Chain-of-Search: Parameter-Efficient Reasoning for Zero-Shot Object Navigation

Hanrui Chen[1*], Liqi Yan[1*†], Qifan Wang[2], Jianhui Zhang[1], Fangli Guan[1], Pan Li[1†]

[1]Hangzhou Dianzi University
[2]Meta AI
ylq@hdu.edu.cn, lipan@ieee.org

Abstract

Zero-shot object navigation tasks agents with locating target objects in unseen environments—a core capability of embodied intelligence. While recent vision-language navigation methods leverage Large Language Models (LLMs) for multimodal reasoning, they suffer from two key limitations: (1) semantic misalignment between language-grounded maps and real-world layouts, and (2) inefficiency due to LLMs' lack of specialization for navigation-specific tasks. To address these challenges, we propose **Chain-of-Search (CoS)**, a novel parameter-efficient framework that enables human-like decision-making via iterative semantic reasoning. **First**, CoS replaces traditional global maps with an optimal-benefit multi-map construction that continuously balances expected gain and cost throughout the navigation process. **Second**, we introduce a Parameter-Efficient Intent Aligner (PEIA), trained via a prompt-guided paradigm to align directional decisions with navigation intent. PEIA injects semantic cues into benefit-aware maps, enabling more rational and goal-consistent exploration. **Finally**, a Reflection-Guided Destination Verifier (RDV) confirms whether the target is reached via language-driven reasoning and corrects potential errors through self-reflection. CoS achieves state-of-the-art performance on HM3D (+2.8% SR) and MP3D (+1.2% SR) without relying on LLMs, demonstrating the effectiveness of lightweight, reasoning-centered navigation.

Introduction

Object navigation (Majumdar et al. 2022; Yuan et al. 2024) in unknown environments presents a multifaceted challenge for robots, relying on a combination of visual perception (Cui et al. 2021), semantic understanding of the surroundings (Yan et al. 2022b), and spatial reasoning (Yan et al. 2024). Effective navigation requires the seamless integration of internal knowledge, real-time sensory observations, and intelligent decision-making. Humans navigate unfamiliar environments by leveraging semantic priors and spatial context—for example, knowing that pillows are likely near beds, not toilets, and that bathrooms are often adjacent to bedrooms. Such intuition also reflects spatial proximity

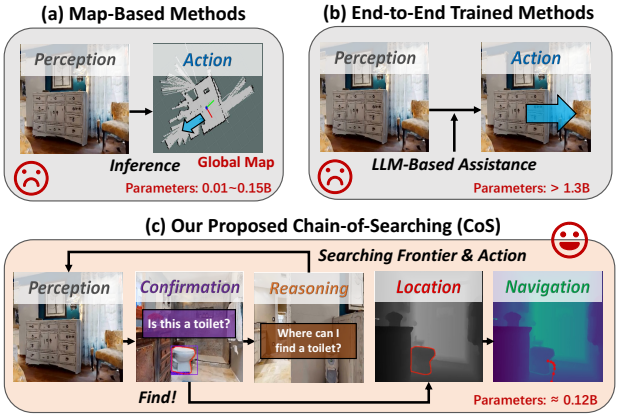

Figure 1: Comparison with existing zero-shot object navigation methods. Map-based approaches lack semantic understanding, while end-to-end LLM methods are computationally expensive. Our Chain-of-Search (CoS) addresses this by enabling efficient reasoning through lightweight vision-language model tuning, achieving strong generalization without relying on large models.

preferences, e.g., moving toward the nearest bedroom when searching for a bed. Emulating this human-like reasoning is crucial for effective robot navigation.

Recent vision-language navigation (VLN) methods take initial steps in this direction by training agents to reason over multimodal inputs and make context-aware decisions grounded in language and visual perception. Among existing VLN models, zero-shot solutions offer a convenient pathway for future deployment in unknown environment. The impressive performance of Large Language Models (LLMs) (Cai et al. 2024; Wu et al. 2024) and Vision-Language Models (VLMs) (Yuan et al. 2024; Nie et al. 2025) demonstrates their potential to reason about objects beyond the current field of view in zero-shot settings. These methods rely on object detectors and language models, such as BERT or LLMs (Shah et al. 2023; Zhou et al. 2023; Yu, Kasaei, and Cao 2023; Dorbala, Mullen, and Manocha 2024; Cai et al. 2024). For example, NavGPT (Zhou, Hong, and Wu 2024) and NavGPT-2 (Zhou et al. 2024) utilize LLMs to guide agents through semantic scene understanding

*These authors contributed equally.
†Corresponding authors: Liqi Yan, Pan Li.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

and goal alignment. Subsequent approaches leverage large vision-language models, such as BLIP2 (Li et al. 2023), to compute cosine similarity as a semantic score (Yokoyama et al. 2024), highlighting the impact of LLMs and VLMs on zero-shot navigation with semantic reasoning.

However, as illustrated in Fig. 1, existing LLM-based zero-shot navigation methods face two critical challenges. ① **Semantic loss in global map construction undermines navigation reliability.** These methods typically build semantic maps by fusing visual observations with prior knowledge. Yet, due to limited field-of-view and imperfect language-grounded perception, the resulting maps often misrepresent the actual environment, leading to cumulative errors. Such misalignment increases the risk of navigating toward irrelevant regions, reducing task efficiency and success rate. ② **LLMs lack specialization for navigation tasks and suffer from computational inefficiency.** Although LLMs offer strong general reasoning, their massive scale and language-centric design are poorly aligned with the fine-grained, spatial-temporal demands of embodied navigation. Without domain adaptation, their reasoning remains inefficient and frequently suboptimal. These limitations highlight the need for lightweight, navigation-aware alternatives that can perform efficient semantic reasoning without relying on full-scale LLMs.

To address the above challenges, we propose **Chain-of-Search (CoS)**, a zero-shot object navigation framework that requires no training on object navigation datasets. **First**, CoS constructs a fused map that jointly models the *benefit* and *cost* of exploration, enabling more informed direction prediction beyond conventional global or purely semantic maps (Yokoyama et al. 2024; Dorbala, Mullen, and Manocha 2024). **Second**, we introduce the Parameter-Efficient Intent Aligner (PEIA), a lightweight semantic reasoning module trained via lightweight prompt tuning. By describing the current view and inferring the likelihood of unseen objects, PEIA enables strong generalization using minimal semantic priors. **Third**, CoS continuously optimizes exploration by balancing action cost and gain, mimicking human-like decision-making. A Reflection-Guided Destination Verifier (RDV) further confirms goal completion through object detection, segmentation, and visual question answering (VQA), enabling robust destination verification. Together, these components allow CoS to perform efficient, interpretable, and generalizable navigation without reliance on large-scale LLMs or labeled training data. Our key contributions are summarized as follows:

- We propose a zero-shot object navigation framework called CoS. This approach enables human-like, scalable, and efficient navigation without relying on large models or complex prompt engineering, thereby improving adaptability and reducing inference overhead.
- We design an optimal-benefit multi-map construction that leverages a global-local direction scoring mechanism based on a multi-map balancing semantic-aware benefit and cost during the exploration for the optimal frontier.
- We reformulate lightweight semantic mapping as a visual reasoning task and introduce the PEIA module, which infers target directions via parameter-efficient prompts. This avoids inaccuracies caused by multi-stage linguistic descriptions and enhances semantic-aware search efficiency.
- We develop a Reflection-Guided Destination Verifier (RDV) to mitigate semantic mismatches by iteratively confirming and correcting navigation decisions, avoiding reliance on globally accurate semantic maps.

Experiments demonstrate that CoS exhibits semantic-aware search capabilities, effectively inferring target directions beyond the robot's current view through visual reasoning. Compared to LLM-based methods, including VLFM (Yokoyama et al. 2024) (CLRA 24 Best Paper), our approach achieves the highest success rates on zero-shot object navigation, e.g., +2.8% on HM3D (Ramakrishnan et al. 2021), +1.2% on MP3D (Chang et al. 2017). Notably, our PEIA module, trained on a small dataset such as MS COCO (Lin et al. 2014), efficiently extracts semantic priors.

Related Works

Object Navigation. Object Navigation is a core task in Vision-Language Navigation. Early methods relied on imitation learning (Karnan et al. 2022) or reinforcement learning (RL) (Wijmans et al. 2020; Zhu et al. 2017; Min et al. 2023; Yadav et al. 2023) to guide agents via visual cues. Subsequent work improved generalization through data augmentation (Maksymets et al. 2021), object-level semantic mapping (Chaplot et al. 2020a; Zhou et al. 2023), and topological reasoning (Chaplot et al. 2020b; Wu et al. 2024). Top-down map-based approaches (Ramakrishnan et al. 2022; Yu, Kasaei, and Cao 2023; Yu et al. 2024; Shah et al. 2023) and hybrid methods combining RL with classical techniques (Ramakrishnan et al. 2022) have shown strong performance. However, RL-based methods remain data-intensive, and many approaches (Ramakrishnan et al. 2022; Zhu, Zhao, and Kong 2022) struggle to generalize to unseen objects, limiting their real-world scalability. To address this, we propose CoS for generalization to novel environments and efficiently locates open-vocabulary objects.

Zero-Shot Object Navigation. Zero-shot object navigation aims to locate unseen objects in unfamiliar environments without scene-specific training. Existing frontier-based methods (Yamauchi 1997; Gadre et al. 2023) guide agents toward the nearest frontier, while map-based approaches construct top-down semantic maps to identify promising unexplored regions (Ramakrishnan et al. 2022; Long et al. 2024; Yu, Kasaei, and Cao 2023; Chaplot et al. 2020a). These can be categorized into implicit (Yokoyama et al. 2024; Long et al. 2024) and explicit (Chaplot et al. 2020a; Zhou et al. 2023) semantic mapping strategies. Recent advances leverage language models for semantic reasoning; for instance, VLFM (Yokoyama et al. 2024) selects frontiers using BLIP2-generated cues, and Instruct-Nav (Long et al. 2024) focuses on trajectory alignment. However, these methods lack effective estimation of the object's potential direction. To address this, we introduce a multi-map strategy for optimal frontier prediction by balancing expected benefit and exploration cost, and further design a reflection-based verifier to confirm goal achievement.

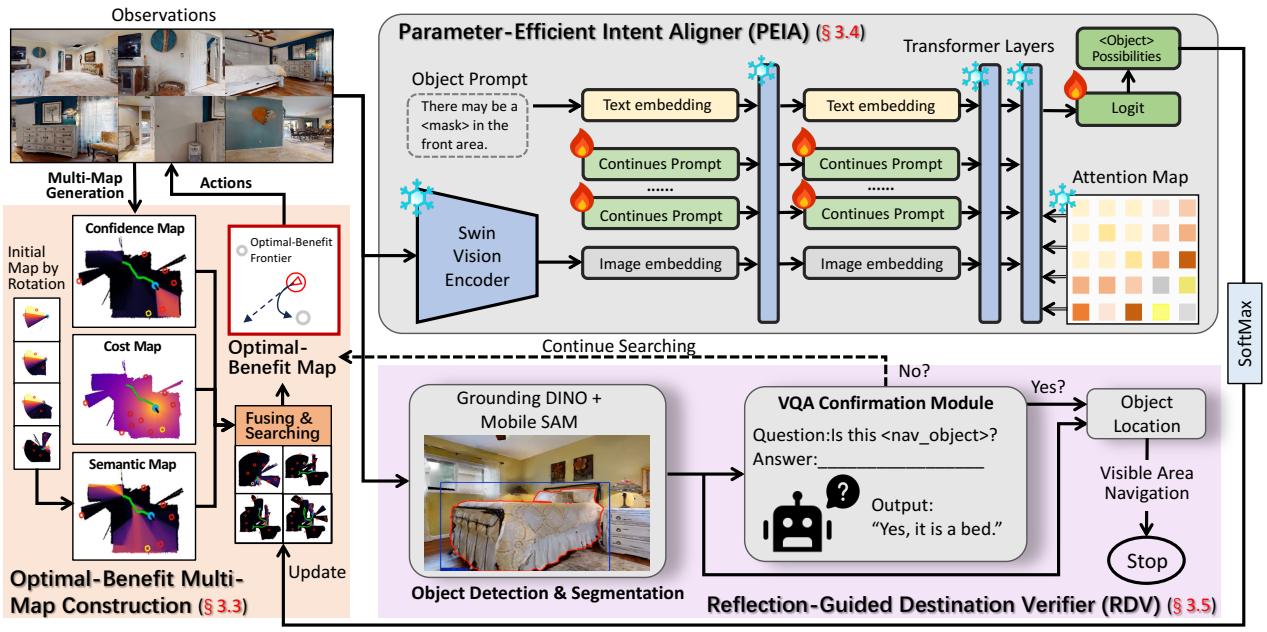

Figure 2: System framework of our CoS. It integrates Optimal-Benefit Multi-Map Construction, Parameter-Efficient Intent Aligner (PEIA), and Reflection-Guided Destination Verifier (RDV). It holistically considers both the *benefit* and *cost* of exploration, enabling scalable navigation in unknown environments for unseen objects without retraining or reliance on LLMs.

Navigation with Large Language Models. Recently, various methods for planning frontiers in exploration maps in the zero-shot task using intuition from the LLMs or scoring and ranking have been proposed (Long et al. 2024; Yu, Kasaei, and Cao 2023; Zhou et al. 2023; Shah et al. 2023; Zhou, Hong, and Wu 2024). However, important parts of these methods: path planning (Majumdar et al. 2022; Cai et al. 2024; Zhou et al. 2024), frontier selection (Zhou et al. 2023; Dorbala, Mullen, and Manocha 2024; Shah et al. 2023), chain-of-though (Long et al. 2024; Shah et al. 2023), are all limited by the capabilities of LLMs. The employment of more sophisticated LLMs has been demonstrated to engender enhanced navigation success rates. However, general-purpose LLMs are not optimized for visual-language navigation; their large scale is primarily designed for general reasoning. Without task adaptation, model weights are underutilized, causing inefficient decision-making. Thus, we introduce a lightweight vision-language reasoning model instead of LLMs.

Chain-of-Search (CoS)

Problem Formulation

We address zero-shot object navigation, where the robot is deployed in an environment $\mathcal{E}$ with unknown layout and unseen target object locations. The target object is initially out of view. At each time step t, the robot captures observation $I_t = (I_t^{rgb}, I_t^{depth})$ using an RGB-D camera, where $I_t^{rgb} \in \mathbb{R}^{H\times W\times 3}$ is the RGB image and $I_t^{depth} \in \mathbb{R}^{H\times W}$ is the depth map. The task is deemed successful if the agent navigates to within 1 meter of any instance of the target object category g and issues a '`STOP`' action within 500 steps.

Overview

We treat zero-shot object navigation as an exploration task, as illustrated in Fig. 2. Inspired by how humans explore unfamiliar environments, CoS begins by rotating in place to build an initial map during the Optimal-Benefit Multi-Map Construction stage. The proposed Parameter-Efficient Intent Aligner (PEIA) then updates the optimal-benefit map throughout the search by fusing the Semantic Map, Confidence Map, and Cost Map to predict the most promising frontier for exploration. Finally, when a potential target is observed, CoS mimics human behavior by approaching while verifying the object using a Reflection-Guided Destination Verifier (RDV), which leverages object detection, segmentation, and VQA for goal confirmation.

Optimal-Benefit Multi-Map Construction

Initial Map and Navigable Frontiers. We construct a top-down 2D initial map using odometry and depth observations by projecting uniformly sampled 3D point clouds onto a global coordinate system. This map is dynamically updated by recording the robot's trajectory and observed data over time. Points not meeting obstacle criteria are filtered out, and the geometric structure of the remaining obstacles provides spatial awareness for downstream path planning and target localization. Navigable frontiers are defined as the midpoints of boundary edges in the explored map, and the most beneficial frontier is selected based on this map.

Semantic Map. Humans infer the likelihood of an object appearing in a given direction based on the visual scene ahead. Inspired by this, we formulate the robot's map exploration as a perception task modeled via image-caption

prediction. Specifically, the model predicts $\mathrm{Logit}(T)$ from the in-view RGB image I_t^{rgb} and the task prompt T. As illustrated in Figure 2, the proposed PEIA module processes I_t^{rgb} at time t using a Swin Transformer encoder to extract visual features. These features are jointly encoded with the prompt T, which includes a <mask> token, and passed into a multimodal Transformer decoder for mixed-attention computation. The resulting probability distribution determines the semantic value $s_{i,j}$ for each pixel (i, j) in the robot's current field of view:

$$s_{i,j}^t = \sum_{k=1}^{m} \left(\frac{e^{l_k(w)}}{\sum_{z \in Z} e^{l_k(z)}} \right) \tag{1}$$

where m represents the length of masked positions, $l_k(\cdot)$ denotes the probability for a specific word at the kth masked position in the Logit and Z represents the entire vocabulary of the model. We rely on the Confidence Map to process overlapping pixels (p, q) at time T. The semantic value $s_{p,q}$ is updated to favor the higher weighted semantic value:

$$s_{p,q} = \frac{s_{p,q}^T \times c_{p,q}^T + s_{p,q} \times c_{p,q}}{c_{p,q}^T + c_{p,q}} \tag{2}$$

Confidence Map. Humans are generally more confident in observations made directly ahead than in peripheral views. To mimic this perceptual bias and address overlapping conical fields of view in the robot's perception, we construct a Confidence Map that models view-dependent observation reliability. Let a pixel (i, j) be observed at time t; then the observed confidence value $c_{i,j}^t$ is defined to decrease as the angle θ_t from the current heading increases:

$$c_{i,j}^t = \cos^2 \left(\frac{\pi}{2} \times \theta_t \times \left(\frac{\theta_0}{2} \right)^{-1} \right) \tag{3}$$

where θ_0 denotes the robot's horizontal Field-of-View (FOV). At time T, suppose the robot's current field of view overlaps with a previously explored region. For each pixel (p, q) in the overlapping area, its confidence value $c_{p,q}$ is updated by retaining the higher of the two confidence scores (after semantic value update):

$$c_{p,q} = \frac{c_{p,q}^T \times c_{p,q}^T + c_{p,q} \times c_{p,q}}{c_{p,q}^T + c_{p,q}} \tag{4}$$

Cost Map. People's choice of exploration direction is often influenced by distance. When the differences among various directions are subtle, most individuals tend to opt for exploring the nearer fork. Therefore, we propose the Cost Map to optimize the robot's path planning and frontier selection. For a pixel (i, j) at time t, the cost value $cs_{i,j}^t$ in the Cost Map is defined as:

$$cs_{i,j}^t = \frac{1}{dist(t, i, j)^k + C} \tag{5}$$

where $dist(t, i, j)$ is defined as the Euclidean Distance between the pixel (i, j) and the robot at time t in the 2D projection, and C is a small constant (e.g., 10^{-5}). Essentially, this function encourages the robot to explore regions further along the current direction, while mitigating the inefficiencies caused by frequently traversing already explored areas. The hyperparameter k in this heuristic function will be discussed in the experimental section.

Map Fusion and Optimal-Benefit Frontier. An Optimal-Benefit Map is generated through multi-map fusion. Let A_n^t denote the set of frontiers in the explored area at time t, and let S_i^t represent the fusion score of frontier i. We define: $S_i^t = s_i^t \times sc_i^t$, where s_i^t and sc_i^t are the average semantic and confidence scores, respectively, computed over all pixels within a radius r meters around frontier i. The most likely frontier toward the navigation target at time t considering both the benefit and cost, referred to as the Optimal-Benefit Frontier, is given by:

$$\arg\max(S_i^t) \quad \forall i \in A_n^t \tag{6}$$

Parameter-Efficient Intent Aligner (PEIA)

In this section, we trained a vision-language reasoning and aligning model for object navigation, Parameter-Efficient Intent Aligner (PEIA), which is trained by both of discrete prompt and continues prompt to update the Semantic Map.

Module Structure. As shown in Figure 2, PEIA adopts a frozen encoder-decoder architecture. The image I^{rgb} is encoded by Swin Transformer layers, and a Texture Token from a single embedding layer is concatenated with Image Tokens. These tokens are processed via Transformer layers with full self-attention to align visual and textual modalities. Components marked with a snowflake icon, including encoder/decoder layers and cross-attention matrices, are trained during pre-training and then frozen. Flame-marked continuous prompts are trained during fine-tuning, while classification head logits are trained in both stages.

Pre-Training with Discrete Prompt. The model acquires salient capabilities by learning high-entropy knowledge words from images (Yan et al. 2023, 2022a). PEIA is pre-trained to extract key information from images using object prompts. These discrete prompts are generated by automatically tagging parts of speech (e.g., Flair (Akbik et al. 2019)), where verbs and nouns are replaced with <mask>. The model's objective is to recover the vocabulary at the <mask> positions based on the image and the remaining caption. The loss function uses the cross-entropy function:

$$\arg\min \left(\sum_{y \in Y} p\left(y\right) \log p\left(\hat{y}|X\right) \right) \tag{7}$$

where y and $\hat{y}$ denote the predicted words and the ground-truth respectively, Y denotes the set of words at the $< mask >$ positions and X represents the image and masked textual context prompt.

Fine-Tuning with Continuous Prompt. The model is fine-tuned to learn the complete content of sentences for a holistic understanding of the image. Following VPT (Jia et al. 2022), continuous prompts are injected into each layer of the Transformer decoder via fully connected layers. For the i-th Transformer layer $\mathcal{T}_i$, the set of N learnable d-dimensional prompt embeddings is represented as:

$$\mathbb{T}^{(i)} = \left\{ t_k \in \mathbb{R}^d \mid T_j = (t_0, t_1, \ldots, t_k, \ldots, t_d) \right\}_{j=1}^{N} \tag{8}$$

The deep prompt self-attention in the decoder is implemented with n multimodal Transformer layers. During this

phase, trainable components include the prompt projection function $f(\cdot)$, the prompt embeddings $\mathbb{T}^{(i)}$, and the final prediction head Logit; all other modules $\mathcal{T}_i$ are kept frozen. The function $f(\cdot)$ projects continuous prompts into the self-attention layer, while Logit maps the output embeddings to log-likelihood scores Z for word prediction:

$$\left[\mathbb{E}^{(i)}_{\text{Text}}, \mathbb{E}^{(i)}_{\text{Prompt}}, \mathbb{E}^{(i)}_{\text{Image}}\right] = \mathcal{T}_i\left(\left[\mathbb{E}^{(i-1)}_{\text{Text}}, f(\mathbb{T}^{(i-1)}), \mathbb{E}^{(i-1)}_{\text{Image}}\right]\right)$$
$$Z = \text{Logit}\left(\mathbb{E}^{(n)}_{\text{Text}}\right) \tag{9}$$

To adapt the model for the navigation intent alignment tasks while preserving pretrained knowledge, random tokens in the input text are replaced with <mask>, and the objective is to reconstruct the original sentence. Let $\mathcal{Z}$ be the vocabulary set, with z as the predicted token and $\hat{z}$ as the ground truth. The fine-tuning loss is defined as:

$$\arg\min\left(\sum_{z\in\mathcal{Z}} p(z)\log p(\hat{z})\right) \tag{10}$$

Reflection-Guided Destination Verifier (RDV)

Object Detection and Segmentation. For object categories in the COCO dataset (Lin et al. 2014), we use YOLOv7 (Wang, Bochkovskiy, and Liao 2023) for detection. To generalize beyond COCO, we incorporate Grounding-DINO (Liu et al. 2024), an open-vocabulary detector capable of identifying arbitrary targets, enabling CoS to navigate toward any specified object. Once detected, a pre-trained Mobile-SAM (Zhang et al. 2023) segmenter generates a mask for object verification and localization.

Object Confirmation and Location. Then, the object segmentation module pairs each bounding box with the prompt: ("Question: Is This <Nav_Object>, yes or no? Answer:"). This is passed to the VQA Confirmation module, which continuously performs visual question answering during target approach. A negative response triggers a return to the Frontier Exploring phase. Here, <Nav_Object> denotes the object specified by the current navigation goal. As shown in Fig. 2, reasoning-based semantic priors help rectify errors from the Detection Module. After VQA confirmation, we fuse the segmented RGB image I_t^{rgb} and depth image I_t^{depth} to construct a point cloud from the robot's perspective. The centroid of the nearest 25% of points is selected as the Object Point Navigation target. The robot navigates to this point via a local verifier and issues the 'STOP' command upon entering a predefined success radius.

Visible Area Navigation. Inspired by human strategies for simplifying complex tasks, we unify all local navigation behaviors, toward frontiers or object goals in visible area, under a point-goal navigation framework. This visible area navigation policy is trained with VER (Wijmans, Essa, and Batra 2022) using the same settings as (Yokoyama et al. 2024). It leverages only depth inputs and the relative position between the robot and target, enabling efficient and robust local planning without semantic supervision. This design circumvents challenges in map-based methods and improves reliability under occlusions.

Experiment

Experience Setup

Datasets. We evaluated on the validation splits of two realistic 3D scan datasets, HM3D (Ramakrishnan et al. 2021) and MP3D (Chang et al. 2017), using the Habitat simulator (Savva et al. 2019). HM3D's validation includes 6 object categories with 2000 episodes over 20 scenes, while MP3D's validation consists of 2195 episodes across 21 object categories and 11 scenes.

Metrics. For all state-of-the-art methods, the Success Rate (SR↑) and Success Weighted by Inverse Path Length (SPL↑) (Anderson et al. 2018) is reported for the purpose of comparison. SR↑ is defined as $\frac{1}{N}\sum_{i=1}^{N} S_i$ and SPL↑ is defined as $\frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(l_i,p_i)}$, where S_i is assigned a value of 1 if the episode is successful and 0 otherwise. N is the number of episodes, l_i is the shortest trajectory length between the start position and one of the successful positions, and p_i is the trajectory length of the current episode i. In order to facilitate a more profound comprehension of the modules within the aforementioned model, Distance to Goal (DTG↓) and Navigation Error (NE↓) are utilised for both the ablation and comparison experiments. The DTG↓ is defined as the distance between the agent and the target at the conclusion of the given episode, and the NE↓ is defined as the rate of failure determination.

Baselines. We compare our method with state-of-the-art zero-shot object navigation approaches: ZSON (Majumdar et al. 2022), CLIP on Wheels (CoW) (Gadre et al. 2023), ESC (Zhou et al. 2023), L3MVN (Yu, Kasaei, and Cao 2023), VLFM (Yokoyama et al. 2024), PixNav (Cai et al. 2024), VoroNav (Wu et al. 2024), ImagineNav (Zhao et al. 2024), GAMap (Yuan et al. 2024), and UniGoal (Yin et al. 2025). Among zero-shot methods utilizing LLMs/VLMs, PixNav prompts LLMs with raw pixels for navigation reasoning. VoroNav uses Voronoi decomposition for exploration. ImagineNav models navigation as a visual-language sequence, optimizing viewpoints via point-goal policies. GAMap scores frontiers by capturing multi-scale geometric and affordance object features. We also compare with non-zero-shot baselines: PONI (Ramakrishnan et al. 2022), SemExp (Chaplot et al. 2020a), and L3MVN (Feed-forward) (Yu, Kasaei, and Cao 2023). PONI combines latent frontier selection with reinforcement learning and mapping. SemExp applies semantic expansion for exploration. L3MVN (Feed-forward) integrates a trained probe head and LLM for object localization reasoning.

Main Results

Zero-Shot Performance Superiority. We collected data from published articles. As shown in Table 1, CoS outperforms all zero-shot methods in terms of SR↑ on both datasets, with an increase of +2.9% on HM3D compared to GAMap and +1.2% on MP3D compared to VLFM. Regarding SPL↑, our method falls short of VLFM by -1.3% SPL↑ on HM3D, but achieves +0.1% SPL↑ on MP3D. Thanks

Method	Conference Published	Reasoning LLM/VLM Needed	Zero-Shot	Semantic Nav Training Free	HM3D		MP3D		Params
					SPL↑	SR↑	SPL↑	SR↑	
Semexp	NeurIPS 20	None	×	×	18.8	37.9	-	-	75M
PONI	CVPR 22	None	×	×	-	-	12.1	31.8	45M
V3MVN (Feed-forward)	IROS 23	None	×	×	25.4	54.2	-	-	190M
PixNav	ICRA 24	GPT-4	✓	✓	20.5	37.9	-	-	1.8T
ESC	ICML 23	InstructGPT	✓	✓	22.3	39.2	14.2	28.7	4.9G
VoroNav	ICML 24	GPT-3.5	✓	✓	26.0	42.0	-	-	76.3G
ImagineNav	ICLR 25	GPT-4o-mini	✓	✓	23.8	53.0	-	-	30.5G
GAMap	NeurIPS 24	GPT-4V	✓	✓	26.0	53.1	-	-	1.37T
UniGoal	CVPR 25	LLaMA-2-7B	✓	✓	25.1	54.5	16.4	41.0	26.7G
ZSON	NeurIPS 22	None	✓	×	12.6	25.5	4.8	15.3	278M
CoW	CVPR 23	None	✓	✓	18.1	32.0	3.7	7.4	175M
V3MVN (Zero-Shot)	IROS 23	None	✓	✓	23.1	50.4	-	-	190M
VLFM	ICRA 24	None	✓	✓	30.4	52.5	17.5	36.4	531M
CoS (Ours)	-	**None**	✓	✓	**29.1**	**55.9**	**17.6**	**37.6**	469M

Table 1: Performance comparison of our method with state-of-the-art (SOTA) methods evaluated on the HM3D and MP3D benchmarks. It is organized into three sections: the upper section encompasses non-zero-shot methods; the middle section includes zero-shot methods that utilize LLMs or VLMs to reasoning why and where to navigation; and the lower section features zero-shot methods that do not employ LLMs or VLMs. Our method demonstrates superior performance over previous zero-shot approaches in terms of SR, even those that use LLMs or VLMs.

Method	HM3D			
	SR↑	SPL↑	DTG↓	Travel Stairs
CoS (Ours)	**55.90**	29.06	**4.09**	**6.60**
VLFM	52.50	**30.40**	4.20	4.60

Table 2: Explaination of the slightly lower SPL compared to VLFM. While outperforming VLFM in other metrics, our method explores more floors and increases the percentage of Travel Stairs, thereby affecting the SPL on HM3D dataset.

to its chain-style reasoning ability, CoS achieves significantly higher SR↑ and SPL↑ on the HM3D dataset compared to MP3D, addressing the semantic loss challenges in zero-shot object navigation tasks without relying on heavyweight LLMs or VLMs. A significant portion of the failures on MP3D is attributed to parts of the depth near the boundaries being set to infinity, which hinders the robot's ability to effectively handle boundary and obstacle issues.

Efficiency Without Heavyweight Reasoning. Notably, CoS's SPL↑ and SR↑ on the HM3D dataset surpass those of SOTA methods that require heavyweight invocation of LLM/VLM to reasoning, achieving +3.1% SPL↑ and +2.8% SR↑ on HM3D compared to GAMap (which relies on GPT-4V), all while operating without semantic navigation training and in a zero-shot manner. This represents lower computational overhead and improved operational efficiency.

Comparison with Non-Zero-Shot Methods. For non-zero-shot methods, CoS also achieves +3.7% SPL↑ and +1.7% SR↑ compared to L3MVN (feed-forward) on HM3D, and +3.2% SPL↑ and +1.6% SR↑ compared to SemExp on MP3D, demonstrating the robust performance of CoS.

Detailed Comparison with VLFM. To further examine performance differences, we reproduced VLFM on the HM3D dataset and compared it with CoS (Table 2), with Travel Stairs indicating the percentage of multi-floor exploration. As shown in Fig. 3, CoS achieves more effective path planning, with higher SR↑ (+3.65%) and lower DTG↓ (-0.107), demonstrating better task completion and object proximity. The slight SPL↑ drop (-1.1%) suggests longer paths on average, potentially due to CoS's greater tendency to explore other floors (+2.0% Travel Stairs). While this may lead to missing nearer targets, it also reflects a more human-like exploratory strategy in unfamiliar environments.

Ablation Study

Impact of Multi-Map Construction and RDV Module. We perform an ablation study on the full HM3D episodes to evaluate the contribution of key components, as summarized in Table 3. These include the Confidence Map, Semantic Map, Cost Map in Multi-Map Construction, and the VQA Confirmation Module. Removing any component leads to decreased SR↑ and increased DTG↓, with the Semantic Map having the greatest impact on DTG↓ due to loss of semantic guidance. Despite this, the Cost Map and other modules help mitigate performance loss. Without semantic cues, fewer objects are detected, reducing NE↓. Replacing the Confidence Map with averaged values or removing the Cost Map degrades all metrics (SR↑, SPL↑, DTG↓, NE↓). Removing the Confirmation Module significantly increases NE↓ (+8.60%), highlighting its role in correcting detection errors. Although SPL↑ improves due to fewer detours, the drop in SR↑ indicates that leveraging VQA confirmation enhances successful target localization, addressing the semantic loss challengs.

Impact of PEIA Module. As shown in Table 4, our PEIA module-trained only on the MS COCO (Lin et al. 2014) dataset-achieves better SR↑ (+0.65%) and DTG↓ (-0.08) in the CoS framework compared to the BLIP2 model (OPT), while having only a fraction of the parameters and train-

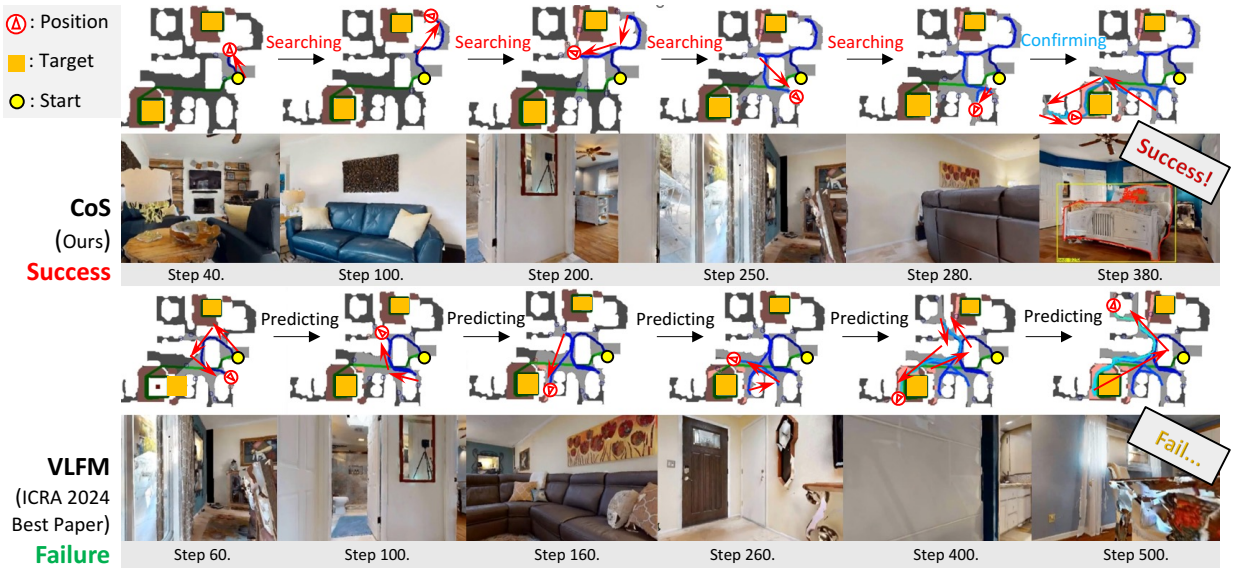

Figure 3: Qualitative comparison of our proposed CoS and previous SOTA. CoS employs a more human-like strategy by performing semantic-aware frontier search that balances exploration benefits and costs, enabling superior path planning compared to VLFM (Yokoyama et al. 2024), which selects only the top-scoring prediction at each step.

Method	HM3D			
	SR↑	SPL↑	DTG↓	NE↓
CoS (Ours)	**55.90**	29.06	**4.09**	10.10
w/o Confidence Map (Avg.)	54.85	28.49	4.16	10.30
w/o Cost Map	54.50	27.78	4.13	10.90
w/o Semantic Map	53.15	27.77	4.38	**9.85**
w/o RDV	52.35	**29.75**	4.22	18.70

Table 3: Effect of Multi-Map Construction and RDV Module. The results are evaluated on the HM3D dataset.

Method	Total Params	Training Datasets	HM3D		
			SR↑	SPL↑	DTG↓
PEIA (Ours)	0.122B	0.33M	**55.90**	29.06	**4.09**
BLIP2_OPT	7.8B	129M	55.25	**30.31**	4.17

Table 4: Effect of PEIA module. The results are evaluated on the HM3D dataset versus the BLIP2 model.

ing data. We attribute the decline in SPL↑ (-1.24%) of PEIA to its relatively weaker distinction between similar frontier observations. In contrast, BLIP2, being based on a LLM, benefits from more extensive knowledge and commonsense reasoning, leading to more deterministic path planning and better SPL↑. However, this same reliance on potentially biased or noisy knowledge could also account for its worse performance in SR↑ and DTG↓.

Impact of Prompt Tuning. To evaluate the effectiveness of two-phase prompt-tuning for PEIA, we conducted ablations by removing either the pre-training or fine-tuning stage. As shown in Table 5, eliminating pre-training significantly reduces Success Rate (-3.55% SR↑ on HM3D), indicating that continuous prompts alone cannot capture image-grounded knowledge. In contrast, omitting fine-tuning leads to only minor drops in SR↑ and DTG↓, suggesting that while the pre-trained model captures semantic cues to some extent, the full PEIA module after fine-tuning more effectively models sentence semantics and image-text alignment, thereby addressing the tuning efficiency challenges.

Method	HM3D		
	SR↑	SPL↑	DTG↓
PEIA (Pre-Training & Fine-Tuning)	**55.90**	29.06	**4.09**
w/o Pre-Training	53.35	29.08	4.16
w/o Fine-Tuning	55.60	**29.38**	4.14

Table 5: Effect of prompt tuning in PEIA, evaluated on the HM3D Dataset. All other components are used.

Conclusion

We present CoS, a novel object searching framework for zero-shot navigation that replaces heavyweight LLMs with a parameter-efficient reasoning framework, built from a prompt tuning methodology for semantic image-captioning and question-answering tasks. By incorporating cost and benefit considerations through Multi-Map Construction, we derive the Optimal-Benefit Frontier as an exploration target to guide the chain of object perception, confirmation, reasoning, localization, and navigation, all without requiring dedicated semantic navigation training. Our approach outperforms previous SOTA methods on both the HM3D and MP3D datasets. These promising results signal a paradigm shift and pave the way for future work to enhance prompt-map alignment strategies, which is crucial for long-term vision-language navigation.

Acknowledgments

This project is supported by the Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020019 and No. LQN25D010005), the Fundamental Research Funds for Provincial Universities of Zhejiang (No. GK249909299001-033 and No. GK249909299001-023), the Key Laboratory of Data Science and Intelligence Education (Hainan Normal University), Ministry of Education (No. DSIE202403), and the National Natural Science Foundation of China (No. 42401517).

References

Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 54–59.

Anderson, P.; Chang, A. X.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; and Zamir, A. R. 2018. On Evaluation of Embodied Navigation Agents. *CoRR*, abs/1807.06757.

Cai, W.; Huang, S.; Cheng, G.; Long, Y.; Gao, P.; Sun, C.; and Dong, H. 2024. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5228–5234. IEEE.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*.

Chaplot, D. S.; Gandhi, D.; Gupta, A.; and Salakhutdinov, R. 2020a. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *In Advances of Neural Information Processing Systems (NeurIPS)*.

Chaplot, D. S.; Salakhutdinov, R.; Gupta, A.; and Gupta, S. 2020b. Neural Topological SLAM for Visual Navigation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cui, Y.; Yan, L.; Cao, Z.; and Liu, D. 2021. Tf-Blender: Temporal Feature Blender for Video Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8138–8147.

Dorbala, V. S.; Mullen, J. F.; and Manocha, D. 2024. Can an Embodied Agent Find Your "Cat-shaped Mug"? LLM-Based Zero-Shot Object Navigation. *IEEE Robotics and Automation Letters*, 9(5): 4083–4090.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2023. CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23171–23181. IEEE.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. arXiv:2203.12119.

Karnan, H.; Warnell, G.; Xiao, X.; and Stone, P. 2022. VOILA: Visual-Observation-Only Imitation Learning for Autonomous Navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, 2497–2503. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *European Conference on Computer Vision (ECCV)*, volume 15105 of *Lecture Notes in Computer Science*, 38–55. Springer.

Long, Y.; Cai, W.; Wang, H.; Zhan, G.; and Dong, H. 2024. InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems (NeurIPS)*.

Maksymets, O.; Cartillier, V.; Gokaslan, A.; Wijmans, E.; Galuba, W.; Lee, S.; and Batra, D. 2021. THDA: Treasure Hunt Data Augmentation for Semantic Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15374–15383.

Min, S. Y.; Tsai, Y. H.; Ding, W.; Farhadi, A.; Salakhutdinov, R.; Bisk, Y.; and Zhang, J. 2023. Self-Supervised Object Goal Navigation with In-Situ Finetuning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7119–7126.

Nie, D.; Guo, X.; Duan, Y.; Zhang, R.; and Chen, L. 2025. WMNav: Integrating Vision-Language Models into World Models for Object Goal Navigation. *CoRR*, abs/2503.02247.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning. In *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*. IEEE.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J. M.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; Savva, M.; Zhao, Y.; and Batra, D. 2021. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; Parikh,

D.; and Batra, D. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Shah, D.; Equi, M. R.; Osinski, B.; Xia, F.; Ichter, B.; and Levine, S. 2023. Navigation with Large Language Models: Semantic Guesswork as a Heuristic for Planning. In Tan, J.; Toussaint, M.; and Darvish, K., eds., *Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, 2683–2699. PMLR.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wijmans, E.; Essa, I.; and Batra, D. 2022. VER: Scaling On-Policy RL Leads to the Emergence of Navigation in Embodied Rearrangement. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems (NeurIPS)*.

Wijmans, E.; Kadian, A.; Morcos, A.; Lee, S.; Essa, I.; Parikh, D.; Savva, M.; and Batra, D. 2020. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. In *8th International Conference on Learning Representations (ICLR)*. OpenReview.net.

Wu, P.; Mu, Y.; Wu, B.; Hou, Y.; Ma, J.; Zhang, S.; and Liu, C. 2024. VoroNav: Voronoi-based Zero-shot Object Navigation with Large Language Model. In *Forty-first International Conference on Machine Learning (ICML)*. OpenReview.net.

Yadav, K.; Majumdar, A.; Ramrakhya, R.; Yokoyama, N.; Baevski, A.; Kira, Z.; Maksymets, O.; and Batra, D. 2023. OVRL-V2: A simple state-of-art baseline for ImageNav and ObjectNav. *CoRR*, abs/2303.07798.

Yamauchi, B. 1997. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)*, 146–151.

Yan, L.; Han, C.; Xu, Z.; Liu, D.; and Wang, Q. 2023. Prompt Learns Prompt: Exploring Knowledge-Aware Generative Prompt Collaboration For Video Captioning. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 1622–1630. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yan, L.; Ma, S.; Wang, Q.; Chen, Y.; Zhang, X.; Savakis, A.; and Liu, D. 2022a. Video Captioning Using Global-Local Representation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(10): 6642–6656.

Yan, L.; Wang, Q.; Ma, S.; Wang, J.; and Yu, C. 2022b. Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 33(1): 393–406.

Yan, L.; Wang, Q.; Zhao, J.; Guan, Q.; Tang, Z.; Zhang, J.; and Liu, D. 2024. Radiance field learners as uav first-person viewers. In *European Conference on Computer Vision (ECCV)*, 88–107. Springer.

Yin, H.; Xu, X.; Zhao, L.; Wang, Z.; Zhou, J.; and Lu, J. 2025. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 19057–19066.

Yokoyama, N.; Ha, S.; Batra, D.; Wang, J.; and Bucher, B. 2024. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 42–48. IEEE.

Yu, B.; Kasaei, H.; and Cao, M. 2023. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3554–3560. IEEE.

Yu, B.; Liu, Y.; Han, L.; Kasaei, H.; Li, T.; and Cao, M. 2024. VLN-Game: Vision-Language Equilibrium Search for Zero-Shot Semantic Navigation. arXiv:2411.11609.

Yuan, S.; Huang, H.; Hao, Y.; Wen, C.; Tzes, A.; and Fang, Y. 2024. GAMap: Zero-Shot Object Goal Navigation with Multi-Scale Geometric-Affordance Guidance. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.; Lee, S.; and Hong, C. S. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. *CoRR*, abs/2306.14289.

Zhao, X.; Cai, W.; Tang, L.; and Wang, T. 2024. ImagineNav: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination. *CoRR*, abs/2410.09874.

Zhou, G.; Hong, Y.; Wang, Z.; Wang, X. E.; and Wu, Q. 2024. NavGPT-2: Unleashing Navigational Reasoning Capability for Large Vision-Language Models. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *European Conference on Computer Vision (ECCV)*, volume 15065 of *Lecture Notes in Computer Science*, 260–278. Springer.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 7641–7649. AAAI Press.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, 42829–42842. PMLR.

Zhu, M.; Zhao, B.; and Kong, T. 2022. Navigating to Objects in Unseen Environments by Distance Prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10571–10578. IEEE.

Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J. J.; Gupta, A.; Fei-Fei, L.; and Farhadi, A. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 3357–3364. IEEE.