, panorama, pose, occupancy map, geometric frontiers, Scene Tree, active wave, and remaining budget, varying only

candidate-location generation. They retain the original geometric frontiers and add either no candidates, K boundary samples, K ordinary reachable points, or K GSFs. Here, K equals the state's valid GSF count, so the last three groups add the same number of candidates at locations determined by their respective rules. Subsequent search and local verification are identical. The system-level ablation in Table 2 instead disables GSFs throughout navigation.

Common treatment of added candidates. In the three $+K$ controls in Table 5, all added candidates use the GSF discovery prior, ranking and admission rules, with identical validity checks and execution budgets. The original geometric frontiers receive the same treatment across groups; only the locations of the added candidates differ.

Candidate events and task metrics. Let $\mathcal{E}_m$ contain events in which method m executes a candidate and obtains a formal panorama, with Y_o annotated by the visibility rules in Appendix E.1. Candidate-event discovery rate is

$$\text{Discovery}_m = \frac{100}{|\mathcal{E}_m|} \sum_{o \in \mathcal{E}_m} Y_o. \tag{26}$$

Each execution uses only its first subsequent formal panorama, so a subtask may contribute several events; executions without a formal observation are excluded from $\mathcal{E}_m$. SR/SPL use all 278 diagnostic subtasks. First-discovery distance accumulates from search start until the target first appears in a formal panorama; Table 5 reports its mean over successful subtasks.

Current and retrieved sources. The source ablation retains the Scene Tree and historical target verification, changing only GSF sources: none, current views during exploration, retrieved views during historical inspection, or both. In Table 8, current cues alone yield 77.70% SR and 55.03% SPL, historical cues alone yield 80.58% and 56.86%, and both yield 81.65% and 58.94%. Removing historical cues reduces SR/SPL by 3.95/3.91 points, more than the 1.07/2.08 points lost without current cues. Because historical target verification remains enabled, the historical-source gain reflects reuse of indirect evidence: related objects and spatial cues can suggest where to search even when the image does not show the target. Using both sources also outperforms historical cues alone, indicating that current-view cues further improve exploration by adding observation locations missed by geometric frontiers.

Table 8: GSF source ablations on the fixed diagnostic subset, retaining historical target verification in all variants. SR and SPL are percentages.

GSF source	SR ↑	SPL ↑
None	77.34	53.02
Current views only	77.70	55.03
Retrieved views only	80.58	56.86
Both sources	81.65	58.94

G.4 Scheduling and Recovery Metrics

Discovery-prior transfer. Test discovery rates and Brier scores in Table 4(a) are computed from the diagnostic subset's reference trajectories. GSF and geometric-frontier discovery rates are 0.27 and 0.09 in development, versus 0.25 and 0.08 in testing. The candidate types remain distinct in testing, and each rate stays close to its prior fixed on development data. Collection and annotation are described in Appendix E.1.

Scheduling comparisons. In Table 2, the no-wave variant retains a/p_{find} ranking and removes only the admission delay for new geometric candidates. Equal priors assign the same prior to GSFs and geometric frontiers; VLM scoring replaces the fixed discovery prior in the ranking denominator with each candidate's direct score. These latter variants retain waves, and all other settings match the full method.

Distances on successful subtasks. Let $\mathcal{S}_m$ be the subtasks completed successfully by policy m, and $d_{m,q}$ the cumulative movement from the shared search start to the first formal observation containing the target. GSF candidate comparisons report the mean of $\{d_{m,q} : q \in \mathcal{S}_m\}$; scheduling comparisons report the median and P90. These distances use each policy's own successful subtasks, while SR/SPL include all subtasks, including failures.

GSF–wave interaction. Let S_{cw} denote SPL for GSF setting c and wave setting w. Define

$$\Delta_{\text{int}} = S_{11} - S_{10} - S_{01} + S_{00}. \tag{27}$$

The four settings have $S_{11} = 58.94$, $S_{10} = 51.88$, $S_{01} = 53.02$, and $S_{00} = 50.10$, yielding 4.14 SPL points. On this diagnostic subset, the joint SPL gain thus exceeds the sum of the gains from enabling each component separately.

Recovery success rate and extra distance. Table 4(c) compares three active recovery policies. Nearest-location revisiting selects the reachable historical observation location with the shortest navigation distance among those not yet visited in the current recovery phase. Sufficiency-only recovery selects endpoints using $1 - C_t(c)$; the full method further weights them by Scene Tree relevance. The latter two revisit underobserved locations in known free space. All three share the subsequent budget, endpoint spacing, and local execution rules, varying only location selection.

Let $\mathcal{R}$ contain subtasks that trigger recovery in the reference run. We save the state at each subtask's first candidate exhaustion and resume each policy independently from it. Let $\mathcal{R}_m^+ \subseteq \mathcal{R}$ be the subtasks subsequently completed by policy m. Recovery success rate and mean extra distance over successful recoveries are

$$\text{RecoverySR}_m = 100\frac{|\mathcal{R}_m^+|}{|\mathcal{R}|}, \qquad \bar{d}_m^{\text{rec}} = \frac{1}{|\mathcal{R}_m^+|}\sum_{q\in\mathcal{R}_m^+} d_{m,q}^{\text{rec}}. \tag{28}$$

Here, $d_{m,q}^{\text{rec}}$ includes all movement from the first recovery trigger to success, including subsequent candidate search. Each subtask is counted once, and policies share the same remaining task budget. Individual movements are capped at 20 m; cumulative distance can exceed this value. The no-recovery policy stops at the trigger, yielding zero recovery SR and no applicable success-conditioned distance.

G.5 Reproducibility Settings

Experiments use the Habitat simulation platform (Savva et al., 2019). Table 9 summarizes model, sensor, navigation execution, and evaluation settings. Retrieval parameters are given in Appendix D; data collection and retrieval comparisons are described in Appendix G.2.

Table 9: Model, execution, and evaluation configuration.

Configuration item	Setting
VLM	Qwen3-VL-8B
Image size and camera intrinsics	640×640; $f_x = f_y = 203.86$, $c_x = c_y = 319.05$
Decoding temperature; token cap	0, 4096
Main navigation seed	100
Shared collection seed	0
Simulator version	Habitat-Sim 0.2.3
Maximum valid depth (m)	5.00
Local approach/stopping (m)	1.00
Observation-sufficiency threshold τ	0.50 (dimensionless)
Recovery movement limit (m)	20.00
Target visibility threshold and review rules	Target-instance area $\geq 0.1\%$ and ≥ 64 pixels; manual review of all positives and a random 10% sample of negatives

Online navigation uses RGB-D observations, acquisition poses, and an incremental navigation map. Simulator instance masks and target ground truth serve offline annotation and evaluation; a separately recorded navigation mesh is used for visualization.

Indexing cost. On an NVIDIA RTX A6000, inserting a new formal panorama into the Scene Tree takes 2.49 s on average, while a complete goal-conditioned regional routing pass takes 4.94 s on average.

G.6 Recovery Example

Figure 5 illustrates recovery after candidate exhaustion. Scene Tree routing selects a relevant bedroom region, and the sufficiency map identifies an underobserved location within it. Revisiting this location reveals the hanging-clothes target.

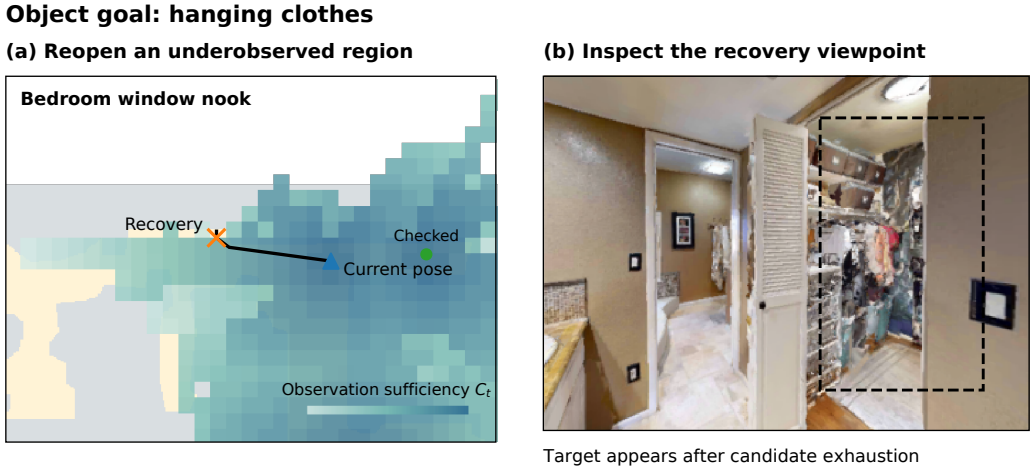

Figure 5: Recovery after candidate exhaustion. Moving to the selected endpoint (cross) reveals the hanging-clothes target. Left: sufficiency map C_t accumulated from formal RGB-D observations. Right: the target seen after the revisit.

G.7 Cross-Backbone Evaluation

We evaluate the backbones in Table 10 on the same 36-episode, 278-subtask diagnostic subset. Each backbone handles all VLM calls, including tree growth, regional summaries, routing, target verification, and cue extraction, and builds its memory online from scratch in each episode. Navigation settings, task and verification budgets, and prompt semantics are shared, with input formatting adapted to each model interface. For each backbone, candidate-type discovery priors are estimated from the same development data using the same procedure in Appendix E.1 and remain fixed throughout testing. SR and SPL include all 278 subtasks.

Table 10: Cross-backbone results on the fixed 36-episode, 278-subtask diagnostic subset. Discovery priors are shown for GSFs and geometric frontiers; SR and SPL are percentages. The Qwen3-VL-8B result repeats Table 2.

Backbone	$\widehat{p}_{\text{context}}$	$\widehat{p}_{\text{geometric}}$	SR ↑	SPL ↑
Qwen3-VL-8B	0.27	0.09	81.65	58.94
Qwen3-VL-32B-Instruct	0.30	0.09	84.89	62.31
InternVL3.5-8B	0.25	0.10	80.58	57.09
GPT-5.6 Luna (max)	0.34	0.08	**86.33**	**64.72**

Backbones with stronger visual understanding achieve higher SR/SPL and higher GSF discovery rates on development data, indicating more effective observation candidates from better cue interpretation. Consistent performance across the Qwen, InternVL, and GPT families supports the transferability of adaptive regional indexing and evidence-guided search across VLM backbones.

H SUPPLEMENTARY HM3DV2 OBJECTNAV RESULTS

We additionally evaluate TreeNav on HM3Dv2 category-level ObjectNav using Habitat-Matterport 3D Semantics (Yadav et al., 2023), with 36 scenes, 1,000 episodes, 6 goal categories, and a 1 m success distance. Here each episode has one category goal, so results are reported separately from GOAT-Bench's sequential multimodal tasks. In Table 11, TreeNav achieves 72.70% SR and 34.89% SPL; published methods retain their reported configurations as system-level references.

Table 11: HM3Dv2 ObjectNav results. SR and SPL are percentages; bold marks the best value in each column. Published rows retain their original numerical precision.

Method	Training-Free	SR ↑	SPL ↑
CompassNav (Li et al., 2026b)	No	61.6	27.8
L3MVN (Yu et al., 2023)	Yes	36.3	15.7
VLFM (Yokoyama et al., 2024)	Yes	63.6	32.5
SG-Nav (Yin et al., 2024)	Yes	49.6	25.5
EvoMemNav (Ge et al., 2026)	Yes	63.8	**39.4**
STEGNav (Chen et al., 2026b)	Yes	69.4	28.2
TreeNav (Ours)	Yes	**72.70**	34.89

I FAILURE ANALYSIS

Table 12: Failure breakdown for the default Qwen3-VL-8B configuration on the 278-subtask diagnostic subset. Percentages use its 51 failed subtasks as the denominator.

Failure category	Count	Share (%)
Similar object mistaken for the target	17	33.33
Target incorrectly rejected	14	27.45
Target found, but stopping too far away	7	13.73
Search exhausted without finding the target	13	25.49
Total	51	100.00

Similar-object confusion and false rejection account for 31 failures (60.78%), making target verification the dominant failure source. The remaining cases highlight two further directions: improving stopping control after target discovery and acquiring informative observations during search.

J GSF CONSTRUCTION DETAILS

GSFs use goal-related cues from historical and current images to expand the set of locations considered for further observation; the procedures below ground these cues and turn their search guidance into reachable inspection endpoints. Historical and current GSFs use the same construction rules. Each cue is grounded using the RGB-D data, camera parameters, and acquisition pose of its source observation.

Cue identification and grounding. For each goal-conditioned four-view panorama, the VLM interface provides at most one point-anchor record and one regional-view record. A point record specifies the source view and an absolute pixel coordinate selected directly on a visible cue. The controller projects this pixel into world coordinates using the aligned depth and recorded camera transform. If the selected pixel cannot be projected, a valid depth sample is sought within a 36-pixel neighborhood; no point-grounded candidate is generated if projection remains unavailable. A regional-view record instead identifies a promising search direction. Reachable geometric frontiers are associated with the source view whose horizontal heading is closest to their bearing from the source position. The grounding branch therefore depends on both the VLM evidence and the available geometric information.

Endpoint sampling and observation. Around a point cue $\mathbf{z}_c$, we sample horizontal positions at the source-pose height using radii $\{0.85, 1.15, 1.45, 1.80, 2.25\}$ m and 45° angular spacing, giving

40 raw candidates. The expanded setting adds radius $2.75\,\mathrm{m}$ and uses 30° spacing, giving 72 candidates. Samples are snapped to navigable space and filtered for valid paths, excessive snapping displacement, and repeated visits. Feasible endpoints are ranked by navigation cost from the source pose, snapping displacement, and their distance and approach direction relative to the cue. Up to three endpoints are retained, with at least $0.7\,\mathrm{m}$ navigable separation. After arrival, the agent acquires four views at yaw offsets of 0°, 90°, 180°, and 270° relative to its acquisition orientation. Sampling angles determine endpoint positions, not camera headings; no additional cue-facing or outward-facing orientation is imposed.

Cue consolidation and scheduling. Each cue group records its source panorama and view identifiers. A regional proposal does not create another search opportunity when its source view is already represented by an existing cue group. Proposals from different observations are associated using floor, anchor or endpoint locations, and navigable distance; nearby endpoints within a group are consolidated, with at most three members retained. At each decision, each active group contributes its nearest reachable, unattempted endpoint to the shared discovery-benefit and navigation-cost ranking. These rules prevent repeated source evidence and nearby endpoints from multiplying independent inspection opportunities.