RATIO: A REASONING AGENT FOR TASK-ADAPTIVE INTERACTION ORCHESTRATION IN ZERO-SHOT OBJECT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot object navigation requires sequential decisions under partial observations. Existing methods increasingly use large language models (LLMs) for high-level reasoning, yet navigation capabilities are still commonly organized by predefined routines, while precise spatial computation is often mixed with semantic decision making. We introduce **RATIO**, a **R**easoning **A**gent for **T**ask-Adaptive **I**nteraction **O**rchestration for zero-shot object navigation. RATIO organizes geometric, semantic, and navigational information into an agent-facing environment representation and an executable navigation graph that the agent can query on demand, and exposes specialized navigation capabilities as tools. At each decision step, the agent selects tools based on the current task context, while dedicated modules perform geometric, retrieval, and candidate computations. By separating capability selection from execution, RATIO enables context-dependent navigation while preserving precise and reproducible spatial computation. We evaluate RATIO on HM3Dv1, HM3Dv2, MP3D, and HM3D-OVON Unseen, achieving a 53.0% success rate and 34.0% success weighted by path length on the open-vocabulary benchmark with GPT-5.6 Luna.

1 INTRODUCTION

Zero-shot object navigation requires an agent to search for a target instance in an unseen environment using only the target category, real-time RGB and depth (RGB-D) observations, and its own pose, and to stop near the target (Batra et al., 2020). As vision-language models (VLMs) and large language models (LLMs) have demonstrated open-vocabulary recognition, commonsense reasoning, and long-horizon planning, navigation systems have increasingly used scene semantics and commonsense knowledge to search for targets in unknown environments (Gadre et al., 2023; Yokoyama et al., 2024a; Yin et al., 2024). The role of VLMs and LLMs in navigation has expanded from semantic matching to map understanding, high-level planning, state selection, and the use of historical experience, giving them a greater role in decision making.

However, existing methods typically follow predefined navigation procedures. ApexNav (Zhang et al., 2025) adjusts exploration according to semantic evidence, CogNav (Cao et al., 2025) uses an LLM to select predefined cognitive states, and TrajRAG (Wang et al., 2026) retrieves historical trajectories to assist candidate selection. Although LLMs and VLMs can participate in high-level decisions, the triggering and combination of capabilities such as relation matching, experience retrieval, target candidate search, and exploration recommendation are usually specified by the system in advance. Models rarely decide which navigation capability to use at a given moment based on the current observation and task evidence.

Giving an LLM decision authority does not make it responsible for every navigation computation. VLMs and LLMs interpret target semantics, scene context, and multiple information sources (Chen et al., 2024a; Zitkovich et al., 2023), whereas coordinate transformations and distance calculations require precise spatial computation. LLM spatial plans can also conflict with geometric constraints. Recent methods separate these roles by letting a VLM select candidates in visual space and using depth and camera geometry to recover their three-dimensional positions (Chen et al., 2025; Peng et al., 2026). Another method geometrically projects a VLM-selected visual target and navigates

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.

with a low-level simultaneous localization and mapping (SLAM) controller (Feng et al., 2026). Accordingly, the agent selects needed information and capabilities, while specialized modules perform geometric, retrieval, and statistical computations and return structured results.

For the agent to select capabilities at each decision point, it must be able to query the evidence needed for that choice. Existing representations distribute traversability, exploration uncertainty, task-related semantic evidence, and reachable observation positions across separate structures, leaving the agent to assemble them.

We introduce RATIO, a **R**easoning **A**gent for **T**ask-Adaptive **I**nteraction **O**rchestration. RATIO organizes this evidence into an agent-facing environment representation and an executable navigation graph with queryable nodes, semantic regions conditioned on the task, and edges with geometric and semantic attributes. Its navigation capabilities cover relation matching, experience retrieval, target candidate search, and exploration recommendation. At each high-level decision point, the LLM selects tools using the target category, robot state, executable navigation graph, and tool feedback history. Specialized modules perform the selected computations, and the LLM integrates their results to choose the next executable navigation node or stop.

We evaluate RATIO on HM3Dv1 (Ramakrishnan et al., 2021), HM3Dv2 (Yadav et al., 2023), MP3D (Chang et al., 2017), and HM3D-OVON Unseen (Yokoyama et al., 2024b) , comparing it with representative zero-shot object navigation methods. Four progressive ablation settings start from a shared navigation backbone and successively add the executable navigation graph, direct LLM decision making, and the full RATIO system to analyze the roles of the graph, high-level LLM decisions, and tool interaction. We also evaluate adaptation to different LLMs and open-vocabulary target settings.

Our main contributions are as follows. **(1)** We introduce an agent decision paradigm for zero-shot object navigation in which the LLM selects navigation capabilities based on the current task context, while specialized tools perform geometric, retrieval, and candidate computations. This separates capability selection from execution. **(2)** We construct an agent-facing environment representation and an executable navigation graph that make task evidence, geometric uncertainty, reachable observation positions, and spatial relations queryable through regions, nodes, and edges, grounding agent decisions in executable navigation goals.

2 Related Work

2.1 Zero-Shot Object Navigation

Zero-shot object navigation uses the open-vocabulary knowledge of pretrained VLMs and LLMs to search unknown environments without retraining for target categories. Early methods paired open-vocabulary perception with exploration. CoW (Gadre et al., 2023) combines VLM target localization with geometric exploration, ESC (Zhou et al., 2023) uses soft commonsense constraints, and VLFM (Yokoyama et al., 2024a) builds a frontier value map from vision-language similarity. L3MVN (Yu et al., 2023), SG-Nav (Yin et al., 2024), and VoroNav (Wu et al., 2024) then brought target commonsense, scene relations, and topology into candidate selection. Recent methods adjust search strategies. ApexNav (Zhang et al., 2025) coordinates geometric and semantic exploration using semantic evidence, UniGoal (Yin et al., 2025b) changes strategies according to matching states between scene and goal graphs, CogNav (Cao et al., 2025) lets an LLM choose predefined cognitive states, and TrajRAG (Wang et al., 2026) retrieves past trajectories to assess candidates. These methods expand the role of VLMs and LLMs in high-level navigation, but still rely on predefined states or triggers to combine navigation capabilities. RATIO shifts capability selection to the LLM, which chooses tools based on the current task context.

2.2 Environment Representation

Persistent environment representations let navigation decisions use information beyond the current view. Semantic maps accumulate task-related evidence (Chaplot et al., 2020), while later work incorporates semantic frontiers (Chen et al., 2023), geometric affordances (Huang et al., 2024), and open-set semantics (Kuang et al., 2024). Belief maps explicitly represent occupancy uncertainty (Zhou et al., 2025). Language-aligned representations organize environmental semantics in implicit lan-

guage maps (Lee et al., 2025) or maintain instance-level object representations for open-vocabulary navigation (Tang et al., 2025). Methods without explicit maps learn latent representations through policy networks (Ramakrishnan et al., 2022; Majumdar et al., 2022) or predict future observations with world models (Bar et al., 2025). For long-horizon navigation, map caching and context compression help control storage and query costs (Yang et al., 2025; Hu et al., 2026).

These representations often store information in dense grids, continuous confidence values, or implicit vectors that must be serialized into prompts for LLM decisions under a limited input budget (Zhou et al., 2024; Chen et al., 2024b). The agent must also gather node traversability, unexplored space around a region, relevance to the current target, and observation positions from separate geometric maps, semantic maps, and retrieval memory. RATIO organizes this information in the same set of executable nodes, semantic regions conditioned on the task, and navigation edges with geometric and semantic attributes for direct querying during agent decisions.

2.3 LLM-Driven Navigation Decisions

ReAct (Yao et al., 2023), SayCan (Ichter et al., 2023), and Inner Monologue (Huang et al., 2023) incorporate external capabilities into LLM decision loops. In navigation, CogNav (Cao et al., 2025) uses an LLM to switch among predefined cognitive states, while Hierarchical Fast–Slow ReAct Agent (Lan et al., 2026) combines reactive control with reasoning and retrieval triggered by events. AgenticNav (Li et al., 2026) explicitly exposes actions, depth, and visual memory as callable tools in continuous vision-language navigation. AgenticNav shows that tool interfaces can connect an LLM with embodied execution, while its interface focuses on perception, memory, and action for instruction following.

Another line of work separates semantic decisions from precise spatial computation. AO-Planner (Chen et al., 2025) lets a VLM select candidate paths in visual space, then uses depth and camera geometry to recover three-dimensional coordinates. Less Is More (Peng et al., 2026) pairs detections from a vision model with VLM reasoning to improve scene understanding, then maps VLM-selected object IDs to coordinates through a deterministic lookup. GC-VLN (Yin et al., 2025a) solves graph constraints derived from linguistic spatial relations to obtain waypoints and paths. Embodied-Navigator (Feng et al., 2026) projects VLM-selected image locations into three dimensions for execution by a SLAM controller. RATIO extends this division of labor by letting the LLM choose navigation capabilities from context while specialized tools perform geometric, retrieval, and candidate computations.

3 Method

3.1 Problem Formulation and Overview

We study zero-shot object navigation without task-specific policy training. Given target category text g, the agent uses real-time RGB-D observations and an odometry pose to search an unknown environment through discrete actions within a step limit, stopping within the success distance of a target instance (Batra et al., 2020). As shown in Fig. 1, RATIO comprises an agent-facing environment representation, an executable navigation graph, and navigation decision making. The representation organizes decision evidence for on-demand queries.

3.2 Agent-Facing Environment Representation

The agent-facing environment representation $\mathcal{M}_t^A = \{\mathcal{M}_t^G, \mathcal{M}_t^{S,A}\}$ comprises a geometric occupancy grid map $\mathcal{M}_t^G$ and a semantic map conditioned on the task $\mathcal{M}_t^{S,A}$. It preserves traversability, exploration uncertainty, and task-related semantic evidence.

3.2.1 Geometric Occupancy Grid Map with Four States

We convert depth images to point clouds, apply radius outlier removal, and use odometry poses to project the points onto a common map plane. Ray casting performs probabilistic occupancy fusion by updating occupied evidence at ray endpoints and free evidence along traversed cells. Each map cell uses a single-channel encoding with four states:

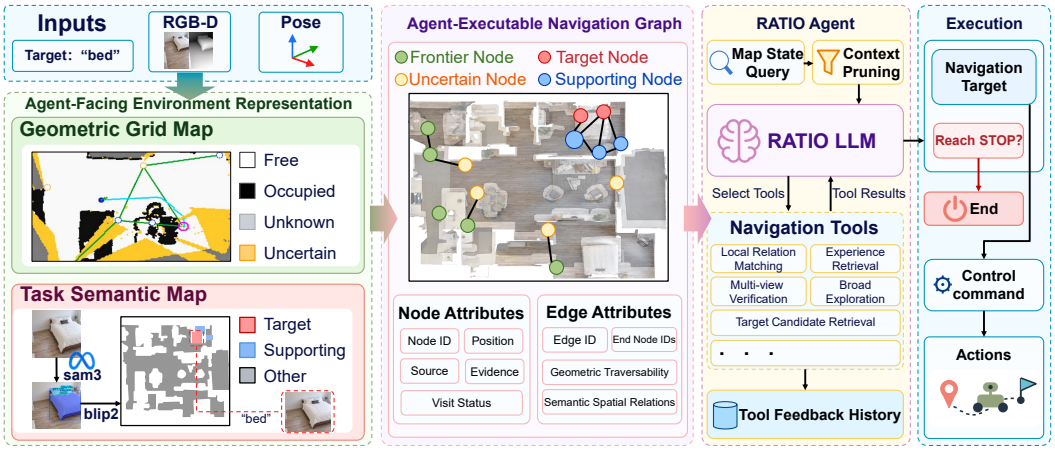

Figure 1: Overview of RATIO. RGB-D observations update the agent-facing environment representation and executable navigation graph. At each high-level decision point, the LLM selects tools from the decision context and uses their structured outputs to choose the next navigation goal. Tool calls and outputs are retained in the tool feedback history for subsequent decisions.

$$\mathcal{M}_t^G(\mathbf{x}) = \begin{cases} \text{unknown}, & n_t^{\text{obs}}(\mathbf{x}) = 0, \\ \text{free}, & n_t^{\text{obs}}(\mathbf{x}) \geq n_{\min},\ p_t^{\text{occ}}(\mathbf{x}) \leq \tau_{\text{free}}, \\ \text{occupied}, & n_t^{\text{obs}}(\mathbf{x}) \geq n_{\min},\ p_t^{\text{occ}}(\mathbf{x}) \geq \tau_{\text{occ}}, \\ \text{uncertain}, & \text{else}, \end{cases} \tag{1}$$

Here, $p_t^{\text{occ}}(\mathbf{x}) \in [0,1]$ is the occupancy probability of cell $\mathbf{x}$ at time t, and $n_t^{\text{obs}}(\mathbf{x}) \in \mathbb{N}_0$ is its cumulative number of valid observations. The minimum count for stable classification is $n_{\min} > 0$, and the state thresholds satisfy $0 < \tau_{\text{free}} < \tau_{\text{occ}} < 1$. An unknown cell has not been observed, whereas an uncertain cell has been observed without sufficient evidence for a stable classification. New or conflicting observations can move a cell into or out of uncertain, so early classifications are not fixed.

3.2.2 Semantic Regions for the Current Task

We organize successive observations and cell-wise semantic scores into semantic regions conditioned on the task. The semantic map retains regions with target or supporting evidence:

$$\mathcal{M}_t^{S,A} = \left\{ \widetilde{U}_k \mid 1 \leq k \leq \widetilde{N}_t,\ \alpha_k \neq \text{unknown} \right\}, \quad \widetilde{U}_k = (R_k, \mathbf{p}_k, \mathbf{e}_k, \mathcal{O}_k, \alpha_k), \tag{2}$$

Here, $\widetilde{U}_k$ is the k-th candidate semantic region conditioned on the task, and $\widetilde{N}_t$ is the number of candidate regions. R_k is its cell set, $\mathbf{p}_k$ its geometric center, $\mathbf{e}_k$ its task evidence, $\mathcal{O}_k$ its associated object clusters, and α_k its functional category.

We construct target-centric object clusters $O_j = \{(P_j^\ell, \gamma_j^\ell, n_j^\ell)\}_{\ell \in \mathcal{L}_{\text{obj}}}$, where $\mathcal{L}_{\text{obj}}$ is the object-label set. For label ℓ, P_j^ℓ is its 3D point cloud, γ_j^ℓ its fused detection confidence, and n_j^ℓ its accumulated fusion weight. Given the goal g, an LLM constructs the object-prompt set $\mathcal{Q}_{\text{obj}} = \mathcal{Q}_{tar} \cup \mathcal{Q}_{sup}$ for target and supporting objects. These prompts serve as semantic queries. Frame-wise fusion uses the cell projection range and viewing-angle weights of the underlying semantic fusion (Zhang et al., 2025). We form candidate regions from free cells whose maximum query score $a_t(\mathbf{x})$ reaches the semantic response threshold and from object-cluster projections:

$$\{R_k\}_{k=1}^{\widetilde{N}_t} = \text{Merge}\left(\left\{\mathbf{x} \mid \mathcal{M}_t^G(\mathbf{x}) = \text{free},\ a_t(\mathbf{x}) \geq \vartheta^{sem}\right\}, \{O_j\}_j\right), \tag{3}$$

Here, ϑ^{sem} is the semantic response threshold. Merge partitions selected cells into connected regions, projects object-cluster point clouds onto the geometric occupancy grid map, and merges connected projection fragments belonging to the same object cluster.

The task evidence $\mathbf{e}_k = (\bar{s}_k^{tar}, \bar{s}_k^{sup}, r_k)$ records target and supporting semantic responses and mean observation confidence. We compute the response to each query category q from the fused cell scores:

$$\bar{s}_k^q = \frac{\sum_{\mathbf{x}\in\mathcal{S}_k} w_t(\mathbf{x}) s_t^q(\mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{S}_k} w_t(\mathbf{x})}, \quad q \in \mathcal{Q}_{tar} \cup \mathcal{Q}_{sup}, \tag{4}$$

Here, $\mathcal{S}_k$ is the set of free cells in R_k with valid semantic scores and positive fusion confidence. $s_t^q(\mathbf{x})$ is the fused semantic score for query category q at cell $\mathbf{x}$, and $w_t(\mathbf{x})$ is its viewing-angle-based fusion confidence. For $c \in \{tar, sup\}$, the corresponding region response is $\bar{s}_k^c = \max_{q\in\mathcal{Q}_c} \bar{s}_k^q$. The mean observation confidence is calculated separately as $r_k = |\mathcal{S}_k|^{-1} \sum_{\mathbf{x}\in\mathcal{S}_k} w_t(\mathbf{x})$. $\mathcal{O}_k$ contains target-centric object clusters whose 2D projections overlap R_k. We assign the functional category using task evidence:

$$\alpha_k = \begin{cases} \text{target}, & r_k \geq \vartheta^{obs},\ \bar{s}_k^{tar} \geq \vartheta^{tar}, \\ \text{supporting}, & r_k \geq \vartheta^{obs},\ \bar{s}_k^{tar} < \vartheta^{tar},\ \bar{s}_k^{sup} \geq \vartheta^{sup}, \\ \text{unknown}, & \text{else}, \end{cases} \tag{5}$$

Here, ϑ^{tar} and ϑ^{sup} are the thresholds for target and supporting semantic responses, and ϑ^{obs} is the mean observation confidence threshold. Target and supporting regions generate nodes of the corresponding types.

3.3 Executable Navigation Graph

We construct an executable navigation graph $\mathcal{G}_t^A = (V_t^A, E_t^A)$ to guide choices among exploration, additional observation, and target verification. Its nodes are navigable positions with task attributes derived from the agent-facing environment representation; its edges encode symbolic spatial relations and geometric traversability.

3.3.1 Executable Navigation Nodes

The geometric occupancy grid map and semantic map conditioned on the task generate candidate positions, as detailed in Appendix B. After projection onto nearby stable free cells and distance-based merging, candidates are matched to existing nodes by position. Matched nodes retain their identifiers and update their attributes; unmatched candidates create new nodes, while nodes without supporting sources are removed. We define the current executable navigation nodes as:

$$V_t^A = \{v_i\}_{i=1}^{N_t^A}, \quad v_i = (\mathrm{id}_i, \mathbf{p}_i, \mathbf{b}_i, \eta_i, \mathcal{E}_i), \tag{6}$$

where N_t^A is the number of current nodes. id_i is the node identifier, $\mathbf{p}_i$ its position, $\mathbf{b}_i$ its source indicator, η_i its state, and $\mathcal{E}_i$ its evidence. The multi-hot source indicator $\mathbf{b}_i = [b_i^F, b_i^U, b_i^T, b_i^S] \in \{0,1\}^4$ records frontier, uncertain, target, and supporting sources, originating from frontier cluster centers, connected uncertain regions, target regions, and supporting regions, respectively. A semantic region conditioned on the task may generate multiple observation positions, and a node may have multiple sources. Both target and supporting nodes lie on stable free cells. Target nodes support target verification and additional observations from multiple viewpoints. Supporting nodes provide observation positions for regions formed from prompts for supporting objects.

The node state $\eta_i = (n_i^{\text{visit}}, t_i^{\text{last}})$ records the cumulative visit count and most recent visit time. The node evidence $\mathcal{E}_i = (\mathcal{E}_i^G, \mathcal{E}_i^S)$ contains geometric and semantic evidence. $\mathcal{E}_i^G = (\pi_i^{\text{free}}, \pi_i^{\text{unknown}}, \pi_i^{\text{uncertain}})$ gives the fractions of free, unknown, and uncertain cells in a local neighborhood centered at $\mathbf{p}_i$, describing nearby explorable space and geometric uncertainty. The semantic evidence is $\mathcal{E}_i^S = \{\mathbf{e}_k, \alpha_k, \mathcal{O}_k\}_{k\in\mathcal{K}_i}$, where $\mathcal{K}_i$ indexes semantic regions conditioned on the task that correspond directly to v_i or can be effectively observed from its position.

3.3.2 Executable Navigation Edges

After geometric traversability verification, we connect node pairs and record their path information and spatial relations:

$$E_t^A = \{e_{ij}\}_{(i,j)\in\mathcal{I}_t}, \quad e_{ij} = \left(\mathrm{id}_{ij}, \mathrm{id}_i, \mathrm{id}_j, \mathbf{g}_{ij}, \mathcal{R}_{ij}^S\right), \tag{7}$$

Here, $\mathcal{I}_t$ is the set of node pairs that pass geometric traversability verification, and id_{ij} is the edge identifier. The geometric traversability information $\mathbf{g}_{ij} = (d_{ij}, u_{ij})$ contains the planned path length between v_i and v_j and the fraction of uncertain cells along that path, as detailed in Appendix C. The relation set $\mathcal{R}_{ij}^S = \{(\ell_k, \rho, \ell_l)\}$ records semantic and spatial relations between regions associated with the two nodes. To describe these relations, we use directional and topological predicates:

$$\begin{aligned} \mathcal{L}_{\mathrm{dir}} &= \{\text{left-of},\ \text{right-of},\ \text{in-front-of},\ \text{behind}\}, \\ \mathcal{L}_{\mathrm{top}} &= \{\text{overlapping},\ \text{inside},\ \text{contains}\}. \end{aligned} \tag{8}$$

For regions R_k and R_l associated with v_i and v_j, ℓ_k and ℓ_l are their representative labels, and $\rho \in \mathcal{L}_{\mathrm{dir}} \cup \mathcal{L}_{\mathrm{top}}$. Each label is the highest-confidence fused object label in the region's associated object clusters, or the query label with the highest region response if no cluster is associated. Directional relations follow the relative positions of region centers, while topological relations follow their cell sets. We record at most one relation of each type per region pair. If either endpoint lacks usable semantic evidence, $\mathcal{R}_{ij}^S = \varnothing$, while the edge retains its geometric traversability information.

3.4 Navigation Decisions Driven by Tool Calls

3.4.1 Navigation Decision Process

RATIO makes a high-level decision after every K steps with valid observations. Before the first K steps accumulate, navigation follows a fixed frontier exploration policy. After this initialization period, if the node currently selected as the navigation goal becomes invalid, RATIO can trigger the next decision early after a grace period of G steps.

At decision point t, the system constructs the decision context $\mathcal{C}_t$ from the goal g, robot position $\mathbf{p}_t^{\mathrm{robot}}$, executable navigation graph $\mathcal{G}_t^A$, and tool feedback history $\mathcal{H}_t$, as shown in Fig. 2. The LLM selects and calls the tool set $\mathcal{U}_t^*$ using $\mathcal{C}_t$. If the context lacks sufficient information, the agent skips that decision round.

After receiving the tool outputs $\mathcal{Y}_t$, the LLM produces a navigation decision d_t selecting a node $v_i \in V_t^A$. The system records the tool calls and outputs in $\mathcal{H}_{t+1}$. The controller executes d_t and updates the map, executable navigation graph, and other state until the next decision point. When the agent is within the success distance of the target, it issues STOP.

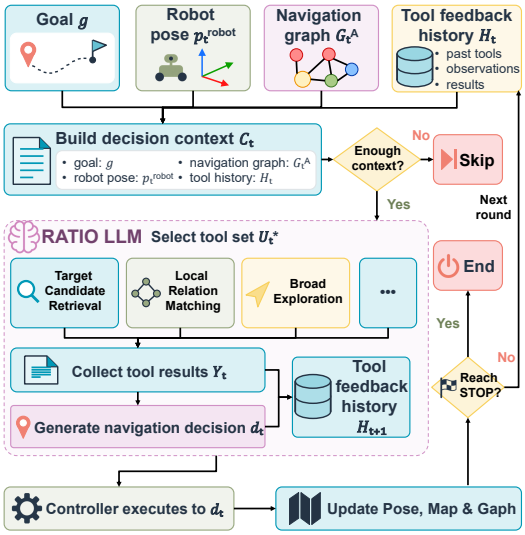

Figure 2: RATIO navigation decision process.

3.4.2 Navigation Tool Set

RATIO organizes seven navigation tools around the executable navigation graph into four categories of system support, evidence acquisition, candidate generation, and navigation recommendation. Map state query and context pruning provide system support. Local relation matching (LRM)

Table 1: Zero-shot object navigation results (%). Unreported results are marked with "–".

Category	Method	HM3Dv1		HM3Dv2		MP3D		HM3D-OVON Unseen	
		SR	SPL	SR	SPL	SR	SPL	SR	SPL
Semantic exploration	VLFM (Yokoyama et al., 2024a)	52.5	30.4	63.6	32.5	36.4	17.5	35.2	19.6
	ApexNav (Zhang et al., 2025)	59.6	33.0	76.2	38.0	39.2	17.8	–	–
	OpenFrontier (Padilla-Cerdio et al., 2026)	–	–	77.3	35.6	40.7	17.8	39.0	20.1
	DRIVE-Nav (Gao et al., 2026)	55.5	35.9	72.4	**41.3**	41.8	**22.6**	50.2	32.6
Map reasoning	SG-Nav (Yin et al., 2024)	54.0	24.9	49.6	25.5	40.2	16.0	–	–
	UniGoal (Yin et al., 2025b)	54.5	25.1	–	–	41.0	16.4	–	–
	BeliefMapNav (Zhou et al., 2025)	61.4	30.6	–	–	37.3	17.6	–	–
	MSGNav (Huang et al., 2026)	–	–	74.1	33.4	–	–	–	–
	SpaceVLN (Deng et al., 2026)	–	–	–	–	–	–	51.6	33.1
LLM decisions	InstructNav (Long et al., 2025)	58.0	20.9	–	–	–	–	–	–
	CogNav (Cao et al., 2025)	72.5	26.2	–	–	46.6	16.1	–	–
	Fast–Slow ReAct (Lan et al., 2026)	68.75	33.97	–	–	47.29	18.01	–	–
RATIO	**RATIO (DeepSeek V4.1 Flash)**	72.3	35.6	77.9	38.6	**48.2**	20.1	52.9	**34.2**
	RATIO (GPT-5.6 Luna)	**72.6**	**36.4**	**78.2**	38.2	47.8	20.6	**53.0**	34.0

and experience retrieval (ER) acquire evidence, target candidate search (TCS) generates candidates, and broad exploration (BE) and multi-view verification (MVV) recommend navigation goals.

At the start of each decision round, the system runs map state query to retrieve nodes, edges, and task evidence relevant to the current goal. Context pruning then retains the current position, necessary history, and tool descriptions within the input length limit. The LLM selects the other five tools using the resulting decision context. The tools handle geometric computation, path planning, visibility checks, and distance checks, allowing the LLM to focus on evidence selection and tool orchestration. Appendix Tables 4 and 5 give the tool functions and their inputs and outputs.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate RATIO on four benchmarks in Habitat (Savva et al., 2019). HM3Dv1 (Ramakrishnan et al., 2021), HM3Dv2 (Yadav et al., 2023), and MP3D (Chang et al., 2017) are used for standard zero-shot object navigation evaluation. For HM3D-OVON (Yokoyama et al., 2024b), we use only the Unseen split to evaluate open-vocabulary generalization to unseen target categories.

Metrics. All main experiments report success rate (SR) and success weighted by path length (SPL). All experiments, including the ablations, use the same settings for shared components. Appendix E provides the parameter values and evaluation criteria.

Implementation. HM3Dv1, HM3Dv2, MP3D, and HM3D-OVON Unseen contain 2,000, 1,000, 2,195, and 3,000 validation episodes, respectively. The agent uses discrete actions to move forward, turn left or right, look up or down, and stop. RATIO uses SAM 3 (Carion et al., 2026) for target detection and instance segmentation and BLIP-2 (Li et al., 2023) for image-text similarity. All models have frozen weights. The two RATIO variants differ only in the high-level LLM.

Baselines. Selected baselines require no task-specific navigation training, prebuilt maps, or offline navigation trajectory libraries. Based on their primary high-level decision mechanism, they fall into semantic exploration, map reasoning, and LLM decision methods. These groups respectively select frontiers using semantic relevance from a VLM, infer target locations from explicit semantic or spatial maps, and directly generate high-level navigation decisions with an LLM.

4.2 Quantitative Analysis

Table 1 shows that RATIO achieves the highest SR on all four benchmarks. With GPT-5.6 Luna, SR reaches 72.6%, 78.2%, and 53.0% on HM3Dv1, HM3Dv2, and HM3D-OVON Unseen, respectively. With DeepSeek V4.1 Flash, SR reaches 48.2% on MP3D. For SPL, GPT-5.6 Luna reaches 36.4% on

Table 2: Configurations and results of the four progressive ablations (%).

Setting	Executable graph	LLM	Tool calls	HM3Dv1		HM3Dv2		MP3D		HM3D-OVON Unseen	
				SR	SPL	SR	SPL	SR	SPL	SR	SPL
A) Backbone	×	×	×	53.1	28.8	61.0	31.0	38.5	15.8	31.5	18.5
B) Graph	✓	×	×	57.8	31.7	66.2	34.5	42.5	18.2	37.8	22.7
C) Direct Agent	✓	✓	×	68.5	34.0	74.5	37.8	45.8	20.1	47.3	30.6
D) Full RATIO	✓	✓	✓	**72.6**	**36.4**	**78.2**	**38.2**	**47.8**	**20.6**	**53.0**	**34.0**

HM3Dv1, and DeepSeek V4.1 Flash reaches 34.2% on HM3D-OVON Unseen; both are the highest values in the table. Semantic exploration methods mainly select exploration regions using predefined scores, whereas RATIO chooses among exploration, candidate search, and additional observation as target evidence changes. On HM3Dv2, ApexNav achieves 76.2% SR and RATIO reaches 78.2%. This result supports organizing navigation capabilities according to the current task state.

Map reasoning methods use scene structure to assist target localization. RATIO organizes task evidence and traversability into an executable navigation graph and provides relation and candidate query results through tools, grounding agent decisions in reachable navigation positions. On HM3D-OVON Unseen, RATIO exceeds SpaceVLN in both SR and SPL. Compared with existing LLM decision methods, RATIO also includes capability selection in high-level decisions and delegates specific computations to specialized tools. On HM3Dv1, RATIO with GPT-5.6 Luna reaches 72.6% SR, slightly above CogNav's 72.5%, and 36.4% SPL, above CogNav's 26.2%. After changing the high-level LLM, the largest SR difference between the two RATIO variants across the benchmarks is 0.4 percentage points, indicating that the navigation representation and tool interface maintain relatively stable performance with the two tested models.

4.3 Ablation Study

We use four progressive settings, shown in Table 2. A) is a non-agent navigation backbone sharing the perception, mapping, and path planning modules. B) adds the executable navigation graph but still selects nodes using deterministic rules. C) enables direct LLM decisions on the same graph without exposing agent tools. D) exposes the complete navigation tool set and retains the tool feedback history, forming the full RATIO system. All settings share the data, perception, map updates, path planning, executor, and evaluation protocol. Settings C and D use GPT-5.6 Luna as the high-level LLM.

Table 2 shows that B) improves both SR and SPL over A) on all four benchmarks, indicating that the executable navigation graph provides useful candidates even before agent reasoning. The gain from B) to C) is especially pronounced on HM3Dv1 and HM3D-OVON Unseen, where SR increases by 10.7 and 9.5%, respectively. This suggests that context-dependent decisions use the graph more effectively than fixed selection. D) further improves SR on every benchmark and SPL on all four benchmarks. On HM3D-OVON Unseen, it raises SR from 47.3% to 53.0% and SPL from 30.6% to 34.0%, providing evidence that tool interaction and feedback history add value beyond direct agent decisions.

4.4 Agent Behavior Analysis

Figure 3 visualizes RATIO's tool choices along three trajectories; Fig. 4 summarizes tool use and token cost across episodes.

Figure 4(a) shows a changing tool-use pattern across navigation stages. TCS has the highest selection rate in the first 200 steps, whereas BE becomes the most frequently selected tool after step 200. ER reaches its highest selection rate at steps 200–299, while LRM and MVV are selected less often throughout. This shift shows that RATIO changes its tool choices as navigation progresses, consistent with its task-adaptive interaction design.

Figure 4(b) shows that mean token consumption per ongoing episode fluctuates rather than increasing with episode step, with local peaks around steps 100 and 250. Mean tool calls also rise near

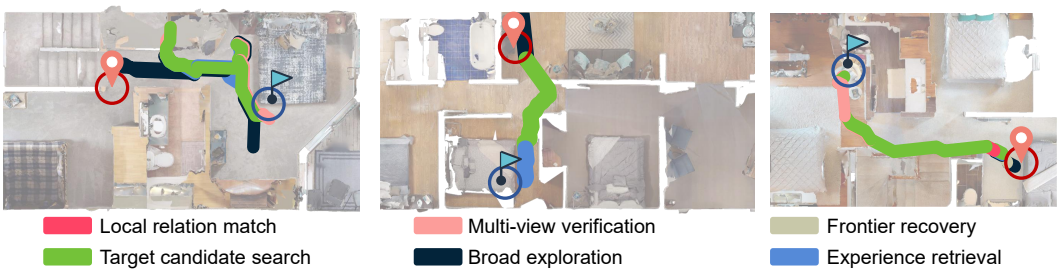

Figure 3: Three representative RATIO navigation trajectories. Path colors identify the navigation tool selected for each segment.

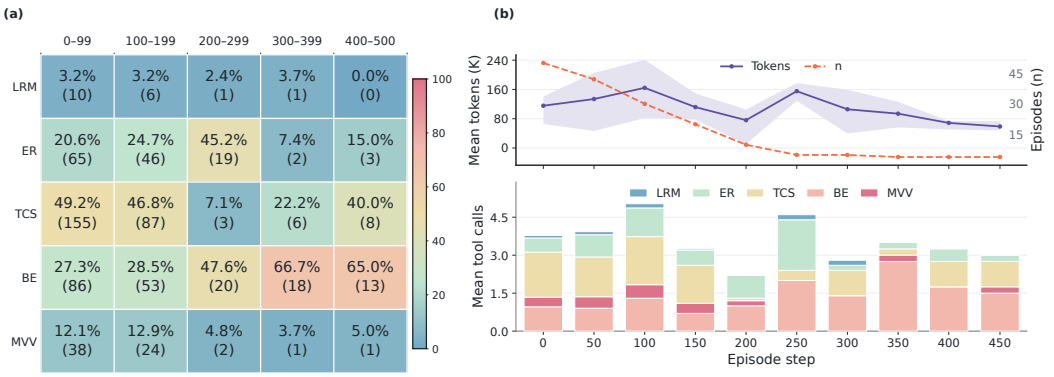

Figure 4: Agent behavior analysis. (a) The proportion of decisions selecting each navigation tool and tool-call counts across navigation stages. (b) Mean token consumption per ongoing episode and the number of ongoing episodes (top), and mean tool calls per ongoing episode in each step interval (bottom). Shading indicates the interquartile range of token consumption.

these steps, suggesting that the token peaks may reflect more tool interactions and the resulting decision context. Both token consumption and tool calls generally decrease later, while the number of ongoing episodes falls sharply. The late-stage averages therefore describe a small set of episodes that remain active.

5 Conclusion

We presented RATIO, a reasoning agent framework for zero-shot object navigation. Existing methods often organize navigation capabilities through predefined procedures and combine semantic reasoning with spatial computation. RATIO organizes environment information into an agent-facing environment representation and an executable navigation graph, and exposes navigation capabilities as tools. At each decision point, the LLM selects tools based on the task context, while dedicated modules perform geometric, retrieval, and candidate computations. This separates capability selection from execution.

We evaluated RATIO on HM3Dv1, HM3Dv2, MP3D, and HM3D-OVON Unseen and tested its adaptation to two high-level LLMs. Taking the better result of the two variants for each metric, SR reaches 72.6%, 78.2%, 48.2%, and 53.0% on the four benchmarks, respectively, and SPL reaches 36.4%, 38.6%, 20.6%, and 34.2%. In the comparison, RATIO achieves the highest SR on all four benchmarks and the highest SPL on HM3Dv1 and HM3D-OVON Unseen. The SR difference between the two variants is at most 0.4 percentage points. Progressive ablations further support the roles of the executable navigation graph, agent decisions, and tool interaction. Future work will extend the executable navigation graph and tool orchestration on demand to multi-target search, compositional language instructions, and real-world robot navigation.

REFERENCES

Amir Bar, Gaoyue Zhou, Danny Tran, et al. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01472. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bar_Navigation_World_Models_CVPR_2025_paper.html.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, et al. ObjectNav revisited: On evaluation of embodied agents navigating to objects. In *arXiv preprint arXiv:2006.13171*, 2020. URL https://arxiv.org/abs/2006.13171.

Yihan Cao, Jiazhao Zhang, Zhinan Yu, et al. Cognav: Cognitive process modeling for object goal navigation with llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9550–9560, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cao_CogNav_Cognitive_Process_Modeling_for_Object_Goal_Navigation_with_LLMs_ICCV_2025_paper.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Angel Chang, Angela Dai, Thomas Funkhouser, et al. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision*, pp. 667–676, 2017. doi: 10.1109/3DV.2017.00067.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, et al. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024a.

Jiaqi Chen, Bingqian Lin, Ran Xu, et al. MapGPT: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9796–9810, 2024b. doi: 10.18653/v1/2024.acl-long.529.

Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Junting Chen, Guohao Li, Suryansh Kumar, et al. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. In *Proceedings of Robotics: Science and Systems XIX*, 2023. doi: 10.15607/RSS.2023.XIX.075.

Yucheng Deng, Pingrui Lai, Xinhai Li, Chenjia Bai, Xiaoheng Deng, Chengnuo Sun, Xuelong Li, and Hua Yang. Spacevln: A zero-shot vision-and-language navigation agent with online spatial cognitive memory and reasoning, 2026. URL https://arxiv.org/abs/2606.08992.

Hongyan Feng, Sunlai Chen, Xuanyu Liu, Miao Pan, Yangfan Xie, Yuxiang Cui, Zhongxiang Zhou, Rong Xiong, Wenqi Zhang, Jianwei Yin, Yueting Zhuang, and Xuhong Zhang. Embodied-navigator: Point, think, memorize, and align for efficient navigation, 2026. URL https://arxiv.org/abs/2608.17512.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, et al. 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. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gadre_CoWs_on_Pasture_Baselines_and_Benchmarks_for_Language-Driven_Zero-Shot_Object_CVPR_2023_paper.html.

Maoguo Gao, Zejun Zhu, Zhiming Sun, Zhengwei Ma, Longze Yuan, Zhongjing Ma, Zhigang Gao, Jinhui Zhang, and Suli Zou. Drive-nav: Directional reasoning, inspection, and verification for efficient open-vocabulary navigation. *IEEE Robotics and Automation Letters*, 11(10):11158–11165, 2026. doi: 10.1109/LRA.2026.3723310.

Junjun Hu, Xinda Xue, Botao Ren, et al. Astranav-memory: Contexts compression for long memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8097–8109, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hu_AstraNav-Memory_Contexts_Compression_for_Long_Memory_CVPR_2026_paper.html.

Hao Huang, Yu Hao, Congcong Wen, et al. GAMap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39386–39408, 2024. URL https://arxiv.org/abs/2409.16997.

Wenlong Huang, Fei Xia, Ted Xiao, et al. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782, 2023. URL https://proceedings.mlr.press/v205/huang23c.html.

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, 2026. URL https://arxiv.org/abs/2511.10376.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Yuxuan Kuang, Hai Lin, and Meng Jiang. OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1510–1523, 2024. URL https://aclanthology.org/2024.findings-naacl.96/.

Zhaochen Lan, Zhi Yang, Yuxiang Fu, et al. Hierarchical fast–slow ReAct agent for zero-shot object-goal navigation. *arXiv preprint arXiv:2608.09816*, 2026. URL https://arxiv.org/abs/2608.09816.

Sibaek Lee, Hyeonwoo Yu, Giseop Kim, et al. LAMP: Implicit language map for robot navigation. *IEEE Robotics and Automation Letters*, 10(12):12365–12372, 2025. doi: 10.1109/LRA.2025.3619820.

Junnan Li, Dongxu Li, Silvio Savarese, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yijian Li, Changze Li, Hantian Shi, et al. AgenticNav: Zero-shot vision-and-language navigation as a tool-calling harness. In *arXiv preprint arXiv:2606.10577*, 2026. URL https://arxiv.org/abs/2606.10577.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, et al. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2049–2060, 2025. URL https://proceedings.mlr.press/v270/long25b.html.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, et al. ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d0a8b91fa097cbd9c9a2d8b0e2b8a1d2-Abstract-Conference.html.

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

Saiqian Peng, Liping Lu, Bing Shi, and Duanfeng Chu. Less is more: A simple yet effective object-centric prompting strategy for vision-language reasoning in autonomous driving. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 97–114, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37556-8.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, et al. Habitat-matterport 3d dataset (HM3D): 1000 large-scale 3d environments for embodied AI. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2109.08238.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, et al. 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.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Savva_Habitat_A_Platform_for_Embodied_AI_Research_ICCV_2019_paper.html.

Yujie Tang, Meiling Wang, Yinan Deng, et al. Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments. *IEEE Robotics and Automation Letters*, 2025. doi: 10.1109/LRA.2025.3592071. URL https://arxiv.org/abs/2501.04279.

Yiyao Wang, Sixian Zhang, Keming Zhang, et al. Trajrag: Retrieving geometric-semantic experience for zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15166–15176, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_TrajRAG_Retrieving_Geometric-Semantic_Experience_for_Zero-Shot_Object_Navigation_CVPR_2026_paper.html.

Pengying Wu, Yao Mu, Bingxian Wu, et al. VoroNav: Voronoi-based zero-shot object navigation with large language model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53757–53775, 2024. URL https://proceedings.mlr.press/v235/wu24aj.html.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4927–4936, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yadav_Habitat-Matterport_3D_Semantics_Dataset_CVPR_2023_paper.html.

Zebin Yang, Sunjian Zheng, Tong Xie, et al. Efficientnav: Towards on-device object-goal navigation with navigation map caching and retrieval. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/067437c6d5d0369b6d09200bef89715b-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Hang Yin, Xiuwei Xu, Zhenyu Wu, et al. SG-Nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5285–5307, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/098491b37deebbe6c007e69815729e09-Abstract-Conference.html.

Hang Yin, Haoyu Wei, Xiuwei Xu, et al. 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, 2025a. URL https://proceedings.mlr.press/v305/yin25a.html.

Hang Yin, Xiuwei Xu, Linqing Zhao, et al. 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, et al. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation*, pp. 42–48, 2024a. doi: 10.1109/ICRA57147.2024.10610712. URL https://openreview.net/forum?id=gdw1zUTABk.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, et al. HM3D-OVON: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4299–4306, 2024b. doi: 10.1109/IROS58592.2024.10802411. URL https://arxiv.org/abs/2409.14296.

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

Mingjie Zhang, Yuheng Du, Chengkai Wu, et al. Apexnav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *IEEE Robotics and Automation Letters*, 10(11):11530–11537, 2025. doi: 10.1109/LRA.2025.3606388.

Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7641–7649, 2024. doi: 10.1609/aaai.v38i7.28597.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, et al. 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, 2023. URL https://proceedings.mlr.press/v202/zhou23r.html.

Zibo Zhou, Yue Hu, Lingkai Zhang, et al. BeliefMapNav: 3d voxel-based belief map for zero-shot object navigation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.00943.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and

Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language refinement and the creation of the framework figure. We have not used generative AI tools for research ideation, methodology design, experimental design, data generation, data analysis, or the generation of citations and scientific claims; the remaining disclosure categories are not applicable to this work. We reviewed all AI-assisted text and visual content, verified their technical consistency with the implemented method, and manually checked the accuracy and attribution of all cited work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

A COMPARISON OF GRAPH REPRESENTATION CAPABILITIES

We compare our representation with existing graph representations in terms of the capabilities needed for agent decisions, as shown in Table 3. A geometric topological graph provides navigable positions and traversability verification but does not contain task semantic evidence or geometric uncertainty. Scene graphs and goal graphs provide semantic relations, but their nodes are not guaranteed to be navigable. A set of frontier candidates is executable but lacks an evidence structure. A trajectory experience graph supports experience retrieval but is not constructed for the current navigation decision. Our executable navigation graph organizes node sources, geometric uncertainty, task semantic evidence, semantic and spatial relations, and traversability verification in one structure, allowing the agent to query any part on demand.

Table 3: Comparison of graph representation capabilities. ✓ indicates explicit support, △ indicates partial support, and a blank entry indicates no direct support.

Graph representation	Representative method	Navigable positions	Node sources	Geometric uncertainty	Task semantic evidence	Semantic and spatial relations	Traversability verification
Geometric topological graph	VoroNav (Wu et al., 2024)	✓					✓
Scene graph	SG-Nav (Yin et al., 2024) UniGoal (Yin et al., 2025b)		△		✓	✓	
Frontier candidate set	ApexNav (Zhang et al., 2025) VLFM (Yokoyama et al., 2024a)	✓					△
Trajectory experience graph	TrajRAG (Wang et al., 2026)	△	△		✓	△	
Executable navigation graph	**RATIO**	✓	✓	✓	✓	✓	✓

B NODE GENERATION FROM MULTIPLE SOURCES

To place a candidate position on a nearby stable free cell, we define a local projection:

$$\Pi_t^{\mathrm{free}}(\mathbf{c}) = \arg \min_{\mathbf{x} \in \mathcal{N}_{r_p}(\mathbf{c}),\ \mathcal{M}_t^G(\mathbf{x}) = \mathrm{free}} \|\mathbf{x} - \mathbf{c}\|_2 , \tag{9}$$

where $\mathbf{c}$ is a candidate position and $\mathcal{N}_{r_p}(\mathbf{c})$ is its local neighborhood of radius r_p. If no stable free cell exists in this neighborhood, no node is generated for the candidate.

For $\mathrm{frontier}$ nodes, the system uses incremental extraction and principal component analysis (PCA) clustering to obtain cluster centers $\mathbf{c}_m^F$ and projects them onto stable free cells, producing the position set for frontier exploration nodes:

$$\mathcal{P}_t^F = \left\{ \Pi_t^{\text{free}}(\mathbf{c}_m^F) \right\}_{m=1}^{N_t^F}, \tag{10}$$

where N_t^F is the number of frontier clusters.

For connected uncertain regions and semantic regions conditioned on the task, we first identify usable observation positions around a region:

$$\text{View}_t(R) = \left\{ \mathbf{x} \;\middle|\; \mathcal{M}_t^G(\mathbf{x}) = \text{free},\ d_{\min} \le d(\mathbf{x}, R) \le d_{\max},\ \text{Vis}(\mathbf{x}, R) = 1 \right\}, \tag{11}$$

where $d(\mathbf{x}, R)$ is the spatial distance from candidate position $\mathbf{x}$ to region R, $\text{Vis}(\mathbf{x}, R)$ indicates that the position satisfies the visibility condition for R, and $d_{\min}$ and $d_{\max}$ bound the valid observation distance.

Connected-region extraction on uncertain cells yields the region set $\mathcal{R}_t^U$. uncertain, target, and supporting regions generate geometric, target, and supporting observation nodes, respectively, with position sets defined as

$$\begin{aligned}
\mathcal{P}_t^U &= \bigcup_{R \in \mathcal{R}_t^U} \text{Select}_{N_{\text{view}}}(\text{View}_t(R)), \\
\mathcal{P}_t^T &= \bigcup_{k:\alpha_k = \text{target}} \text{Select}_{N_{\text{view}}}(\text{View}_t(R_k)), \\
\mathcal{P}_t^S &= \bigcup_{k:\alpha_k = \text{supporting}} \text{Select}_{N_{\text{view}}}(\text{View}_t(R_k)),
\end{aligned} \tag{12}$$

where N_{view} is the maximum number of observation positions retained per region. $\text{Select}_{N_{\text{view}}}(\cdot)$ requires any two selected positions to be at least d_{sep} apart, avoiding overly dense nodes around the same region.

We combine the four candidate position sets:

$$\mathcal{P}_t^A = \text{Unique}_{\text{grid}}\left(\mathcal{P}_t^F \cup \mathcal{P}_t^U \cup \mathcal{P}_t^T \cup \mathcal{P}_t^S\right), \tag{13}$$

where $\text{Unique}_{\text{grid}}(\cdot)$ retains only one candidate position for a cell. When a cell has multiple sources, the system records all of them in the node-source vector $\mathbf{b}_i$. It then matches $\mathcal{P}_t^A$ to existing nodes by position and updates the current executable navigation node set V_t^A.

C Geometric Traversability Verification

For path planning, the grid traversal cost $C_t^G(\mathbf{x})$ is 1 for free cells and $1 + \lambda_u$ for uncertain cells, where $\lambda_u > 0$ is the additional traversal cost. occupied and unknown cells are impassable.

For any two distinct current executable navigation nodes v_i and v_j, the system uses A* on this cost map to plan a connecting path:

$$P_{ij} = \text{A}^*\left(C_t^G, \mathbf{p}_i, \mathbf{p}_j\right). \tag{14}$$

Here, P_{ij} is the planned path connecting the nodes and is $\varnothing$ if the search fails. In implementation, existing path results may be reused when the relevant map area has not changed, reducing repeated searches. For a successful path, we compute the fraction of uncertain cells to measure geometric uncertainty:

$$u_{ij} = \frac{1}{|P_{ij}|} \sum_{\mathbf{x} \in P_{ij}} \mathbf{1}\left[\mathcal{M}_t^G(\mathbf{x}) = \text{uncertain}\right], \tag{15}$$

where $|P_{ij}|$ is the number of cells in the path and $\mathbf{1}[\cdot]$ is the indicator function. We select node pairs eligible for navigation edges by defining

$$\mathcal{I}_t = \{(i,j) \mid i \neq j, \quad P_{ij} \neq \varnothing, \quad u_{ij} \leq \tau_u\}, \tag{16}$$

where τ_u is the upper bound on the fraction of $\mathrm{uncertain}$ cells along a path.

D NAVIGATION TOOLS

Table 4 lists the navigation tools' categories, sources, and functions, while Table 5 lists their inputs and outputs. Tool outputs contain evidence identifiers so that the agent can revisit the original records.

Table 4: Categories, sources, and functions of the RATIO navigation tools.

Category	Tool	Source	Main function
System support	Map state query	Ours	Query nodes, edges, and task evidence relevant to the current decision
System support	Context pruning	EfficientNav (Yang et al., 2025)	Compress the context within the input budget while preserving reference consistency
Evidence acquisition	Local relation matching	UniGoal (Yin et al., 2025b)	Return semantic and spatial relations between regions and the results of condition matching
Evidence acquisition	Experience retrieval	TrajRAG (Wang et al., 2026)	Retrieve similar historical trajectory segments and the observations added afterward
Candidate generation	Target candidate search	CogNav (Cao et al., 2025)	Return target and supporting candidates and their detection statistics across frames
Navigation recommendation	Broad exploration	CogNav (Cao et al., 2025)	Recommend frontier or uncertain-region nodes using weighted path distance
Navigation recommendation	Multi-view verification	CogNav (Cao et al., 2025)	Determine whether a candidate needs another scan and provide an additional observation position

Table 5: Inputs and outputs of the navigation tools.

Tool	Main input	Output
Map state query	Current goal, robot position, executable navigation graph, tool feedback history	Relevant node and edge records, brief object-cluster records, recent entries from the tool feedback history
Context pruning	Query results, tool descriptions, input length limit	Pruned context, identifiers of retained and deleted records, actual token count
Local relation matching	Specified nodes or target candidates, optional relation conditions	Relation triples, source regions and edges, satisfied, unsatisfied, or unknown status for each condition
Experience retrieval	Semantic evidence sequence for current nodes, candidate nodes, task navigation history	Historical segment identifiers, similarity and its components, historical actions and actual outcomes
Target candidate search	Current goal, object clusters, task semantic evidence, associated nodes	Object identifiers, fused detection confidence, associated region evidence, node identifiers
Broad exploration	$\mathrm{frontier}$ nodes, geometric evidence, visit state, traversability information	Candidate node identifiers, ranking values, path information, reasons for exclusion
Multi-view verification	Target candidates, previous observations, associated nodes	Node identifiers, minimum viewpoint difference, object distance, path information, reasons for exclusion

E PARAMETER SETTINGS

Geometric and semantic maps. The geometric occupancy grid map uses free and occupied probability thresholds of 0.35 and 0.80. The minimum observation count is $n_{\min} = 2$ in all settings. The additional A* traversal cost for an $\mathrm{uncertain}$ cell is $\lambda_u = 1.0$. For semantic regions conditioned on the task, the semantic response, target response, supporting response, and mean observation confidence thresholds are $\vartheta^{sem} = 0.50$, $\vartheta^{tar} = 0.40$, $\vartheta^{sup} = 0.50$, and $\vartheta^{obs} = 0.30$, respectively.

Executable navigation graph. Candidate positions are projected onto stable free cells within 0.50 m and merged within 0.75 m. Observation positions are selected 0.20–1.50 m from their associated regions. Node pairs within 3.00 m are considered for an edge, and the fraction of $\mathrm{uncertain}$ cells along a verified path cannot exceed 0.50.

Navigation decisions and tools. RATIO uses a decision interval of $K = 5$ steps. The currently selected navigation goal can trigger an earlier decision after $G = 3$ consecutive steps of invalidity. The tool feedback history stores up to 32 entries, of which up to 8 enter the decision context. ER returns at most three historical segments per candidate. BE weights $\mathrm{frontier}$ and $\mathrm{uncertain}$ sources by 0.7 and 0.3, respectively. MVV requires at least two observation frames and both the mean and latest detection confidence to be at least 0.5.

Evaluation protocol. Episodes are limited to 500 steps. The success distance is 0.20 m on HM3Dv1, HM3Dv2, and MP3D, and 0.25 m on HM3D-OVON Unseen.