This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TrajRAG: Retrieving Geometric-Semantic Experience for Zero-Shot Object Navigation

Yiyao Wang[1,2], Sixian Zhang[1,2], Keming Zhang[1,2], Xinhang Song[1,2]*, Songjie Du[2], Shuqiang Jiang[2,3]
[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing
[2]University of Chinese Academy of Sciences, Beijing, [3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing
{yiyao.wang, sixian.zhang, keming.zhang, xinhang.song}@vipl.ict.ac.cn,
dusongjie25@mails.ucas.ac.cn, sqjiang@ict.ac.cn

Abstract

Existing zero-shot Object Goal Navigation (ObjectNav) methods often exploit commonsense knowledge from large language or vision-language models to guide navigation. However, such knowledge arises from internet-scale text rather than embodied 3D experience, and episodic observations collected during navigation are typically discarded, preventing the accumulation of lifelong experience. To this end, we propose Trajectory RAG (TrajRAG), a retrieval-augmented generation framework that enhances large-model reasoning by retrieving geometric–semantic experiences. TrajRAG incrementally accumulates episodic observations from past navigation episodes. To structure these observations, we propose a topological-polar (topo-polar) trajectory representation that compactly encodes spatial layouts and semantic contexts, effectively removing redundancies in raw episodic observations. A hierarchical chunking structure further organizes similar topo-polar trajectories into unified summaries, enabling coarse-to-fine retrieval. During navigation, candidate frontiers generate multiple trajectory hypotheses that query TrajRAG for similar past trajectories, guiding large-model reasoning for waypoint selection. New experiences are continually consolidated into TrajRAG, enabling the accumulation of lifelong navigation experience. Experiments on MP3D, HM3D-v1, and HM3D-v2 show that TrajRAG effectively retrieves relevant geometric–semantic experiences and improves zero-shot ObjectNav performance.

1. Introduction

Object Goal Navigation (ObjectNav) requires an embodied agent to locate a user-specified object category in an unseen environment, based only on egocentric RGB-D observations. To navigate efficiently without exhaustive exploration, the agent needs to acquire semantic–geometric priors that encode how objects are typically arranged within spatial layouts. Prior works train models in seen environments to learn such priors, including object co-occurrence relations [8, 10, 56], RL-based navigation policies [7, 9, 20], object-location prediction functions [16, 30, 37, 50, 55], and feature and layout adaptations for unseen environments [53, 54]. Recently, large-scale pretrained models (LLMs and VLMs) have demonstrated strong zero-shot common-sense reasoning abilities, inspiring methods that leverage them as general priors to perform ObjectNav in a zero-shot manner.

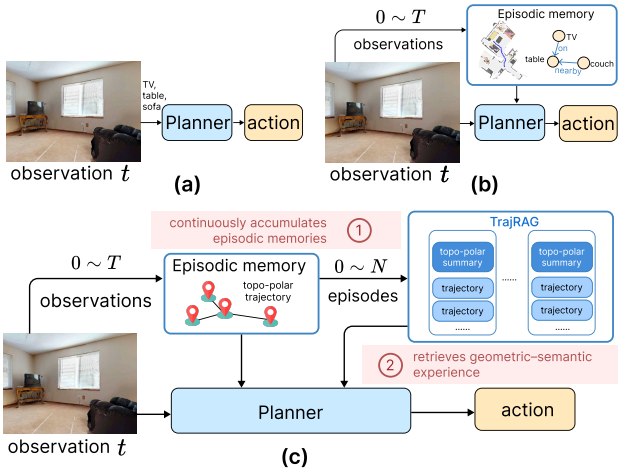

Figure 1. Comparisons with LLM/VLM-based Zero-shot ObjectNav methods. (a) **Single-step context**. the planner (LLM/VLM) receives raw textual observations from the single timestep. (b) **Episodic context**. episodic memory is structured into prompts for reasoning but discarded after each episode. (c) **Episode + experience context**. our TrajRAG serves as long-term memory that continuously accumulates episodic memory and retrieves geometric–semantic experience for planning.

Existing LLM/VLM-based methods for zero-shot ObjectNav as illustrated in Fig. 1. Some methods (Fig. 1a) [2, 14, 58] directly query large models with observations at

*Corresponding author.

each timestep for action planning, but the absence of episodic memory often results in redundant exploration and revisiting previously seen areas. Other methods (Fig. 1b) introduce structured episodic memories, e.g., similarity maps [18, 47, 57, 60], scene graphs [45, 46, 57], landmark graphs [42], or 3D-language feature fields [36], providing richer semantic cues for reasoning. However, while these episodic memories are scene-specific, the large models used for decision-making rely on scene-agnostic knowledge learned from web-scale text rather than 3D spatial experience. Furthermore, such episodic memories are typically discarded after each episode, limiting their scalability toward lifelong navigation. In contrast, human navigation relies on both short-term and long-term memory [12, 28, 40]: the former encodes immediate environmental details, while the latter retrieves related experiences to support decision-making. Moreover, short-term memories are gradually consolidated into long-term memory [23, 27, 34], enabling continual learning [31]. To emulate this capability, an embodied agent requires a systematic internal representation [19], to continuously accumulate past experiences for lifelong navigation.

Motivated by this, we propose Trajectory RAG (TrajRAG), which aims to build a "long-term" memory of navigation experience (Fig. 1c) that (1) continuously accumulates episodic memories and (2) retrieves geometric–semantic experience to enhance the reasoning of large models. For the first objective, raw episodic memories (trajectories of RGB-D observations) are redundant both within trajectories (due to revisits or local loops) and across trajectories (from spatial overlap among runs in the same scene). To achieve memory compactness, we propose the topological-polar (topo-polar) trajectory to organize these raw trajectories. Specifically, we first build a semantic map based on RGB-D observations, then skeletonize the navigable region to extract topological nodes. Around each node, space is discretized into polar sectors (30° per sector), each sector records the observed semantic labels. In the incremental construction of TrajRAG, the topo-polar trajectory offers two benefits: (1) The topological skeleton provides a structured representation of raw trajectories, enabling self-checking through the detecting and pruning of redundant segments, ensuring stored trajectories are efficient; and (2) The polar-sector node captures relative geometric–semantic layouts as a distinctive fingerprint for matching new trajectories with existing ones in TrajRAG, determining whether the new one is redundant. This guarantees that the content of TrajRAG remains compact and informative. Compared with other representations, the topo-polar trajectory achieves more accurate spatial matching than scene graphs, while being more flexible and requiring less computation cost than map-based or point-cloud–based matching.

For the second objective, TrajRAG adopts a hierarchical chunking architecture to enable efficient retrieval, in which each chunk corresponds to a topo-polar trajectory. Chunks with similar topological structures are grouped and merged into a unified topo-polar summary, serving as a descriptor for the group's geometric–semantic layout. During retrieval, a coarse-to-fine strategy is employed. Coarse matching between the query and topo-polar summaries first identifies relevant groups with similar layout, followed by fine-grained matching within these groups to locate the similar trajectory chunks. As the number of trajectory chunks far exceeds that of summaries, a trajectory encoder is proposed to accelerate fine-grained retrieval. It embeds topo-polar trajectories, ensuring that trajectories from the same group with the same navigation goal lie close in the embedding space, while unrelated ones remain distant.

In this paper, we propose TrajRAG, a retrieval-augmented generation framework that retrieves geometric–semantic experiences for zero-shot object navigation. During TrajRAG inference in navigation, the agent incrementally constructs a semantic map and transforms it into a topo-polar trajectory. At each timestep, candidate exploration points are selected as the centers of individual frontiers. Based on these candidates and the existing topo-polar trajectory, a set of candidate trajectories is generated. These candidate trajectories then perform coarse-to-fine retrieval over the TrajRAG, retrieving trajectory chunks that contain relevant geometric–semantic experiences. The retrieved experiences are provided to assist the policy model (e.g., LLMs) in selecting the next waypoint for exploration. After completing an episode, the entire trajectory is incorporated into the TrajRAG, enabling lifelong experience accumulation. We evaluate TrajRAG on zero-shot ObjectNav benchmarks across MP3D, HM3D-v1, and HM3D-v2. Experimental results demonstrate that integrating TrajRAG improves navigation accuracy and efficiency. Furthermore, we validate the effectiveness of TrajRAG, demonstrating that it can effectively retrieve relevant geometric–semantic experiences during the decision-making process of large models. This capability bridges scene-agnostic commonsense reasoning and scene-specific experience, thereby enhancing generalization in zero-shot ObjectNav.

2. Related Works

2.1. Zero-shot Object Navigation

In zero-shot object navigation, with the rise of pre-trained models, recent methods utilize knowledge from pre-trained models to assist navigation decision-making and can be broadly categorized into two types: those leveraging Large Language Models (LLMs) and Vision-Language Models (VLMs) [14, 26, 47, 59]. LLM-based methods primarily leverage the reasoning capabilities of LLMs to assist navigation. They extract and represent environmental informa-

tion as natural-language structures, enabling the LLM to infer potential target locations and select the next exploration waypoint [42, 46, 48, 58]. For instance, VoroNav [42] represents traversable paths as candidate trajectories and formulates the objects and scenes along these trajectories into natural-language descriptions, allowing the LLM to reason about and select the optimal path. Beyond using LLMs to infer target locations or assisting path selection, CogNav [3] elevates the LLM to determine the current navigation phase of the agent, thereby deciding which exploration policy the agent should execute. VLM-based methods [18, 47, 60] extracts frontier points on a semantic map; then the VLM converts the similarity between frontier observations and goal text into frontier values, thereby guiding path planning. Recent advancements empower VLMs with 3D understanding; e.g., Dynam3D [38] utilizes dynamic 3D tokens to provide long-term memory within a single environment.

Although prior methods use pretrained models, their knowledge is not grounded in real scene experience and cannot accumulate transferable memory. We propose TrajRAG, which retrieves geometric–semantic experiences from past trajectories and enables long-term memory through the continual accumulation of episodic experiences.

2.2. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) [21] was initially introduced to enhance large language models (LLMs) by retrieving relevant document fragments, thereby injecting domain-specific knowledge to improve factuality and relevance. Traditional RAG frameworks embed user queries and document chunks into a shared vector space, retrieving the top-k semantically similar passages to expand the model's context window [4, 15]. Recent RAG works introduced iterative retrieval and knowledge-graph-based methods. Models such as GraphRAG [11] and LightRAG [17] extract entities and relations to construct graph structures, enabling more holistic and globally-aware retrieval.

Recently, researchers have begun adapting RAG to embodied intelligence [1, 39, 43]. EmbodiedRAG [1] and Embodied-RAG [43] build non-parametric, structured memory for the current scene in an online fashion, organizing scene information via RAG to enhance retrieval efficiency and accuracy over both the present environment and previously collected historical scenes. NavRAG [39] proposes a retrieval-augmented framework that automatically generates user-oriented navigation instructions from structured simulator data without manual annotation, producing large-scale, high-quality instruction–trajectory pairs for Vision–Language–Navigation training, assuming that comprehensive scene information is already available.

Existing embodied-RAG approaches retrieve within the current scene and cannot transfer knowledge or leverage historical experience. TrajRAG instead continually indexes knowledge along novel trajectories and uses a coarse-to-fine retrieval pipeline to inject relevant past experience into the LLM, enabling cross-scene informed decisions.

3. Method

3.1. TrajRAG Definition

Raw episodic memories, represented as trajectories of RGB-D observations $\mathcal{T} = \{(I_t, D_t, \mathbf{p}_t)\}_{t=1}^{T}$ (where I_t, D_t, and $\mathbf{p}_t$ denote RGB, depth, and agent pose, respectively), are inherently redundant due to local revisits, small inter-frame variations, and cross-episode overlaps. To achieve compact yet semantically meaningful memory, we propose the topological–polar (topo-polar) trajectory, which transforms continuous raw observations into a structured representation.

Topo-polar trajectory. Given the raw observations $\mathcal{T} = \{(I_t, D_t, \mathbf{p}_t)\}_{t=1}^{T}$, the map module first constructs an open-vocabulary semantic map $\mathcal{M}_t$ (see Sec. 3.3). Based on the explored regions in the semantic map $m_t^{free} \in \mathcal{M}_t$, a morphological thinning operation $\mathcal{S}(\cdot)$ is applied to obtain a one-pixel-wide skeleton, formally $\mathcal{G}_{\text{skel}} = \mathcal{S}(m_t^{free})$.

Then, candidate topological nodes are defined by selecting pixels whose 8-neighborhood contains at least three connected components:

$$\mathcal{V}_{\text{cand}} = \{v \in \mathcal{G}_{\text{skel}} \mid |\mathcal{N}_8(v)| \geq 3\} \tag{1}$$

Since candidate nodes $\mathcal{V}_{\text{cand}}$ may exhibit spatial redundancy from local skeleton noise or nearby junctions, we apply non-maximum distance suppression, merging nodes within a threshold distance to form the final node set $\mathcal{V} = \{v_k\}_{k=1}^{N_v}$ that defines the discrete topology of the explored region.

For each node v_k, we adopt a polar-coordinate representation to characterize its surrounding geometric-semantic context. Since different trajectories arise from varying agent initial poses, the absolute positions of observed objects are inconsistent across runs. In contrast, the polar representation captures relative spatial relationships, making it more suitable for this setting. Moreover, the division of space into multiple angular sectors allows for a clear distinction of node-specific characteristics, which facilitates subsequent node matching. Specifically, for each node v_k, we cast polar sampling rays centered at its position in $\mathcal{M}_t$. Each ray $\mathcal{R}(\theta)$, with direction $\theta \in [0, 2\pi)$ and step $\Delta\theta = 30°$, searches for the first non-free pixel (i.e., an object, obstacle, or unknown region) within range R. The sampling function is defined as:

$$\phi_k(\theta) = \begin{cases} c, & \text{if hits object } c \\ \text{obstacle}, & \text{if hits obstacle} \\ \text{unknown}, & \text{if hits unknown region} \\ \text{free}, & \text{if no hit within } R \end{cases} \tag{2}$$

Before sampling, each semantic channel is morphological dilation to prioritize semantic pixels are prioritized over obstacles, preventing incorrect associations caused by thin wall boundaries. The sector vector around v_k is:

$$\mathbf{s}_k = [\phi_k(\theta_1), \phi_k(\theta_2), \ldots, \phi_k(\theta_{12})] \tag{3}$$

Sector indices increase counterclockwise relative to the agent's left-facing direction. This vector $\mathbf{s}_k$ encodes local semantic–geometric structure, and episode-specific world coordinate $\mathbf{p}_k$ of each node v_k is recorded for alignment.

For each segment $\tau = (I_t, D_t, \mathbf{p}_t)$ in the observations $\mathcal{T} = \{(I_t, D_t, \mathbf{p}_t)\}_{t=1}^{T}$, we normalize it by assigning it to its nearest node according to the agent's pose measured by Euclidean distance:

$$v_t^* = \arg\min_{v_k \in \mathcal{V}} \|\mathbf{p}_t - \mathbf{p}_k\|_2 \tag{4}$$

Consecutive identical nodes are first merged to eliminate short-term revisits within the same topological region. To further remove long-term cycles, we prune loops (denoted as f_{PL}) by retaining only the last occurrence of each node in temporal order, i.e., when a node reappears, all nodes between its previous and latest occurrence are discarded. This ensures that the stored trajectories are loop-free and efficient, avoiding redundancy within each individual trajectory. The resulting sequence defines a topo-polar trajectory $\mathcal{T}_{\text{tp}}$ for the raw observations $\mathcal{T}$:

$$\mathcal{T}_{\text{tp}} = (\mathcal{V}, \mathcal{E}), \quad \mathcal{V} = f_{\text{PL}}(\{v_t^*\}) \tag{5}$$

where $\mathcal{E}$ denotes directed edges between consecutive nodes.

TrajRAG structure. TrajRAG adopts a hierarchical chunking architecture to organize the topo-polar trajectories, consisting of the following key components:

1) Chunk. Each chunk χ_i corresponds to a topo-polar trajectory, which includes a topological–polar trajectory $\mathcal{T}_{\text{tp}}^i$, a natural language description $L(\mathcal{T}_{\text{tp}}^i)$, and the corresponding trajectory embedding $\mathbf{z}_i = f_{\text{E}}(\mathcal{T}_{\text{tp}}^i)$.

2) Coarse index. Chunks with high geometric–semantic similarity are grouped and merged into a unified topo-polar summary. The similarity between two chunks is measured by the geometric–semantic consistency between their topo-polar trajectories (see Sec. 3.2 for details). The topo-polar summary is defined as $\mathcal{G}_{\text{sum}} = (\mathcal{V}_{\text{uni}}, \mathcal{E}_{\text{mrg}})$, where $\mathcal{V}_{\text{uni}}$ and $\mathcal{E}_{\text{mrg}}$ denote the fused nodes and edges after removing duplicates. The summary graph $\mathcal{G}_{\text{sum}}$ captures the group's overall geometric–semantic layout and serves as a **coarse index** during retrieval. Given a query, TrajRAG first retrieves a set of relevant summaries before performing fine-grained trajectory search. This process narrows down the trajectory search to relevant groups, enhancing the relevance of the retrieved historical trajectory knowledge to the current scene.

3) Fine index. Since the number of chunks is much larger than the number of topo-polar summaries, each chunk is encoded as a hidden vector embedding to enable efficient retrieval during real-time navigation. Each topo-polar node v_k is represented by its sector vector $\mathbf{s}_k$. We employ an encoder-only transformer (e.g., DistilBERT) to obtain node embeddings, as its bidirectional attention and position-insensitive properties make it well-suited for encoding polar representations: $\mathbf{h}_k = \mathcal{E}_{\text{node}}(\mathbf{s}_k)$. Given the ordered sequence of node embeddings $[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L]$, we encode their sequential dependencies using a decoder-only transformer (e.g., DistilGPT2) denoted as $\mathcal{D}_{\text{traj}}$. The order-sensitive $\mathcal{D}_{\text{traj}}$ captures temporal correlations along the traversal order. We then concatenate the final token representation with the goal semantic embedding $\mathbf{h}_{\text{g}}$ to form the trajectory embedding:

$$\mathbf{z} = f_{\text{E}}(\mathcal{T}_{\text{tp}}) = \mathbf{h}'_L \oplus \mathbf{h}_{\text{g}}, \quad [\mathbf{h}'_1, ..., \mathbf{h}'_L] = \mathcal{D}_{\text{traj}}([\mathbf{h}_1, ..., \mathbf{h}_L]) \tag{6}$$

where $\oplus$ denotes concatenation. We optimize the trajectory encoder f_{E} (including both the encoder-only and decoder-only transformers) using a contrastive learning objective:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j^+)/\tau)}{\sum_k \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{7}$$

where $\text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \mathbf{z}_j / (\|\mathbf{z}_i\| \|\mathbf{z}_j\|)$ denotes cosine similarity and τ is a temperature parameter. Positive pairs $(\mathbf{z}_i, \mathbf{z}_j^+)$ are sampled from trajectories within the same topological group or sharing the same navigation goal, while negatives are randomly drawn from different groups. Note that during training, the encoder-only transformer is initialized with pretrained weights and kept frozen, i.e., its parameters are not updated.

3.2. Incremental Construction

During navigation, TrajRAG is incrementally constructed. Given a new episode $\mathcal{T} = \{(I_t, D_t, \mathbf{p}_t)\}$, we first derive its topo-polar trajectory $\mathcal{T}_{\text{tp}}$ as defined in Eq. 5. The resulting trajectory is then matched against existing topo-polar summaries for potential integration.

Semantic matching. We compute the semantic similarity matrix $\mathbf{S} \in \mathbb{R}^{N_{\text{new}} \times N_{\text{sum}}}$ between the nodes of $\mathcal{T}_{\text{tp}}$ and those in each summary graph:

$$S_{ij} = \max_{\Delta\theta} \text{sim}\big(\text{Rot}(\mathbf{s}_i, \Delta\theta), \mathbf{s}_j\big) \tag{8}$$

where $\text{Rot}(\mathbf{s}_i, \Delta\theta)$ cyclically rotates the 12-dimensional sector vector to compensate for heading misalignment, which ensures that nodes with similar surrounding semantics are matched even when the agent's facing direction differs across trajectories. After computing $\mathbf{S}$, we adopt a bidirectional K-nearest-neighbor (mutual KNN) strategy to select consistent correspondences, i.e., a pair (v_i, v_j) is retained only if v_j ranks among the top-K nearest neighbors of v_i and vice versa. This symmetric constraint filters out spurious matches while preserving high-confidence node pairs for subsequent geometric matching.

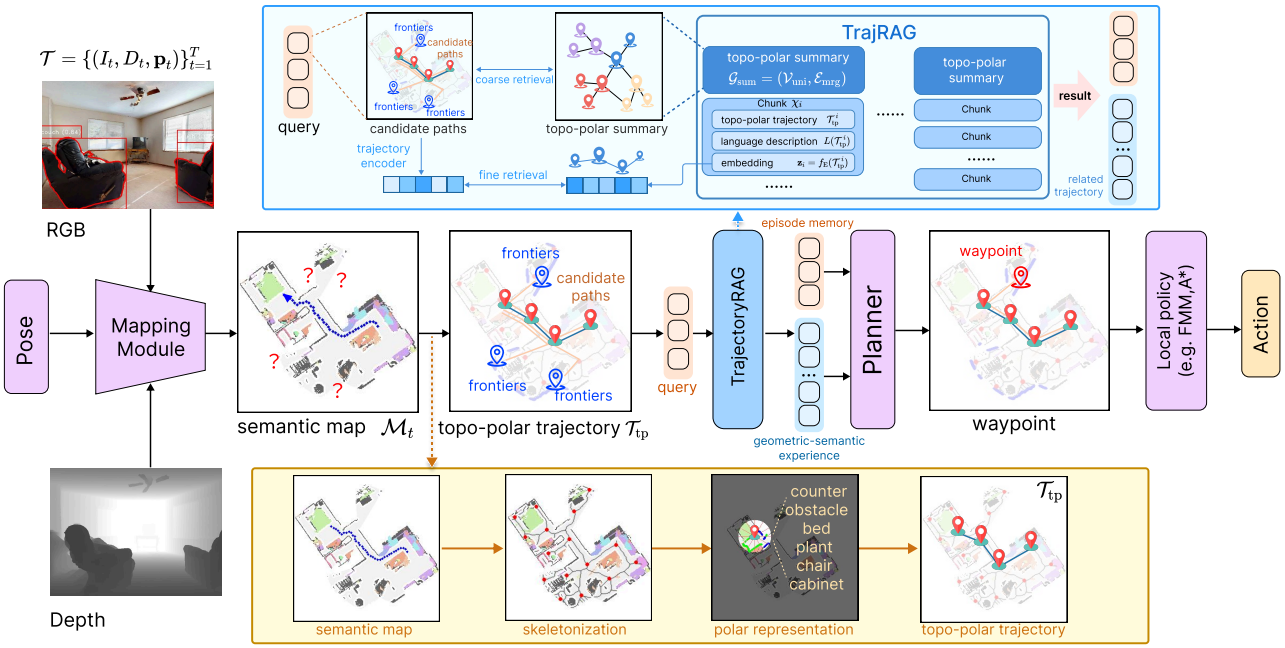

Figure 2. **Navigation Framework of TrajRAG.** The agent incrementally maintains a semantic map during navigation. Based on this map, we convert the explored area into a topo-polar trajectory. Candidate trajectories are then generated according to the potential frontiers. For each candidate, TrajRAG retrieves relevant experiences to help the planner estimate which trajectory can reach the goal more efficiently. A coarse-to-fine retrieval strategy is adopted to accelerate retrieval.

Geometric matching. Using the world coordinates $\{\mathbf{p}_k\}$ stored in matched nodes, we apply RANSAC [13] estimation to find a robust geometric transformation $\mathbf{T} \in SE(2)$ aligning the new trajectory to the summary:

$$\mathbf{T} = \arg\min_{\mathbf{T}} \sum_{(v_i, v_j) \in \mathcal{C}} \rho\big(\|\mathbf{T}\mathbf{p}_i - \mathbf{p}_j\|\big) \tag{9}$$

where $\mathcal{C}$ is the set of matched node pairs and $\rho(\cdot)$ is a robust penalty. We set the geometric similarity score to the inlier ratio $|\mathcal{C}_{\text{in}}|/|\mathcal{C}|$ returned by RANSAC, with $\mathcal{C}$in defined as the set of consistent pairs under the estimated transformation. If a valid $\mathbf{T}$ is found, $\mathcal{T}_{\text{tp}}$ is merged into the corresponding summary group, extending $\mathcal{V}_{\text{unique}}$ and $\mathcal{E}_{\text{merged}}$ with newly discovered nodes and edges.

Redundancy checking and TrajRAG update. After semantic and geometric matching, we further compare the new trajectory against existing members within each group to prevent redundant storage. Specifically, if two trajectories share the same navigation goal and their node sequences exhibit a strict containment relationship (i.e., one trajectory's node list is a subsequence of the other with identical order), the shorter one is regarded as redundant and discarded. Otherwise, the new trajectory enriches the group by contributing additional geometric-semantic context.

Finally, the group's summary graph $\mathcal{G}_{\text{sum}}$ is updated by integrating the new trajectory. Given the estimated transformation $\mathbf{T} \in SE(2)$, the trajectory's nodes and edges are first transformed into the global frame. Newly discovered nodes and edges absent from the existing summary are then identified and added, allowing $\mathcal{G}$sum to expand incrementally as navigation progresses. Additionally, if a new topo-polar trajectory cannot be matched with any existing group in terms of geometric and semantic similarity, a new group is created, and this trajectory is registered as the group's initial topo-polar summary and added to TrajRAG.

We collected navigation trajectories on the training splits of HM3Dv1 [29] and MP3D [5] to build TrajRAG. From the HM3Dv1 ObjectNav Train dataset (3.9M episodes, 6 categories, 80 scenes), we uniformly sample over 200k trajectories, ensuring exhaustive coverage of all floors and categories per scene. Similarly, from the MP3D ObjectNav Train dataset (2.6M episodes, 21 categories, 56 scenes), we gather more than 150k trajectories, again achieving coverage of all floor layouts and object categories in every scene.

3.3. Navigation with TrajRAG

In the zero-shot ObjectNav task, an agent is initialized in an unseen environment and must locate an object from a specified semantic category (e.g., chair) through active exploration, using real-time RGB-D observations and sensor pose. Success requires issuing `stop` within a distance threshold (e.g., 1 m) of the target and within the step budget (e.g., 500 steps).

As shown in Fig. 2, navigation with TrajRAG comprises constructing the current-scene trajectory and retrieving and exploiting trajectory knowledge from the knowl-

edge base. At timestep t of an episode, we utilize a **semantic map module** to incrementally construct a semantic map based on RGB-D observations and agent poses. GroundingDINO [24] detects object categories, and MobileSAM [51] provides more precise segmentation masks. The semantic map is formulated as $\mathcal{M}_t \in \mathbb{R}^{(2+N_o)\times H\times W}$, where N_o denotes the number of object categories, which can be arbitrarily specified. The remaining two channels correspond to obstacles and explored areas.

Based on the semantic map, a topo-polar trajectory is constructed following Eq. 5. From the obstacles and explored areas of the semantic map, frontiers are extracted, defined as the boundaries between explored and unexplored areas. Each frontier is represented by its geometric centroid and filtered to remove short and noisy segments. Similar to the node in topo-polar trajectory, each frontier is further represented by a sector vector as defined in Eq. 3. A breadth-first search (BFS) is performed over the current topo-polar trajectory to compute paths leading to all valid frontiers. This process yields candidate paths $\{\Pi_i\}_{i=0}^{N}$.

To provide the navigation planner with experience on what typically follows a candidate path (i.e., which path may lead closer to the goal), we perform hierarchical retrieval of these candidate trajectories within TrajRAG to locate the most relevant experiences. Specifically, a coarse retrieval stage first identifies top-ranked similar groups by comparing stored topo-polar summaries with the current candidate topo-polar trajectory, using both semantic and geometric matching results (as described in Eq. 8 and Eq. 9).

Within the selected groups, a fine-grained retrieval stage encodes the current topo-polar trajectory with the trajectory encoder f_E, and retrieves the top-K most similar trajectories. For each retrieved trajectory $\mathcal{T}_j$, we utilize its corresponding language description $\mathcal{L}_j$ as the experience. The current candidate paths $\{\Pi_i\}$ and retrieved experience $\{\mathcal{L}_j\}$ are jointly provided to a navigation planner (e.g., LLM-based model) to select the optimal navigation path Π^*. Whenever the agent reaches a new topo-polar node, the planner performs an inference. The frontier corresponding to Π^* is then set as the next waypoint for the local policy. Our local policy computes the shortest path from the current position to the waypoint using the A* algorithm, and then determines the next action for the agent by following this trajectory through its discrete action space.

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. For **zero-shot ObjectNav**, we evaluate our method on the HM3Dv1 [29], HM3Dv2 [44] , and MP3D [5] within the Habitat simulator [32]. Following previous works [46, 52], we report Success Rate (SR) and Success weighted by normalized inverse Path Length (SPL) as the evaluation metrics.

Table 1. Ablation on node representations in HM3Dv1. TNA: textual neighbor aggregation; TPS-G: topo-polar sector geometry; TPS-S: topo-polar sector semantics.

TNA	TPS-G	TPS-S	SR(%)	SPL(%)
✓			53.9	25.7
	✓		48.1	22.3
		✓	57.3	30.6
	✓	✓	**61.7**	**33.2**

Table 2. Ablation study on retrieval strategy in HM3Dv1. TE: text embedding; SE: our sequence embedding.

Coarse	Fine	SR(%)	SPL(%)
	SE	54.3	25.6
✓	TE	57.8	29.7
✓	SE	**61.7**	**33.2**

Implementation Details. During navigation, we employ GroundingDINO [24] and MobileSAM [51] to detect and segment images and generate a semantic map, while Qwen3-32B [35] is used as the planner to select the path. In the free map, topological keypoints are deduplicated with a distance threshold $d_{\min}$ of 0.5 meters. For generating textual descriptions of each sector in the semantic map, a sampling radius of 1.5 meters is employed. To ensure fair evaluation and prevent test-set leakage, TrajRAG is pre-built solely on training data and kept frozen during testing, although it natively supports dynamic test-time updates.

4.2. Evaluation Results

Ablation on Node Representations. Tab. 1 presents the ablation on node representations. Our full model, which integrates both Topo-Polar Sector geometry and semantics (TPS-G & TPS-S), achieves the best performance, demonstrating the complementarity of geometric and semantic cues. In contrast, the Textual Neighbor Aggregation (TNA) baseline (Line 1), which aggregates textual descriptions without spatial order for trajectory description and node embedding, performs worse. Furthermore, using only Topo-Polar Sector-Geometry (TPS-G) yields the weakest performance, as geometry alone lacks semantic cues. Employing only Topo-Polar Sector-Semantics (TPS-S) surpasses both TNA and TPS-G, demonstrating the importance of structured semantic description and encoding. In summary, our topo-polar format provides a clear advantage by preserving spatial relationships and structuring semantic information.

Ablation Study on the Retrieval Strategy. Tab. 2 evaluates the impact of each component in our retrieval strategy. TrajRAG employs a coarse-to-fine representation and retrieval pipeline. The coarse stage retrieves scenes based on topo-polar geometric-semantic relationships, while the fine stage retrieves specific trajectories using embeddings. In the first row, we remove coarse-level matching between

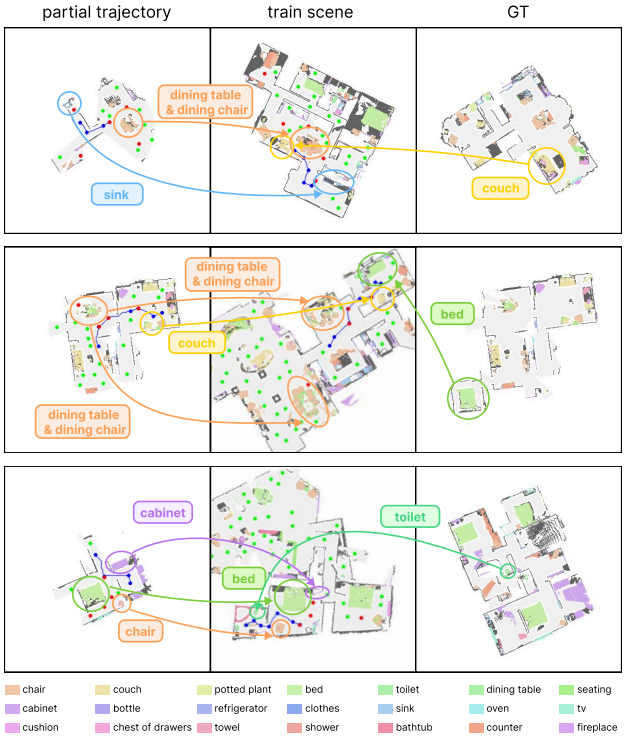

Figure 3. **Retrieved Examples**. Given a partial trajectory, TrajRAG uses its accumulated keypoints (green) to retrieve a layout-consistent scene and the assistant goal-reaching trajectory (blue). For pre-retrieval, the top-5 matched keypoints within the visible map are highlighted in red.

the current scene and database summaries, retrieving trajectories directly with sequence embeddings. This leads to a clear performance drop, confirming that pre-retrieval based on scene layout is essential for filtering irrelevant contexts effectively. The second row retains scene-level matching but replaces our trained trajectory encoder with a pre-trained text model to encode structured trajectory descriptions. This also degrades results, underscoring the limitation of generic text embeddings compared with our dedicated sequence encoder in capturing trajectory semantics. The best performance is achieved by integrating both retrieval stages, which validates the design of TrajRAG's hierarchical retrieval system.

Qualitative Analysis. To better understand how our TrajRAG supports navigation decisions, we visualize the trajectory retrieval results in Fig. 3. The keypoints accumulated from the partial trajectory are used as the query for retrieval. We highlight the top-5 keypoints pre-retrieved based on scene layout (shown in red) as well as the retrieved trajectory leading toward the target object (shown in blue). As illustrated, the retrieved trajectory exhibits a strong latent spatial alignment with the test environment. Although it is retrieved from a different scene, its geometric structure and goal-directed progression are highly compatible with the query layout. This alignment indicates that TrajRAG captures scene-invariant spatial relations and retrieve trajectories that reflect how agents typically approach similar targets. Consequently, the retrieved trajectory provides meaningful guidance for downstream navigation, demonstrating the effectiveness of TrajRAG.

Table 3. Comparison with Other Types of RAG in HM3Dv1.

Method	Retrieval	Content	SR(%)	SPL(%)
TrajTextRAG	text embedding	text description	53.3	25.6
GraphRAG	graph embedding	textual scene graph	55.2	30.7
TrajRAG(Ours)	hierarchical retrieval	topo-polar trajectory	**61.7**	**33.2**

Table 4. Cross-dataset evaluation results on HM3Dv1 and MP3D.

	HM3Dv1 train	MP3D train	SR(%)	SPL(%)
HM3Dv1 val	✓		61.7	33.2
		✓	59.8	31.6
	✓	✓	**62.5**	**33.9**
MP3D val	✓		39.4	16.2
		✓	41.7	17.6
	✓	✓	**42.6**	**18.0**

In addition, Fig. 4 shows that the agent can effectively choose the right waypoint based on the TrajRAG.

Comparison with other types of RAG. As shown in Tab. 3, we compare our method with two alternative RAG formulations. Our TrajRAG employs a hierarchical retrieval structure that accounts for scene layout, and retrieves topo-polar trajectories to assist the LLM in selecting suitable candidate paths. TrajTextRAG encode historical trajectories as textual sequences and perform retrieval in a semantic embedding space. The method retrieves trajectory descriptions based on current candidate paths and uses them to guide the LLM. However, its performance is lower than ours, suggesting that the textual semantic space struggles to distinguish sequential relationships in trajectories and lacks scene-aware pre-retrieval, leading to suboptimal matching with the current scene. GraphRAG construct scene graphs where objects serve as nodes and spatial relations (e.g., "next to") as edges. During navigation, relevant scene graphs are retrieved to prompt the LLM for frontier scoring. This variant performs worse than TrajRAG. A plausible explanation for this gap is that the retrieved scene graphs often contain generic information and lack trajectory continuity, thus providing limited actionable knowledge for the LLM.

Cross-dataset Evaluation. Table 4 evaluates the cross-dataset generalization of our retrieval corpus. The results demonstrate that a retrieval corpus built from one dataset (e.g., HM3Dv1) generalizes effectively to guide navigation in another (e.g., MP3D), and vice versa, with only a minor performance drop compared to in-domain retrieval. This indicates that the semantic and topological representations in our retrieval corpus capture universal navigation cues that transfer across different environments. Furthermore, com-

Figure 4. **Navigation with TrajRAG**. The left column shows the agent's ego-view RGB images. The middle column presents the skeletonization and detected keypoints (red) on the skeleton map, where the location icon indicates the ground-truth location of the target ("bed") for visualization purposes only; the target location is unknown to the agent during navigation. The right column illustrates the agent's traversed trajectory from the start to the pre-navigation position (light gray), candidate trajectories from the current location to various frontiers (light-colored, with blue dots indicating frontiers), and the trajectory selected by the model (burgundy).

Table 5. Comparisons with the related works on MP3D, HM3Dv1 and HM3Dv2 datasets. "OV" denotes if the method supports open-vocabulary object goals.

Method	OV	MP3D		HM3Dv1		HM3Dv2	
		SR(%)	SPL(%)	SR(%)	SPL(%)	SR(%)	SPL(%)
DD-PPO[41]	×	8.0	1.8	27.9	14.2	-	-
SemExp [6]	×	36.0	14.4	-	-	-	-
SGM [55]	×	37.7	14.7	60.2	30.8	-	-
T-Diff [49]	×	39.6	15.2	-	-	-	-
GOAL [22]	×	41.7	15.5	-	-	-	-
ZSON [26]	✓	15.3	4.8	25.5	12.6	-	-
PSL [33]	✓	-	-	42.4	19.2	-	-
ESC [59]	✓	28.7	14.2	39.2	22.3	-	-
VLFM [47]	✓	36.4	17.5	52.5	30.4	63.6	32.5
VoroNav [42]	✓	-	-	42.0	26.0	-	-
InstructNav [25]	✓	-	-	-	-	58.0	20.9
SG-Nav [45]	✓	40.2	16.0	54.0	24.9	49.6	25.5
UniGoal [46]	✓	41.0	16.4	54.5	25.1	-	-
ApexNAV [52]	✓	39.2	17.8	59.6	33.0	76.2	38.0
BeliefMapNav [60]	✓	37.3	17.6	61.4	30.6	-	-
TrajRAG (Ours)	✓	**42.6**	**18.0**	**62.5**	**33.9**	**78.1**	**40.2**

bining the retrieval corpora from both datasets consistently enhances performance on both validation sets, achieving the best results. This shows that our method not only transfers well but also benefits from a more diverse and comprehensive set of navigation experiences.

4.3. Comparison with SOTA Methods

Zero-shot ObjectNav. We compare our method with prior works on three challenging datasets in Tab. 5, where competing methods are categorized by their support for open-vocabulary (OV) object goals. Our proposed TrajRAG establishes new state-of-the-art performance across all benchmarks. It achieves the highest Success Rate (SR) and Success weighted by Path Length (SPL) on the MP3D, HM3Dv1, and HM3Dv2.

TrajRAG's superior performance stems from its core innovation of leveraging historical navigation experiences. By retrieving the most relevant trajectories from an external knowledge base that accounts for both scene context and navigational relevance, the agent can make more informed decisions. This retrieval-augmented strategy guides agent's exploration and long-horizon planning, helping it avoid myopic behavior and navigate more efficiently towards open-vocabulary goals in complex, unseen environments.

5. Conclusion

In this paper, we introduced Trajectory RAG (TrajRAG), a retrieval-augmented generation framework for zero-shot ObjectNav that enhances navigation reasoning by retrieving geometric–semantic experiences. To ensure compact and efficient storage, we propose a topological-polar (topo-polar) trajectory representation that enables self-checking by detecting and pruning redundant segments, while supporting duplicate detection during trajectory integration. Furthermore, TrajRAG employs a hierarchical chunking architecture to enable efficient retrieval of relevant experiences. Experiments on MP3D, HM3D-v1, and HM3D-v2 demonstrate that TrajRAG effectively retrieves geometric–semantic experiences and improves navigation performance in zero-shot ObjectNav.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62125207, Grant 62495084, Grant 62272443, and Grant U23B2012, in part by the Beijing Natural Science Foundation under Grant L242020, in part by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number BX20250391, and in part by the Suzhou Science and Technology Plan Project under grant SYG2024082.

References

[1] Meghan Booker, Grayson Byrd, Bethany Kemp, Aurora Schmidt, and Corban Rivera. Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning. *arXiv preprint arXiv:2410.23968*, 2024. 3

[2] Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE, 2024. 1

[3] Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. Cognav: Cognitive process modeling for object goal navigation with llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9550–9560, 2025. 3

[4] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024. 3

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017. 5, 6

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 8

[7] Shizhe Chen, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. Object goal navigation with recursive implicit maps. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7089–7096. IEEE, 2023. 1

[8] Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Jiagui Tang, Chengju Liu, and Qijun Chen. Search for or navigate to? dual adaptive thinking for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8250–8259, 2023. 1

[9] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 1

[10] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *European Conference on Computer Vision*, pages 19–34, 2020. 1

[11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 3

[12] Thomas W Elston and Joni D Wallis. Context-dependent decision-making in the primate hippocampal–prefrontal circuit. *Nature Neuroscience*, 28(2):374–382, 2025. 2

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[14] 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*, pages 23171–23181, 2023. 1, 2

[15] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, 2023. 3

[16] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022. 1

[17] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024. 3

[18] Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, Yi Fang, et al. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. *Advances in Neural Information Processing Systems*, 37:39386–39408, 2024. 2, 3

[19] Shuqiang Jiang, Sixian Zhang, Shida Tao, Xihong Zhu, Tianliang Qi, and Xinhang Song. Self model for embodied artificial intelligence. *Journal of Computer Science and Technology*, 2026. 2

[20] 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*, pages 16373–16383, 2024. 1

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp

tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3

[22] Badi Li, Renjie Lu, Yu Zhou, Jingke Meng, and Wei-Shi Zheng. Distilling LLM prior to flow model for generalizable agent's imagination in object goal navigation. *Advances in Neural Information Processing Systems*, 2025. 8

[23] John Lisman and A David Redish. Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1193–1201, 2009. 2

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024. 6

[25] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment, 2024. 8

[26] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022. 2, 8

[27] Ida Momennejad. Learning structures: predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, 32:155–166, 2020. 2

[28] Yael Niv. Learning task-state representations. *Nature neuroscience*, 22(10):1544–1553, 2019. 2

[29] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 5, 6

[30] 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*, pages 18890–18900, 2022. 1

[31] Jon W Rueckemann, Marielena Sosa, Lisa M Giocomo, and Elizabeth A Buffalo. The grid code for ordered experience. *Nature Reviews Neuroscience*, 22(10):637–649, 2021. 2

[32] Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9338–9346. IEEE, 2019. 6

[33] Xinyu Sun, Lizhao Liu, Hongyan Zhi, Ronghe Qiu, and Junwei Liang. Prioritized semantic learning for zero-shot instance navigation. In *European Conference on Computer Vision*, pages 161–178. Springer, 2024. 8

[34] Wenbo Tang, Justin D Shin, and Shantanu P Jadhav. Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *Cell reports*, 42(3), 2023. 2

[35] Qwen Team. Qwen3 technical report, 2025. 6

[36] Zihan Wang and Gim Hee Lee. g3d-lf: Generalizable 3d-language feature fields for embodied tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14191–14202, 2025. 2

[37] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762, 2024. 1

[38] Zihan Wang, Seungjun Lee, and Gim Hee Lee. Dynam3d: Dynamic layered 3d tokens empower vlm for vision-and-language navigation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[39] Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *arXiv preprint arXiv:2502.11142*, 2025. 3

[40] James CR Whittington, David McCaffary, Jacob JW Bakermans, and Timothy EJ Behrens. How to build a cognitive map. *Nature neuroscience*, 25(10):1257–1272, 2022. 2

[41] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Ddppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*. 8

[42] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. In *International Conference on Machine Learning*, pages 53757–53775. PMLR, 2024. 2, 3, 8

[43] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*, 2024. 3

[44] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023. 6

[45] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in Neural Information Processing Systems*, 37:5285–5307, 2024. 2, 8

[46] 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 Computer Vision and Pattern Recognition Conference*, pages 19057–19066, 2025. 2, 3, 6, 8

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

ternational Conference on Robotics and Automation (ICRA), pages 42–48. IEEE, 2024. 2, 3, 8

[48] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023. 3

[49] Xinyao Yu, Sixian Zhang, Xinhang Song, Xiaorong Qin, and Shuqiang Jiang. Trajectory diffusion for objectgoal navigation. *Advances in Neural Information Processing Systems*, 37:110388–110411, 2024. 8

[50] Albert J Zhai and Shenlong Wang. Peanut: Predicting and navigating to unseen targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10926–10935, 2023. 1

[51] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 6

[52] Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. Apexnav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *IEEE Robotics Autom. Lett.*, 10(11):11530–11537, 2025. 6, 8

[53] Sixian Zhang, Weijie Li, Xinhang Song, Yubing Bai, and Shuqiang Jiang. Generative meta-adversarial network for unseen object navigation. In *European Conference on Computer Vision*, pages 301–320. Springer, 2022. 1

[54] Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, and Shuqiang Jiang. Layout-based causal inference for object navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10792–10802. IEEE, 2023. 1

[55] Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16414–16425, 2024. 1, 8

[56] Sixian Zhang, Xinhang Song, Xinyao Yu, Yubing Bai, Xinlong Guo, Weijie Li, and Shuqiang Jiang. Hoz++: Versatile hierarchical object-to-zone graph for object navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[57] Sixian Zhang, Xinyao Yu, Xinhang Song, Yiyao Wang, and Shuqiang Jiang. Function-centric bayesian network for zero-shot object goal navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19535–19545, 2025. 2

[58] 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*, pages 7641–7649, 2024. 1, 3

[59] 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 *International Conference on Machine Learning*, pages 42829–42842. PMLR, 2023. 2, 8

[60] Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation. *Advances in Neural Information Processing Systems*, 2025. 2, 3, 8