LANDER-NAV: PROGRESSIVE LANDMARK-SCHEMA GROUNDING FOR MULTI-GRANULARITY OPEN-VOCABULARY NAVIGATION

Anonymous authors
Paper under double-blind review

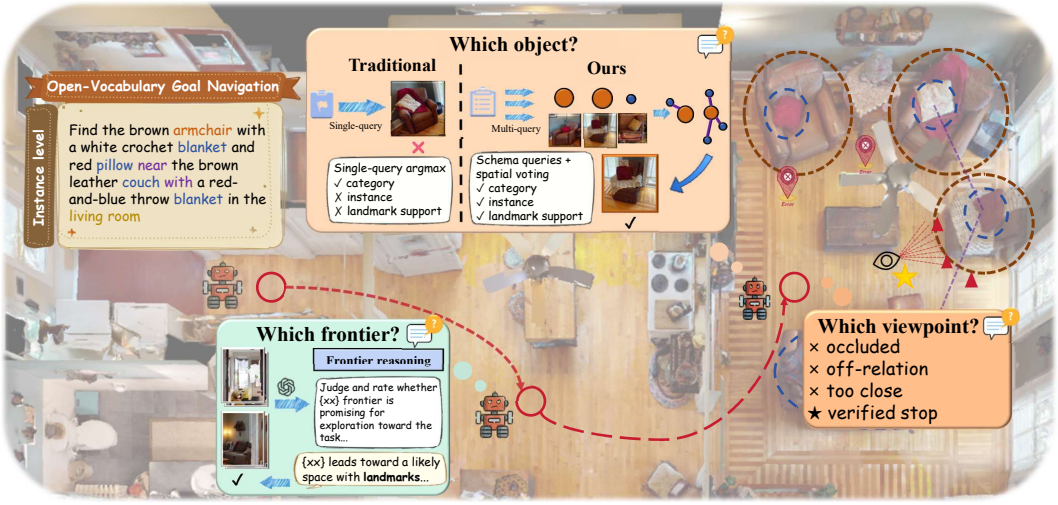


Figure 1: Overview of LANDER-Nav. Target, anchor, context, and relation cues specify the language goal. Progressive grounding connects exploration for supporting evidence, landmark-supported target binding, and selection of an executable stopping viewpoint.

ABSTRACT

Language-conditioned goal navigation is extending beyond category-level search toward multi-granularity goals disambiguated by landmarks, context, and spatial relations. Current methods encode instructions as a single joint query, conflating these constraints and often causing agents to bind the wrong same-category instance or stop at an unverifiable viewpoint. We disentangle each instruction into a landmark schema comprising target, anchor, context, and relation fields, and propose LANDER-Nav, a progressive grounding framework that preserves these role-specific constraints throughout navigation. Under this schema, LANDER-Nav coordinates three coupled decisions: Evidence Grounding decides where to explore by selecting frontiers likely to reveal missing evidence; Entity Grounding decides which entity to bind through landmark-anchored spatial voting over object memory; and Endpoint Grounding decides where to stop by choosing an executable viewpoint that verifies the required relation. Extensive experiments on LangMap and HM3D-OVON demonstrate that LANDER-Nav improves language-conditioned navigation, with particularly strong gains on long-horizon multi-goal tasks and context- and relation-dependent goals, while improving category-level open-vocabulary performance. Our code will be released.

1 INTRODUCTION

Embodied open-vocabulary navigation requires robotic agents to locate language-specified goals in unseen environments. Early benchmarks Ramakrishnan et al. (2021); Chang et al. (2017); Anderson

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.

et al. (2018); Qi et al. (2020); Ku et al. (2020); Zhu et al. (2021); Yadav et al. (2023) primarily focus on route following, predefined object categories, or single referring targets. Recent open-vocabulary goal navigation methods Gadre et al. (2023); Zhou et al. (2023); Yokoyama et al. (2024a); Padilla-Cerdio et al. (2026) relax the target space to free-form language goals. In practice, visual attributes, surrounding landmarks, scene context, and spatial relations extend goal specification beyond category names and identify the requested instance. To capture this complexity, LangMap Miao et al. (2026) formalizes the problem as multi-granularity open-vocabulary navigation across Scene, Room, Region, and Instance goals, where agents localize target objects under increasingly specific contextual constraints.

Existing open-vocabulary goal navigation methods Gadre et al. (2023); Zhou et al. (2023); Yokoyama et al. (2024a); Padilla-Cerdio et al. (2026) typically construct semantic maps or object-centric memories, and use vision-language models (VLMs) for frontier selection, candidate ranking, and semantic reasoning. These approaches improve open-vocabulary retrieval and exploration. Their decisions often compress the entire instruction into a single query and rank candidates using holistic sentence-level similarity. This design makes it challenging to explicitly disentangle and continuously reason about the target, reference landmarks, contextual cues, and spatial relations. For instance, the query "the lamp on the nightstand beside the bed" requires matching the lamp description and using the nightstand and bed constraints to identify the requested instance. Multi-granularity open-vocabulary navigation therefore calls for structured language constraints that persist across exploration, entity binding, and stop verification.

Motivated by the role of landmarks in human localization and spatial organization Epstein & Vass (2014); Epstein et al. (2017), we reinterpret landmarks as executable language roles that guide navigation decisions. From this perspective, context-dependent navigation requires agents to progressively resolve three coupled questions: First, before the target or supporting evidence is observed, *which frontier* is most likely to reveal the missing target, anchor, or contextual cues? Second, when multiple category-consistent objects coexist in memory, *which entity* satisfies the target description jointly with the required landmarks and spatial relations? Third, once an entity is bound, *which reachable viewpoint* enables the agent to observe the target under the intended target–anchor configuration? Together, these questions suggest that exploration, object binding, and stopping should be governed by a unified structured language representation.

Based on this insight, we propose LANDER-Nav, a progressive landmark-schema grounding framework for multi-granularity open-vocabulary navigation. As illustrated in Fig. 1, LANDER-Nav parses each instruction into a landmark schema that explicitly represents the target, reference landmarks, contextual regions, and spatial relations, and maintains this structure across three grounding stages. Within our framework, *Evidence Grounding* selects frontiers likely to reveal missing schema evidence; *Entity Grounding* binds the target under joint target–anchor–context–relation constraints; *Endpoint Grounding* employs Landmark-Ray Endpoint Search to identify a reachable, visible, and relation-consistent stopping viewpoint based on target–anchor geometry. Extensive experiments on LangMap and HM3D-OVON demonstrate that LANDER-Nav outperforms strong baselines, with pronounced gains on context- and relation-dependent goals while also improving category-level open-vocabulary performance. These results support preserving explicit role-specific constraints across multi-granularity navigation decisions.

Our contributions are summarized as follows:

- We introduce LANDER-Nav, which formulates multi-granularity open-vocabulary navigation as a progressive landmark-schema grounding framework, and uses a unified landmark schema to coordinate frontier selection, target binding, and stopping-viewpoint search.
- We introduce a landmark-anchored object binding mechanism based on spatial voting. By explicitly modeling reference landmarks and spatial relations, this reasoning framework evaluates and ranks candidate objects under structured constraints and retains role-level semantics.
- We demonstrate that explicit landmark-schema grounding yields consistent performance gains and strong generalization in open-vocabulary navigation. Extensive experiments on LangMap and HM3D-OVON show improved SR/SPL for multi-granularity goals, while maintaining robust performance on category-level open-vocabulary navigation.

2 Related Work

2.1 Goal-Oriented Navigation

Goal-oriented navigation (GN) requires agents to explore and localize targets from a goal specification Batra et al. (2020). Instruction-following vision-and-language navigation aligns route descriptions with navigation trajectories Anderson et al. (2018); Ku et al. (2020). Language-conditioned GN has progressed from remote referring expressions and scenario descriptions Qi et al. (2020); Zhu et al. (2021) to zero-shot open-vocabulary object goals Gadre et al. (2023); Zhou et al. (2023); Yokoyama et al. (2024b); Majumdar et al. (2022) and instance-level language goals Sun et al. (2024). GOAT-Bench extends this scope to lifelong sequences of category, instance description, and image goals, while LangMap evaluates coarse-to-fine language goals across four semantic levels Khanna et al. (2024); Miao et al. (2026). Accordingly, the central challenge has shifted from category retrieval to context-dependent target disambiguation.

GN methods broadly follow either learned end-to-end policies Ramrakhya et al. (2023); Zeng et al. (2025); Wijmans et al. (2020); Ye et al. (2021) or modular pipelines Chaplot et al. (2020b); Yu et al. (2023); Georgakis et al. (2022); Chen et al. (2023b); Ramakrishnan et al. (2022); Zhang et al. (2023); Chaplot et al. (2020a). Uni-NaVid and MTU3D learn unified navigation or grounding–exploration models from large-scale trajectories Zhang et al. (2025); Zhu et al. (2025), whereas ESC and VLFM separate semantic grounding, mapping, frontier selection, and planning Zhou et al. (2023); Yokoyama et al. (2024a). Many open-vocabulary pipelines encode language goals as category prompts or holistic relevance scores Gadre et al. (2023); Zhou et al. (2023); Yokoyama et al. (2024a). UniGoal provides the closest structured alternative, unifying category, image, and text goals through goal–scene graphs and matching-guided exploration Yin et al. (2025b). Our framework preserves target, anchor, context, and relation roles within multi-granularity language goals across evidence-guided exploration, entity binding, and relation-verifiable stopping.

2.2 Open-Vocabulary Navigation Decision

Open-vocabulary navigation increasingly uses explicit spatial representations to organize visual evidence Wu et al. (2026); Chen et al. (2023a). VLMaps grounds compositional spatial goals in visual–language maps, OVSG grounds context-dependent free-form queries over open-vocabulary 3D scene graphs, and 3D-Mem preserves multi-view observations and frontier snapshots for active reasoning Huang et al. (2023); Chang et al. (2023); Yang et al. (2025). OpenFrontier further treats visually grounded frontiers as actionable semantic anchors Padilla-Cerdio et al. (2026). These methods enrich navigation evidence through spatial maps, scene graphs, and multi-view memory. We organize frontier selection, entity binding, and stopping-viewpoint selection around explicit target, anchor, context, and relation roles.

Structured systems provide the closest precedents Pal et al. (2021). UniGoal matches goal and online scene graphs to drive subgraph search, anchor-pair alignment, and goal verification. MSGNav uses relation-bearing multimodal scene graphs for closed-loop target and frontier reasoning and selects final viewpoints based on target visibility Yin et al. (2025b); Huang et al. (2026). GC-VLN likewise decomposes route-following instructions into waypoint–object graph constraints for instruction-guided VLN Yin et al. (2025a). LANDER-Nav maintains a typed landmark schema throughout the decision loop: Evidence Grounding seeks missing schema evidence, Entity Grounding performs landmark-anchored spatial voting, and Endpoint Grounding selects a reachable, visible, and relation-verifiable viewpoint along target–anchor geometry. Our contribution is the persistent coordination of role-specific constraints across these three decisions.

3 Method

3.1 Progressive Landmark-Schema Grounding

We formulate language-conditioned navigation as progressive landmark-schema grounding: acquiring goal evidence, binding the target instance, and selecting an executable stopping viewpoint. We organize the instruction's language roles as

$$\mathcal{S}(I) = (t, \mathcal{A}, \mathcal{C}, \mathcal{R}), \tag{1}$$

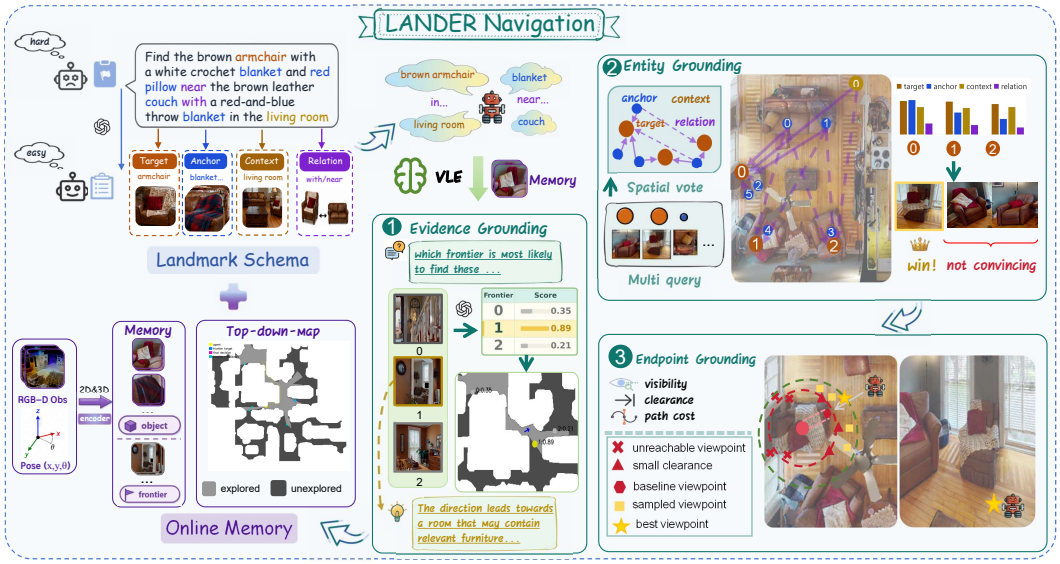


Figure 2: Progressive grounding in LANDER-Nav. Evidence Grounding evaluates current directional views for target, anchor, and contextual evidence; Entity Grounding binds the target through role-specific retrieval and local spatial voting; Endpoint Grounding converts the bound instance into a reachable stopping viewpoint with target visibility.

where t is the target description, $\mathcal{A}$ the reference landmarks, $\mathcal{C}$ the room or region context, and $\mathcal{R}$ the target–landmark relations; unspecified fields are empty.

Posed RGB-D observations update the online state $\mathcal{Z}_t = (\mathcal{O}_t, \mathcal{M}_t, \mathcal{F}_t)$: object entries, an occupancy/exploration map, and frontiers (Fig. 2). At frontier proposals, Evidence Grounding evaluates current views for target, anchor, and context evidence. At final object proposals, Entity Grounding explicitly decomposes the instruction and binds an instance through role-specific retrieval and spatial voting; Endpoint Grounding selects a navigable viewpoint for that object slot. Accumulated memory connects these event-triggered decisions from language evidence to physical approach.

3.2 Evidence Grounding: Which Frontier?

Evidence Grounding selects frontiers that are likely to reveal evidence required by the landmark schema. At each decision step, a budgeted set of candidate frontiers is selected from the backbone proposals and assessed using their already observed facing views. For each candidate j, the VLM receives the instruction, frontier metadata, and panorama observation, and returns target, anchor, room-context, negative-evidence, and confidence scores $c_j^t, c_j^a, c_j^c, c_j^-, \kappa_j$, each clipped to $[0, 1]$. We aggregate the semantic evidence as

$$h_j = \mathrm{clip}_{[0,1]}\left(w_t c_j^t + w_a c_j^a + w_c c_j^c - w_- c_j^-\right). \tag{2}$$

The nonnegative weights control the contribution of each evidence type.

Let $\bar{\ell}_j$ denote the min–max normalized backbone score over the current frontier candidates, with zero assigned when all scores coincide. For frontiers with valid VLM assessments $\mathcal{J}_{\mathrm{VLM}}$, we combine backbone preference and semantic evidence as

$$s_j = \begin{cases} \alpha\bar{\ell}_j + \beta h_j, & j \in \mathcal{J}_{\mathrm{VLM}}, \\ \bar{\ell}_j, & j \notin \mathcal{J}_{\mathrm{VLM}}. \end{cases} \tag{3}$$

Let j_b denote the original backbone proposal and $j^* = \arg\max_j s_j$. Evidence Grounding selects j^* when its VLM assessment is sufficiently confident and its fused score exceeds the backbone proposal by at least τ_{gain}; otherwise, the agent follows j_b. Candidate and VLM-call budgets restrict semantic assessment to leading or close-scoring frontiers, keeping the decision lightweight.

3.3 Entity Grounding: Which Object?

Entity Grounding identifies the intended target among competing object candidates through local configuration selection followed by target binding. At final object decisions, the instruction is decomposed into target, primary/supporting anchor, context, and relation phrases. Full-instruction retrieval preserves contextual and relational cues, while separate target and nonempty anchor queries provide role-specific object evidence.

Let $p_{q,i}$ be the softmax retrieval score for object o_i, T_q the top-retrieval set, and $\tilde{p}_{q,i} = p_{q,i}\mathbf{1}[i \in T_q]$. Candidates in $\bigcup_q T_q$ connect when $d_{ij}^{\mathrm{fp}} = \max(0, \|x_i - x_j\|_2 - r_i - r_j) \le \varepsilon_{\mathrm{cl}}$, where x_i is the planar box center and r_i a clipped half of its larger horizontal box width. Components L represent local target–support configurations.

For full, target, and available anchor queries $\mathcal{Q}_{\mathrm{req}}$, component evidence is $P_q(L) = 1 - \prod_{i \in L}(1 - \tilde{p}_{q,i})$. The landmark-consensus vote is

$$\begin{aligned} S_{\mathrm{reg}}(L) = &\sum_{q \in \mathcal{Q}_{\mathrm{req}}} w_q \log(\epsilon_{\log} + P_q(L)) + \lambda_{\mathrm{cov}} \operatorname{Cov}(L) \\ &+ \lambda_{\mathrm{cmp}} \operatorname{Cmp}(L) - D(L), \end{aligned} \tag{4}$$

where $w_q, \lambda_{\mathrm{cov}}, \lambda_{\mathrm{cmp}}$ weight evidence, coverage, and compactness, and $\epsilon_{\log} > 0$ stabilizes logarithms. $D(L)$ penalizes crowded components with sparse query coverage. Their definitions are provided in the appendix B. Selecting $L^* = \arg\max_L S_{\mathrm{reg}}(L)$ favors spatially concentrated configurations supported by complementary descriptions.

Binding is restricted to $T_t \cap L^*$. Spatial support $R(i, L^*)$ is the maximum $\exp[-(d_{ia}^{\mathrm{fp}})^2/(2\sigma_R^2)]$ over other anchor/full-query-supported members, or zero for an empty supporting set. The target score is

$$\begin{aligned} S_{\mathrm{obj}}(i) = &\eta_t \log(\epsilon_{\log} + p_{t,i}) + \eta_f \log(\epsilon_{\log} + p_{\mathrm{full},i}) + \eta_R R(i, L^*) \\ &+ \eta_m \log(\max(m_i, \epsilon_{\log})) + \eta_n \log(\max(n_i, n_{\mathrm{floor}})), \end{aligned} \tag{5}$$

where η weights each term, m_i is the merged object score, n_i the observation count, and n_{floor} its floor. Language evidence, horizontal co-localization, and observation reliability jointly rank target instances. Full-instruction evidence carries the complete description, target evidence identifies eligible objects, and anchor evidence supports configuration selection. This division preserves the identity of each retrieved role while combining its support with the surrounding objects. The maximizing candidate replaces a different backbone slot when coverage, target support, region margin, and selection gain satisfy $\tau_{\mathrm{cov}}, \tau_{\mathrm{tgt}}, \tau_{\mathrm{reg}}, \tau_{\mathrm{sel}}$. Otherwise the preceding slot is retained. Detailed gate definitions and the base reference are provided in the appendix B. Target-only binding converts configuration consensus into an instance-level decision.

3.4 Endpoint Grounding: Which Viewpoint?

Endpoint Grounding transforms the selected target instance into a geometrically valid stopping viewpoint. After Entity Grounding determines which object satisfies the instruction, navigation still requires a viewpoint from which that decision can be physically realized and visually supported. We therefore search for a stopping location that jointly preserves target observability, navigability, and sufficient spatial margin for final approach.

Given sampled world-coordinate target points P_g, let $c_g = (c_x, c_y, c_z)$ denote their centroid and p_t^y the agent height. We construct a polar set of candidate viewpoints around the target:

$$v_{km} = \big(c_x + r_k \cos\theta_m,\ p_t^y,\ c_z + r_k \sin\theta_m\big), \quad \theta_m = 2\pi m / N_\theta, \tag{6}$$

where r_k samples $[r_{\min}, r_{\max}]$ at interval Δr, and $m \in \{0, \ldots, N_\theta - 1\}$. Candidates are projected onto the navigable component when necessary and duplicate locations are removed, yielding a set of reachable viewpoints distributed around the grounded target.

We evaluate each viewpoint according to how much of the target remains observable from the corresponding camera position. For a candidate v, rays from $v + (0, h, 0)$ to valid target samples $P_g^{\mathrm{val}}(v)$ are tested against observed non-target scene points P_b:

$$V_g(v) = \frac{1}{|P_g^{\mathrm{val}}(v)|} \sum_{x \in P_g^{\mathrm{val}}(v)} \mathbf{1}\left[\min_{y \in P_b} \min_{s \in \mathrm{Ray}(v,x)} \|s - y\|_2 \ge \tau\right], \tag{7}$$

where h is the camera-height offset and τ controls occlusion tolerance. $V_g(v)$ measures the fraction of target evidence that remains geometrically visible from viewpoint v under the currently observed scene structure, linking endpoint selection to the same grounded entity established in the previous stage.

Visibility is considered together with target distance and obstacle clearance to form feasible viewpoint regions around the selected entity. These regions capture spatially coherent areas from which the target can be approached and observed under suitable geometric conditions. The feasibility criteria are progressively relaxed when the initial constraints produce insufficient support, allowing the search to adapt to cluttered or partially observed environments without discarding the target geometry established by Entity Grounding.

Among the resulting connected regions, Endpoint Grounding favors large, visually reliable regions with short geodesic access and selects an interior viewpoint with strong surrounding support. This region-level selection reduces sensitivity to isolated feasible samples and favors stopping locations that remain stable under small spatial perturbations. When connected regions remain too small, candidate-level ranking provides a complementary selection mechanism. The relaxation schedule, region ranking, and endpoint replacement criteria are detailed in the appendix B.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate LANDER-Nav on LangMap Miao et al. (2026) and HM3D-OVON Yokoyama et al. (2024b). LangMap contains over 18K tasks built from human-verified annotations, spanning 414 object categories and four goal granularities: Scene, Room, Region, and Instance. These levels progressively constrain the target from an object category to a room type, a specific region, and a unique instance; Region and Instance goals additionally provide concise and detailed descriptions, with the latter incorporating richer contextual cues and discriminative attributes. Built on HM3D-Sem, HM3D-OVON contains over 15K annotated household-object instances from 379 categories in photorealistic scans of real indoor environments.

Metrics. We report Success Rate (SR), the fraction of goals completed under the evaluation protocol, and Success weighted by Path Length (SPL), which scales each success by the ratio of geodesic shortest-path length to traveled length, capped at one. For LangMap multi-goal navigation, each subgoal is evaluated separately. Overall SR and SPL are the arithmetic means over all evaluated subgoals, with the same aggregation within each semantic level.

Implementation Details. Consistent with the simulation configurations used in prior work Yokoyama et al. (2024b); Zhu et al. (2025), we instantiate the agent with the Stretch embodiment. The robot has a height of 1.41 m and a base radius of 0.17 m, with an RGB-D camera positioned 1.31 m above the ground. At each time step, the agent receives an RGB observation at a resolution of 360×640 and selects an action from a discrete set consisting of MoveForward (0.25 m), TurnLeft and TurnRight (30°), LookUp, LookDown, and Stop. We employ Qwen-VL-Plus Qwen Team (2024) as the primary vision-language model for semantic reasoning, use DINOv2 Oquab et al. (2024) for visual feature encoding, and perform object segmentation with SAM 2 Ravi et al. (2025) using the SAM 2.1 checkpoints Meta (2024).

Frontier and object proposals come from a two-stage 3D grounding and exploration backbone: RGB-D observations yield fused object representations, and a learned matching head scores objects and frontiers against language queries. Geodesic planning and endpoint feasibility use the simulator's navmesh path finder; observed point-cloud ray tests provide visibility scores.

4.2 Quantitative Results

Table 1 reports results on LangMap across single- and multi-goal settings with concise and detailed descriptions. In multi-goal navigation, LANDER-Nav achieves the best overall performance under both description settings. With detailed descriptions, it reaches 45.5% SR and 28.6% SPL, exceeding MTU3D by 3.7 and 3.5 points. The largest SR gains appear on Region and Instance goals, where

Table 1: Evaluation on LangMap across Scene, Room, Region, and Instance goals. We report SR and SPL for single-goal and multi-goal navigation with concise and detailed descriptions. The published single-goal concise baseline scores for PSL, SenseAct-M, 3D-Mem, Uni-NaVid, and MTU3D are from Miao et al. (2026).

Setting	Method	Overall		scene		room		region		instance	
		SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Single-goal Concise	PSL Sun et al. (2024)	6.6	1.8	6.0	1.4	6.6	1.9	7.3	2.1	6.4	1.9
	SenseAct-M Khanna et al. (2024)	8.7	4.6	10.2	5.6	8.3	4.4	8.5	4.3	7.7	4.0
	3D-Mem-3B Yang et al. (2025)	15.3	2.8	20.7	3.7	20.3	4.0	13.4	2.3	10.2	1.8
	3D-Mem-7B Yang et al. (2025)	21.2	8.7	21.3	8.4	18.8	8.1	21.7	8.8	22.7	9.2
	Uni-NaVid Zhang et al. (2025)	30.3	15.3	33.8	16.2	33.2	16.5	30.1	15.5	**26.2**	**13.8**
	MTU3D Zhu et al. (2025)	29.7	**15.4**	33.1	**16.7**	31.4	15.9	**32.7**	16.5	23.8	13.3
	PlaNaVid Miao et al. (2026)	**31.4**	15.2	**34.9**	16.5	**35.6**	**16.7**	31.6	15.5	**26.2**	13.1
	LANDER-Nav	28.6	15.2	32.3	16.2	31.0	15.0	29.5	**17.3**	23.9	13.1
Single-goal Detailed	MTU3D Zhu et al. (2025)	30.4	15.8	**32.4**	**17.0**	**31.2**	**15.7**	31.6	16.2	27.5	14.9
	LANDER-Nav	**31.1**	**16.2**	32.0	16.8	30.4	14.5	**33.0**	**17.9**	**29.3**	**15.6**
Multi-goal Concise	MTU3D Zhu et al. (2025)	40.4	24.4	51.3	33.8	48.5	28.6	36.9	21.3	24.2	13.7
	LANDER-Nav	**41.5**	**26.4**	**52.7**	**36.9**	**48.8**	**29.1**	**38.4**	**23.7**	**25.5**	**15.2**
Multi-goal Detailed	MTU3D Zhu et al. (2025)	41.8	25.1	52.6	34.5	48.3	27.8	38.1	23.4	27.8	14.3
	LANDER-Nav	**45.5**	**28.6**	**55.5**	**37.5**	**49.6**	**31.1**	**43.2**	**26.8**	**33.3**	**18.9**

Table 2: Category-level open-vocabulary navigation on HM3D-OVON. SR and SPL are percentages across the seen, seen-synonym, and unseen evaluation splits.

Method	Val Seen		Val Seen Synonyms		Val Unseen		Overall	
	SR	SPL	SR	SPL	SR	SPL	SR	SPL
Uni-NaVid Zhang et al. (2025)	41.3	21.1	43.9	**21.8**	39.5	**19.8**	41.6	**20.9**
VLFM Yokoyama et al. (2024a)	35.2	18.6	32.4	17.3	35.2	19.6	34.3	18.5
MTU3D Zhu et al. (2025)	55.0	**23.6**	45.0	14.7	40.8	12.1	46.9	16.8
LANDER-Nav	**56.7**	21.1	**46.7**	14.5	**41.7**	14.1	**48.3**	16.6

contextual landmarks, discriminative attributes, and spatial relations provide richer evidence for fine-grained target grounding. Progressive landmark-schema grounding organizes these cues through target and anchor retrieval within local object configurations.

With concise multi-goal descriptions, LANDER-Nav improves both SR and SPL over MTU3D across all four goal granularities. Observations accumulated across successive subgoals enrich object memory and support evidence acquisition, entity selection, and physical approach throughout the episode. In single-goal navigation, LANDER-Nav reaches 28.6% SR and 15.2% SPL with concise descriptions, while PlaNaVid and MTU3D achieve the highest overall SR and SPL, respectively. With detailed descriptions, LANDER-Nav improves to 31.1% SR and 16.2% SPL, outperforming MTU3D and achieving the best overall performance. The strongest advantage therefore emerges in detailed multi-goal navigation, where contextual and relational cues combine with observations accumulated across subgoals.

On HM3D-OVON, Table 2 shows that LANDER-Nav achieves the highest overall SR of 48.3%, surpassing MTU3D by 1.4 points, with an overall SPL of 16.6%. These results extend the framework to category-level open-vocabulary navigation, where contextual landmarks support disambiguation when available, while target evidence and endpoint selection support category-level target search.

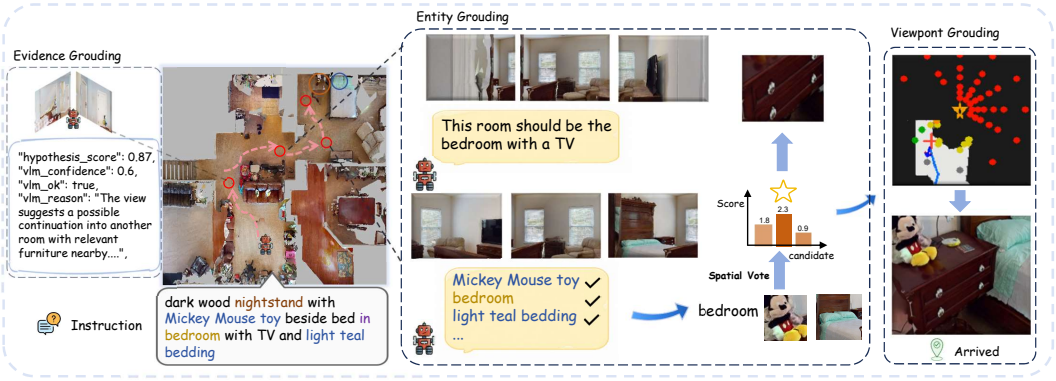


Figure 3: Progressive grounding for a LangMap Instance goal. Exploration accumulates bedroom and landmark evidence; role-specific retrieval and spatial voting bind the requested dark-wood nightstand; endpoint selection provides a navigable viewpoint with the target visible beside the bed.

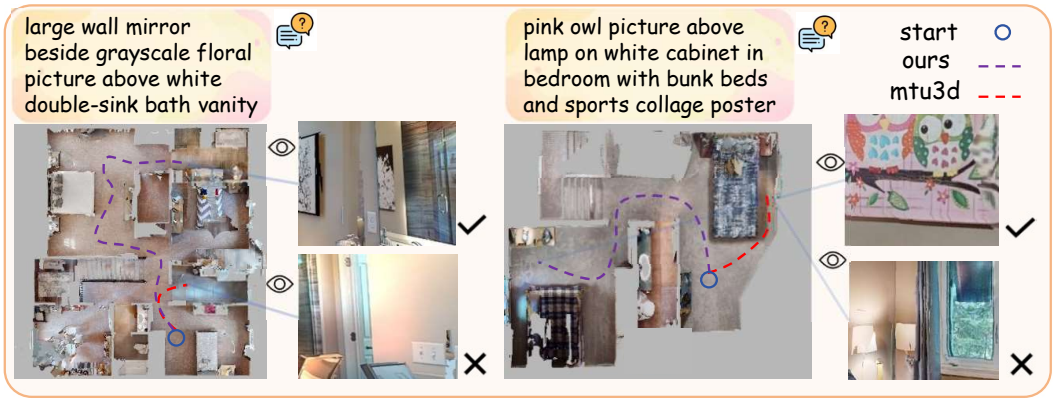


Figure 4: Trajectory and terminal-view comparisons for the wall-mirror and owl-picture LangMap goals requiring contextual instance disambiguation. Purple and red dashed paths denote LANDER-Nav and MTU3D, respectively.

4.3 Qualitative Analysis

Figure 3 illustrates grounding for a LangMap Instance goal. Exploration reveals a bedroom with a television, a Mickey Mouse toy, a bed, and light teal bedding, placing the requested dark-wood nightstand in a distinctive local configuration. Evidence Grounding assesses directional views; Entity Grounding combines nightstand, toy, and bed retrievals through spatial voting; Endpoint Grounding selects a navigable viewpoint with the bound nightstand visible. The final view connects these contextual observations with the selected instance.

Figure 4 compares trajectories and terminal views for a wall mirror beside a grayscale floral picture and an owl picture above a lamp on a white cabinet. In both examples, LANDER-Nav reaches the specified instance, while MTU3D terminates at another candidate. Nearby objects and room details distinguish plausible alternatives, illustrating contextual binding followed by physical approach.

4.4 Ablation Analysis

4.4.1 Component Ablation

As shown in Table 3, the three grounding stages exhibit distinct functional gains. Entity Grounding yields the strongest standalone SR, particularly on Region goals, where local object configurations provide critical cues for resolving target ambiguity. Target–anchor retrieval and spatial voting

Table 3: Component ablation on LangMap. Each row enables one grounding component or the complete framework. SR and SPL are percentages.

Components			Overall (SR/SPL)	scene (SR/SPL)	room (SR/SPL)	region (SR/SPL)	instance (SR/SPL)
Endpoint	Evidence	Entity					
✓			43.6 / 27.9	55.4 / 36.0	49.6 / 30.7	41.5 / **29.1**	27.2 / 14.7
	✓		42.7 / 26.3	53.3 / 35.6	49.4 / 29.3	39.1 / 24.8	28.6 / 15.2
		✓	44.4 / 27.3	52.0 / 32.7	47.7 / 29.2	**44.0** / 28.5	33.2 / 18.0
✓	✓	✓	**45.5 / 28.6**	**55.5 / 37.5**	49.6 / **31.1**	43.2 / 26.8	**33.3 / 18.9**

Table 4: VLM backbone comparison on a 100-task subset of the LangMap detailed multi-goal evaluation.

VLM Backbone	SR	SPL
`qwen-vl-plus` Qwen Team (2024)	**62.0**	**43.7**
`gpt-4o-mini` OpenAI (2024)	60.0	42.0
`qwen3-vl-8b-instruct` Bai et al. (2025)	59.0	40.1
`gemini-2.5-flash` Comanici et al. (2025)	56.0	39.9

strengthen this contextual binding. Endpoint Grounding achieves the strongest standalone SPL by grounding termination in target visibility, navigability, and spatial clearance. Evidence Grounding supports semantic exploration by prioritizing frontier views that contain target-, anchor-, and context-relevant evidence. Their combination achieves the strongest overall and Instance-level performance, forming a progressive chain from evidence acquisition to entity binding and physical approach.

4.4.2 VLM Backbone Comparison

Table 4 evaluates the framework across four VLM backbones. Qwen-VL-Plus achieves the strongest performance, followed closely by GPT-4o-mini, while Qwen3-VL-8B-Instruct and Gemini-2.5-Flash also maintain competitive navigation performance. In our framework, the VLM provides directional evidence assessment and instruction-role decomposition, so the quality of these semantic judgments directly affects subsequent evidence acquisition and entity selection. Stronger semantic grounding provides more informative cues for frontier evaluation and target disambiguation, which is reflected in the downstream SR and SPL differences. The results further show that the Evidence–Entity–Endpoint grounding process can be instantiated with different VLM backbones while preserving its structured decision flow.

5 Conclusion

We presented LANDER-Nav, a progressive landmark-schema grounding framework for multi-granularity open-vocabulary navigation. Distinguishing the target from supporting landmarks and contextual cues connects evidence-guided exploration, landmark-anchored binding, and executable endpoint selection. LangMap results demonstrate improved multi-goal navigation, especially for context-dependent Region and Instance goals, while HM3D-OVON extends the framework to category-level search. These findings support connecting role-specific language evidence with spatial navigation decisions.

AI use statement

Generative AI tools were used to assist with manuscript revision, LaTeX formatting, and literature retrieval and discovery. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final text, citations, methodological claims, and experimental results.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv preprint arXiv:2006.13171*, 2020. doi: 10.48550/arXiv.2006.13171. URL https://arxiv.org/abs/2006.13171.

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. In *Proceedings of the International Conference on 3D Vision*, pp. 667–676, 2017. doi: 10.1109/3DV.2017.00081. URL https://arxiv.org/abs/1709.06158.

Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Pu Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, and Abdeslam Boularias. Context-aware entity grounding with open-vocabulary 3d scene graphs. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1950–1974. PMLR, 2023. URL https://proceedings.mlr.press/v229/chang23b.html.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning To Explore Using Active Neural SLAM. In *International Conference on Learning Representations*, 2020a. URL https://iclr.cc/virtual_2020/poster_HklXn1BKDH.html.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4247–4258, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary Queryable Scene Representations for Real World Planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11509–11522. IEEE, 2023a. doi: 10.1109/ICRA48891.2023.10161534. URL https://doi.org/10.1109/ICRA48891.2023.10161534.

Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. How To Not Train Your Dragon: Training-free Embodied Object Goal Navigation with Semantic Frontiers. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023b. doi: 10.15607/RSS.2023.XIX.075. URL https://www.roboticsproceedings.org/rss19/p075.html.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. doi: 10.48550/arXiv.2507.06261.

Russell A. Epstein and Lindsay K. Vass. Neural systems for landmark-based wayfinding in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1635):20120533, 2014. doi: 10.1098/rstb.2012.0533.

Russell A. Epstein, Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017. doi: 10.1038/nn.4656.

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

Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to Map for Active Semantic Goal Navigation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=swrMQttr6wN.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the 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*, pp. 37154–37163, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_MSGNav_Unleashing_the_Power_of_Multi-modal_3D_Scene_Graph_for_CVPR_2026_paper.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.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352, 2022. doi: 10.52202/068431-2343. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d0b8f0c8f79d3a621af945cafb669f4b-Abstract-Conference.html.

Meta. SAM 2.1 developer suite. Software release, 2024. URL https://github.com/facebookresearch/sam2. Released September 30, 2024.

Bo Miao, Weijia Liu, Jun Luo, Lachlan Shinnick, Jian Liu, Thomas Hamilton-Smith, Yuhe Yang, Zijie Wu, Vanja Videnovic, Feras Dayoub, and Anton van den Hengel. LangMap: A human-verified benchmark for hierarchical open-vocabulary goal navigation. *arXiv preprint arXiv:2602.02220*, 2026. URL https://arxiv.org/abs/2602.02220.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2026-07-22.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Esteban Padilla-Cerdio, Boyang Sun, Marc Pollefeys, and Hermann Blum. OpenFrontier: General navigation with visual-language grounded frontiers. *arXiv preprint arXiv:2603.05377*, 2026. URL https://arxiv.org/abs/2603.05377.

Anwesan Pal, Yiding Qiu, and Henrik Christensen. Learning hierarchical relationships for object-goal navigation. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 517–528. PMLR, 2021. URL https://proceedings.mlr.press/v155/pal21a.html.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Qwen Team. Introducing Qwen-VL. Qwen official blog, January 2024. URL https://qwenlm.github.io/blog/qwen-vl/.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3d environments for embodied ai. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18890–18900, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ramakrishnan_PONI_Potential_Functions_for_ObjectGoal_Navigation_With_Interaction-Free_Learning_CVPR_2022_paper.html.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. PIRLNav: Pretraining With Imitation and RL Finetuning for ObjectNav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17896–17906, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ramrakhya_PIRLNav_Pretraining_With_Imitation_and_RL_Finetuning_for_ObjectNav_CVPR_2023_paper.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

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

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_H1gX8C4YPr.html.

Yi Wu, Junjie An, Xiao Liu, Yiqun Zhou, Yuechen Wu, Xiaoqing Guan, Shuyang Yu, You Wang, and Guang Li. Softnav: Injecting 3d scene tokens into vlms for embodied navigation, 2026. URL https://arxiv.org/abs/2607.14586.

Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mottaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2023. https://aihabitat.org/challenge/2023/, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17294–17303, 2025.

Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary Tasks and Exploration Enable ObjectGoal Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16117–16126, 2021. doi: 10.1109/ICCV48922.2021.01581. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ye_Auxiliary_Tasks_and_Exploration_Enable_ObjectGoal_Navigation_ICCV_2021_paper.html.

Hang Yin, Haoyu Wei, Xiuwei Xu, Wenxuan Guo, Jie Zhou, and Jiwen Lu. GC-VLN: Instruction as graph constraints for training-free vision-and-language navigation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 1809–1824. PMLR, 2025a. URL https://proceedings.mlr.press/v305/yin25a.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, 2025b. 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 *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 42–48, 2024a. doi: 10.1109/ICRA57147.2024.10610712. URL https://arxiv.org/abs/2312.03275.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. HM3D-OVON: A dataset and benchmark for open-vocabulary object goal navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5543–5550, 2024b.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Frontier Semantic Exploration for Visual Target Navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4099–4105. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10161059. URL https://doi.org/10.1109/ICRA48891.2023.10161059.

Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 408–432. PMLR, 2025. URL https://proceedings.mlr.press/v270/zeng25a.html.

Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 3D-Aware Object Goal Navigation via Simultaneous Exploration and Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6672–6682, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_3D-Aware_Object_Goal_Navigation_via_Simultaneous_Exploration_and_Identification_CVPR_2023_paper.html.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2412.06224.

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 International Conference on Machine Learning*, pp. 42829–42842, 2023.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. SOON: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12689–12699, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhu_SOON_Scenario_Oriented_Object_Navigation_With_Graph-Based_Exploration_CVPR_2021_paper.html.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8120–8132, 2025. URL https://arxiv.org/abs/2507.04047.

APPENDIX

A IMPLEMENTATION DETAILS

We use SAM 2 with SAM 2.1 checkpoints for image segmentation, DINOv2-Large for visual feature extraction, and the CLIP ViT-L/14 text encoder for language embeddings. Qwen-VL-Plus is accessed through its API for visual evidence assessment and instruction decomposition. Experiments were conducted on a workstation with four NVIDIA RTX A6000 GPUs (48 GB each), using independent single-GPU workers for parallel evaluation. We use random seed 1234 for sampling and for Python, NumPy, PyTorch, and CUDA initialization. Each experimental configuration is evaluated twice, and the reported SR and SPL are the arithmetic means of the two runs. LANDER-Nav requires no additional task-specific training, and pretrained model parameters remain fixed throughout evaluation.

EVALUATION DATA AND PROTOCOLS

LangMap. LangMap provides human-verified navigation goals across 36 HM3D-Sem validation scenes. Its four semantic levels progressively specify the target: Scene goals identify an object category; Room goals additionally specify a room type; Region goals distinguish a particular room instance through contextual descriptions; and Instance goals identify a specific object through discriminative attributes and spatial relations. Region and Instance goals have concise and detailed descriptions. Concise descriptions provide the main distinguishing cues, while detailed descriptions add visual attributes, nearby objects, and spatial context. Scene and Room goal construction is shared across the two description settings. The main single-goal evaluation covers the complete available annotation set, comprising 9,420 goals across all 36 scenes, under both settings. Multi-goal evaluation uses 720 five-goal episodes drawn from LangMap, totaling 3,600 subgoals. The navigation and evaluation settings are shared with single-goal evaluation, and each subgoal is scored separately. Each single-goal task and each subgoal in a multi-goal episode has a budget of 500 actions.

HM3D-OVON. We evaluate 360 episodes across 36 validation scenes, with 120 episodes each from Val Seen, Val Seen Synonyms, and Val Unseen. These splits cover seen categories, unseen categories semantically similar to training categories, and semantically distinct unseen categories, respectively. Each episode has a budget of 500 actions. Success is determined by a terminal geodesic distance of at most 0.25 m to a valid goal viewpoint associated with the target category. We use the Stretch embodiment with a height of 1.41 m and a base radius of 0.17 m. Its RGB-D camera is mounted at 1.31 m, provides observations of 360×640 pixels (width $\times$ height), and has a horizontal field of view of 42°. The agent moves forward by 0.25 m or turns left or right by 30°.

Ablation evaluation. Component ablations use the LangMap multi-goal tasks with detailed instructions. For the VLM backbone comparison, we sample two five-goal episodes from each of ten scenes, yielding 20 episodes and 100 subgoals, and use detailed instructions throughout. These evaluations use seed 1234 and report the mean results of two runs for each configuration.

ONLINE STATE MAINTENANCE

Object entries store fused geometry and visual–language features. Their horizontal footprints support the spatial grouping used by Entity Grounding. Current panoramic views, object point clouds, the observed agent pose, and navmesh path-finder queries provide the navigation interfaces. The frontier module receives the current task text. At an object decision, the object module decomposes this text. The viewpoint module receives the selected object slot and its observed point cloud.

FRONTIER VIEW ASSOCIATION AND SCORE REFRESH

Evidence Grounding associates each frontier with an already observed panoramic view facing its direction. Let I_t^k denote the k-th current view with unit horizontal heading vector h_k, and let u_j be the unit horizontal direction from the agent to frontier f_j. When view headings are available, view

association selects

$$k_j = \arg \max_{k \in \{1,\dots,K\}} h_k^\top u_j, \qquad I_j = I_t^{k_j}. \tag{8}$$

For uniformly ordered panoramas without view headings, the relative frontier bearing determines the nearest view index by rounding. The VLM receives frontiers with valid bearings above the near-zero threshold and their associated current panoramic views.

Evidence Grounding evaluates a bounded subset of leading or ambiguous frontiers and retains the base policy's frontier when its override gate fails. The default configuration computes semantic-evidence scores without caching. Enabling the optional score cache reuses valid scores for remaining candidates. The base policy's selected frontier is denoted by f_b.

Subgoal Metrics

For each evaluated subgoal i, let $S_i \in \{0, 1\}$ denote success under the evaluation protocol, $d_i^\star$ the geodesic shortest-path length from the subgoal start to its valid goal set, and L_i the traveled length. The per-subgoal metrics are

$$\begin{aligned} \mathrm{SR}_i &= S_i, \\ \mathrm{SPL}_i &= \begin{cases} S_i \dfrac{d_i^\star}{\max(d_i^\star, L_i)}, & 0 < d_i^\star < \infty, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{9}$$

Overall multi-goal SR and SPL average these values over *subgoal rows*, with the same rule within each semantic level.

B Algorithms

The algorithms detail how Progressive Landmark-Schema Grounding selects frontiers, binds target instances, and chooses stopping viewpoints.

Online Decision Procedure

The three grounding modules act at different decisions of the base navigation policy. Evidence Grounding refines a frontier proposal using the current panorama and task text. At a final object proposal, Entity Grounding refines the selected object slot through gated role-conditioned retrieval; Endpoint Grounding then evaluates navigable viewpoints for that slot. The following procedure separates these branches and counts active-scan and follower actions against the subgoal budget B.

Evidence Grounding. Let ℓ_j be the base frontier logit, $\bar{\ell}_j$ its min–max normalization over current candidates (zero when all logits coincide), and j_b the base frontier index. A bounded assessment set includes the base choice and high-logit or close-scoring alternatives when the call interval permits a VLM query. Each assessed frontier is associated with an *already observed* panorama view facing its direction. For a valid assessment, let $c_j^t, c_j^a, c_j^c, c_j^-$ and κ_j denote the returned target, anchor, room, negative-evidence, and confidence fields. The view score and fusion are

$$h_j = \mathrm{clip}_{[0,1]} \left(w_t c_j^t + w_a c_j^a + w_c c_j^c - w_- c_j^- \right), \tag{10}$$

$$s_j = \begin{cases} \alpha \bar{\ell}_j + \beta h_j, & j \in \mathcal{J}_{\mathrm{VLM}}, \\ \bar{\ell}_j, & j \notin \mathcal{J}_{\mathrm{VLM}}, \end{cases} \tag{11}$$

where $\mathcal{J}_{\mathrm{VLM}}$ contains candidates with valid view scores; w_t, w_a, w_c, w_- and α, β are fixed scoring weights. Writing $j^* = \arg\max_j s_j$, the accepted frontier is

$$j_{\mathrm{sel}} = \begin{cases} j^*, & j^* \neq j_b,\ j^* \in \mathcal{J}_{\mathrm{VLM}},\ \kappa_{j^*} \geq \tau_{\mathrm{conf}},\ s_{j^*} - s_{j_b} \geq \tau_{\mathrm{gain}}, \\ j_b, & \text{otherwise.} \end{cases} \tag{12}$$

The confidence and gain thresholds are fixed decision parameters. The assessment budget and call interval limit VLM use; score caching is optional.

Input: instruction I, initial online state $\mathcal{Z}_0$, base policy, path finder, action budget B.
Output: executed subgoal trajectory and terminal pose.

$a \leftarrow 0$; initialize the current pose and object memory.
while $a < B$ **do**
  Observe posed RGB-D views; update $\mathcal{Z}$ and object memory.
  Charge any active-scan actions to a; stop if $a \geq B$.
  Obtain base proposal $(x_b, b_{\mathrm{obj}}, \mathrm{aux})$.
  if $b_{\mathrm{obj}} = \mathrm{false}$ **then**
    Select budgeted frontiers and associate current facing views.
    Score valid views; compute h_j, s_j and gated j_{sel}.
    $x \leftarrow f_{j_{\mathrm{sel}}}$; follow toward x.
    Charge follower actions to a; on failure, mark frontier visited.
    If the follower failed and $a < B$, retry from the next decision.
  else
    Decompose I; retrieve full, target, and available anchor roles.
    Form footprint components; rank eligible region and target hits.
    Apply the object-binding gate to obtain object decision x_{obj}.
    Evaluate viewpoints for the selected object slot and path finder.
    $x \leftarrow$ accepted viewpoint if available, else x_{obj}.
    Follow toward x; increase a by the executed actions.
    End the subgoal after this final decision or follower failure.
  end if
end while

Figure 5: Online grounding at frontier and final-object decisions. Module gates select between grounded proposals and the preceding valid decision; subgoal SR and SPL use the terminal trajectory under the run's evaluation protocol.

Entity Grounding. At a final object decision, the instruction is decomposed into target, primary and supporting anchor, room-context, and relation fields. The object queries are the full instruction, target phrase, and nonempty anchor phrases. The full-instruction query retains contextual and relational language. Given current object-memory set $\mathcal{O}_t$ and logits $z_{q,i}$ for role q, the retrieval probability and its top-k restriction are

$$p_{q,i} = \frac{\exp(z_{q,i}/T_{\mathrm{sm}})}{\sum_{u \in \mathcal{O}_t} \exp(z_{q,u}/T_{\mathrm{sm}})}, \qquad \tilde{p}_{q,i} = p_{q,i}\mathbf{1}[i \in T_q], \tag{13}$$

where T_{sm} is the softmax temperature and T_q is the role's top-retrieval set. Candidates from $\bigcup_q T_q$ are connected when their horizontal footprint distance is at most $\varepsilon_{\mathrm{cl}}$:

$$d_{ij}^{\mathrm{fp}} = \max\left(0, \|x_i - x_j\|_2 - r_i - r_j\right). \tag{14}$$

Here x_i is a planar box center and r_i its clipped horizontal half-width. For each connected component L, define role evidence, coverage, and compactness as

$$\begin{aligned} P_q(L) &= 1 - \prod_{i \in L}(1 - \tilde{p}_{q,i}), \\ \mathrm{Cov}(L) &= \frac{1}{|\mathcal{Q}_{\mathrm{req}}|} \sum_{q \in \mathcal{Q}_{\mathrm{req}}} \mathbf{1}[P_q(L) > \epsilon_{\mathrm{cov}}], \\ \mathrm{Cmp}(L) &= \exp\left[-\frac{\sum_{i \in L} \|x_i - \bar{x}_L\|_2^2}{|L|\sigma^2}\right], \quad \sigma = \max(\varepsilon_{\mathrm{cl}}, \sigma_{\min}), \end{aligned} \tag{15}$$

where $\bar{x}_L$ is the planar component center, and $\mathcal{Q}_{\mathrm{req}}$ contains the full and target queries and each available anchor query. If $n_q(L) = |L \cap T_q|$, the sparse-coverage penalty is

$$D(L) = \lambda_{\mathrm{pen}}\mathbf{1}\left[\mathrm{Cov}(L) < \tau_{\mathrm{sparse}} \ \wedge \ \max_q n_q(L) \geq n_{\mathrm{crowd}}\right]. \tag{16}$$

The component vote and selected region are

$$S_{\mathrm{reg}}(L) = \sum_{q \in \mathcal{Q}_{\mathrm{req}}} w_q \log(\epsilon_{\log} + P_q(L)) + \lambda_{\mathrm{cov}} \operatorname{Cov}(L) + \lambda_{\mathrm{cmp}} \operatorname{Cmp}(L) - D(L), \qquad L^* = \arg\max_L S_{\mathrm{reg}}(L). \tag{17}$$

The constants ϵ_{cov}, $\epsilon_{\log}$, and $\sigma_{\min}$, role weights w_q, and region weights and thresholds are fixed implementation parameters. The coverage threshold used in $D(L)$ is distinct from the later object-binding gate.

Final binding selects among target-role hits in L^*. Let $\mathcal{A}_i(L)$ contain other members of L retrieved for an anchor or full-instruction role, and define

$$R(i, L) = \begin{cases} \max\limits_{u \in \mathcal{A}_i(L)} \exp\left[-\dfrac{(d_{iu}^{\mathrm{fp}})^2}{2\sigma_R^2}\right], & \mathcal{A}_i(L) \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

The target score is

$$S_{\mathrm{obj}}(i) = \eta_t \log(\epsilon_{\log} + p_{t,i}) + \eta_f \log(\epsilon_{\log} + p_{\mathrm{full},i}) + \eta_R R(i, L^*) + \eta_m \log(\max(m_i, \epsilon_{\log})) + \eta_n \log(\max(n_i, n_{\mathrm{floor}})), \tag{19}$$

where m_i is the merged object score, n_i its observation count, and n_{floor} the count floor. Let i^* maximize this score over target-role hits in L^*, and let i_b be the base object slot. The binding gate compares the proposed score to a base reference containing its target, full-query, and spatial-support terms when the base slot is valid:

$$S_b(i_b) = \eta_t \log(\epsilon_{\log} + p_{t,i_b}) + \eta_f \log(\epsilon_{\log} + p_{\mathrm{full},i_b}) + \eta_R R(i_b, L^*). \tag{20}$$

The proposed binding is accepted when $i^* \neq i_b$, $\operatorname{Cov}(L^*) \geq \tau_{\mathrm{cov}}$, $p_{t,i^*} \geq \tau_{\mathrm{tgt}}$, the margin over the second region is at least τ_{reg}, and $S_{\mathrm{obj}}(i^*) - S_b(i_b) \geq \tau_{\mathrm{sel}}$ for a valid base slot. The memory score and observation count weight proposed targets; S_b comprises target, full-query, and spatial-support evidence. Full-instruction retrieval retains contextual and relational language. The support term R measures horizontal co-localization with retrieved supporting objects. For a single region, the region-margin condition is satisfied.

Endpoint Grounding. Given the selected object's observed point cloud, sample up to N_g target points P_g and let $c_g = (c_x, c_y, c_z)$ be their mean. With the agent's current height p_t^y, generate polar-grid locations

$$v_{km} = (c_x + r_k \cos\theta_m,\ p_t^y,\ c_z + r_k \sin\theta_m), \qquad \theta_m = 2\pi m / N_\theta, \quad r_k \in [r_{\min}, r_{\max}], \tag{21}$$

with radial spacing Δr. Keep points that are navigable in the agent's connected navmesh component, directly or after a bounded snap; remove duplicate locations. Assess the current object decision as an incumbent. Form polar-grid components from the generated grid locations.

For candidate v, cast discrete rays from $v + (0, h, 0)$ to valid target points, and let P_b be observed non-target scene points. Estimated target visibility is

$$V_g(v) = \frac{1}{|P_g^{\mathrm{val}}(v)|} \sum_{x \in P_g^{\mathrm{val}}(v)} \mathbf{1}\left[\min_{y \in P_b} \min_{s \in \mathrm{Ray}(v,x)} \|s - y\|_2 \geq \tau\right], \tag{22}$$

where h is the camera-height offset and τ the occlusion tolerance. The score is zero if no valid ray exists; for empty P_b, the inner minimum is $+\infty$. Let $d_g(v)$ be minimum planar distance to sampled target points, $d_b(v)$ minimum planar clearance from observed non-target points, and $d_{\mathrm{geo}}(v)$ the path-finder geodesic distance from the agent. At relaxation stage ℓ, the feasible grid locations are

$$\mathcal{V}_\ell = \left\{v : V_g(v) \geq \nu_\ell, \quad \rho_\ell^- \leq d_g(v) \leq \rho_\ell^+, \quad \left(\chi_\ell = 0 \ \vee\ d_b(v) \geq \tau_b\right)\right\}, \tag{23}$$

where χ_ℓ indicates whether clearance is required. The ordered stages retain the initial visibility, shell, and clearance bounds, then relax visibility, shell, and clearance in that order.

Feasible locations are adjacent when their radial and cyclic angular indices differ by at most one, excluding the identical index pair, or when their planar distance is at most $\lambda_{\mathrm{adj}} \max(\Delta r, \epsilon_{\mathrm{adj}})$. Here λ_{adj} and ϵ_{adj} set the proximity rule used after navmesh snapping; angular indices wrap around the grid. These graph edges form connected components. At the first stage with a component of at least $n_{\min}$ points, let M_ℓ be the size of the largest such component and retain only components with $|C| \geq \lceil \zeta M_\ell \rceil$. Their ranking is lexicographic:

$$C^* = \arg\min_{C} \left(-Q_\xi\{V_g(v) : v \in C\},\ \bar{d}^{\mathrm{fin}}_{\mathrm{geo}}(C),\ -|C|,\ \mathrm{id}(C) \right), \tag{24}$$

where the minimization ranges over the retained components and Q_ξ is a fixed visibility quantile. Here $\bar{d}^{\mathrm{fin}}_{\mathrm{geo}}(C)$ averages the finite geodesic distances in C, or equals $+\infty$ if none is finite. Within C^*, mark a location as boundary when one of its eight neighboring grid indices is absent from C^* or outside the radial range. If no such location exists, use all members as boundary. Let $d_{\partial C^*}(v)$ be shortest graph distance to this boundary. The chosen viewpoint is

$$v^* = \arg\min_{v \in C^*} \left(-d_{\partial C^*}(v),\ -V_g(v),\ d_{\mathrm{geo}}(v),\ \mathrm{order}(v) \right), \tag{25}$$

again in lexicographic order. This selects a viewpoint interior to a feasible connected region, with observed visibility and path length resolving ties. An endpoint is applied only when its selected viewpoint has positive estimated visibility. When component selection is unavailable, visibility ranks the reachable candidates for endpoint acceptance.

C Additional Experiments

Backbone Comparison

We compare LANDER-Nav segmentation configurations on the same paired subset of LangMap multi-goal tasks. Table 5 reports SR and SPL as percentages under the concise and detailed instruction settings.

Table 5: Segmentation-backbone comparison on a paired subset of LangMap multi-goal tasks. SR and SPL are percentages.

Configuration	Instruction form	SR (%)	SPL (%)	Avg. steps
Ours + FastSAM	Detailed	37.11	19.88	197.1
Ours + FastSAM	Concise	33.76	18.49	187.7
Ours + SAM	Detailed	39.27	21.03	196.8
Ours + SAM	Concise	37.11	20.06	188.3
Ours + SAM 2.1	Detailed	40.75	21.67	199.7

Among the detailed LANDER-Nav variants, SAM 2.1 attains the highest SR and SPL in this comparison. This result motivates its use in the main configuration. Object segmentation supplies the object-centric entries ranked by Entity Grounding and the target geometry evaluated by Endpoint Grounding.

D Prompt Templates and Text Queries

This section presents the two VLM instruction templates and their query interface at the respective navigation decisions. Angle-bracketed fields denote template substitutions. The navigation instruction is built from the current Scene, Room, Region, or Instance goal. The first template is sent with a current panoramic image; the second is text-only.

Evidence Grounding: frontier-facing view assessment. The complete instruction template is:

```
You evaluate one CURRENT robot panorama view for test-time frontier
reranking.

The image is already observed at the current decision step; it is not
a rendered future view.

Navigation task: <NAVIGATION_INSTRUCTION>

Frontier/view metadata: <FRONTIER_VIEW_METADATA>

Judge whether moving/exploring in this viewed direction is promising
for reaching the task target.

Do not require the target object to already be visible. Reward
concrete visual cues that make this direction task-facing: likely
target category, named anchor objects, semantically compatible
furniture, room context, signs of traversable continuation, and
spatial layout that plausibly leads toward the target. Penalize dead
ends, blank walls, blocked/narrow views, unrelated rooms, clutter that
prevents progress, or cues that contradict the task. Keep scores
calibrated: 0.0=no support, 0.5=weak/ambiguous support, 1.0=strong
support.

Return strict JSON only with this exact schema:

{"target_context": number between 0 and 1, "anchor_context": number
between 0 and 1, "room_context": number between 0 and 1, "negative":
number between 0 and 1, "confidence": number between 0 and 1,
"reason": "one short sentence"}
```

The frontier/view metadata is a comma-separated subset of `frontier_index`, `frontier_xyz`, `relative_bearing_rad`, and `current_panorama_view_index`; if none is available, its literal value is `not provided`. The scored fields yield the semantic evidence in Eq. 10, which is combined with base frontier logits by Eq. 11.

Entity Grounding: instruction decomposition. The complete text-only instruction template is:

```
You decompose a RefHM3D navigation instruction for multi-query 3D
object grounding.

The instruction specifies full-task semantics, the target, anchors,
and room context for downstream grounding.

Return strict JSON only with this exact schema:

{"target_desc": string, "target_aliases": [string], "anchor_primary":
[string], "anchor_support": [string], "room_context": [string],
"relations": [string]}

Definitions and hard rules:

1. target_desc is the primary object to navigate to, not a location
or anchor.

2. target_aliases are short alternative names for the same target
object; keep 0-3.

3. anchor_primary are the strongest nearby objects/furniture/fixtures
that identify the target area; keep 0-2.

4. anchor_support are other useful nearby concrete objects or visual
area cues; keep 0-4.

5. room_context contains room or region labels only when explicit or
strongly implied; keep 0-2.
```

```
6. Do not put the target object itself into anchor_primary,
anchor_support, or room_context.

7. Avoid generic structural anchors such as wall, floor, ceiling,
door, window unless modified into a distinctive object phrase.

8. Use lowercase concise noun phrases. No instructions, no
sentences, no markdown.

task_type: <TASK_TYPE>

navigation_instruction: <NAVIGATION_INSTRUCTION>
```

The object-grounding model uses the complete instruction, target phrase and aliases, and nonempty primary and supporting anchor phrases as text queries. The text queries and contextual fields are:

```
Text queries:

full: <NAVIGATION_INSTRUCTION>

target: <TARGET_DESC>[, <TARGET_ALIAS> ...]

anchor_primary: <PRIMARY_ANCHOR>[, <PRIMARY_ANCHOR> ...]

anchor_support: <SUPPORT_ANCHOR>[, <SUPPORT_ANCHOR> ...]

Contextual fields:

room_context: [<ROOM_OR_REGION>, ...]

relations: [<SPATIAL_RELATION>, ...]
```

Within each role, nonempty phrases are deduplicated and joined by a comma and a space. Each populated anchor role contributes a query. The full-instruction query carries room and relational context into object grounding, and the role queries specify the target and supporting objects. Endpoint Grounding uses observed point geometry and path finding to evaluate navigable target viewpoints.

E Additional Qualitative Analysis

Figure 6 shows LANDER-Nav grounding two context-dependent LangMap goals. In the first example, LANDER-Nav uses the bed and Mickey Mouse plush toy as contextual cues to ground the intended wardrobe; in the second, it combines the TV cabinet and colorful box cues to identify the target sculpture. Contextual landmarks guide fine-grained instance grounding and navigation toward the requested targets.



Figure 6: Additional qualitative results on LangMap. Each row shows the language goal, the corresponding top-down navigation trajectory, and the target observation and navigation map at the terminal step.