IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 June 2026, accepted 8 July 2026, date of publication 13 July 2026, date of current version 17 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3712357

RESEARCH ARTICLE

ForesightNav: Adaptive Semantic-Geometric Planning for Universal Zero-Shot Goal-Oriented Navigation

YU LIU[1], CHENGJIE FANG[2], AND LIKUN LU[1]
[1]School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

Corresponding author: Likun Lu (lklu@bigc.edu.cn)

This work was supported by the Disciplinary Construction and Postgraduate Education Project of Beijing Institute of Graphic Communication (Grant Nos. 21090126017, 21090126005, 21090325003), the Key Laboratory Project of Signal and Information Processing for High-End Printing Equipment of Beijing (Grant No. 20190226003), and the Fundamental Research Funds for the Municipal Universities of Beijing (Grant No. Ee202535).

ABSTRACT Universal zero-shot goal-oriented navigation requires an agent to locate object categories, target instances, or text-described goals in unseen environments without task-specific policy training. Existing methods usually rely on either broad semantic exploration or fine-grained target verification, but they can become inefficient when semantic cues are sparse, ambiguous, or insufficient for reliable graph matching. This paper presents ForesightNav, a training-free universal zero-shot navigation framework that adaptively combines context-aware semantic value mapping, graph-based target verification, and TSP-based frontier ordering. Guided by semantic-score dispersion and graph-matching confidence, the planner switches among geometric exploration, semantic exploration, partial-match exploration, and target verification, balancing broad goal-directed search with fine-grained target confirmation. Experiments on MP3D, HM3D, and RoboTHOR across Object Navigation (ObjectNav), Instance Navigation (InstanceNav), and Text Navigation (TextNav) show favorable performance across different goal modalities. Compared with UniGoal, ForesightNav improves ObjectNav performance on three evaluated benchmarks, with the largest gains on RoboTHOR ObjectNav: 18.6 percentage points in Success Rate (SR) and 11.4 in Success weighted by Path Length (SPL). On HM3D, it achieves 65.3% SR on InstanceNav and 29.6% SR on TextNav, suggesting that the same training-free planning process remains applicable to instance- and language-specified goals. Foundation-model analysis and ablation results further support the effectiveness of adaptive switching, context-aware semantic mapping, and TSP-based frontier ordering. Code is available at https://anonymous.4open.science/r/ForesightNav

INDEX TERMS Zero-shot navigation, goal-oriented navigation, semantic mapping, scene graph.

I. INTRODUCTION

Enabling embodied agents to navigate toward user-specified goals in unseen environments remains a fundamental challenge for embodied AI. In goal-oriented navigation, the agent receives only a target specification rather than step-by-step route instructions, and must actively explore the environment to infer the target location. Representative formulations include Object Navigation (ObjectNav) [2], [3], [4], Instance Navigation (InstanceNav) [5], [6], and Text Navigation (TextNav) [8], where the target is specified by an object category, an instance image, or a textual description, respectively. A universal navigation framework should support these goal modalities within a unified decision-making process, without designing separate policies for each task.

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

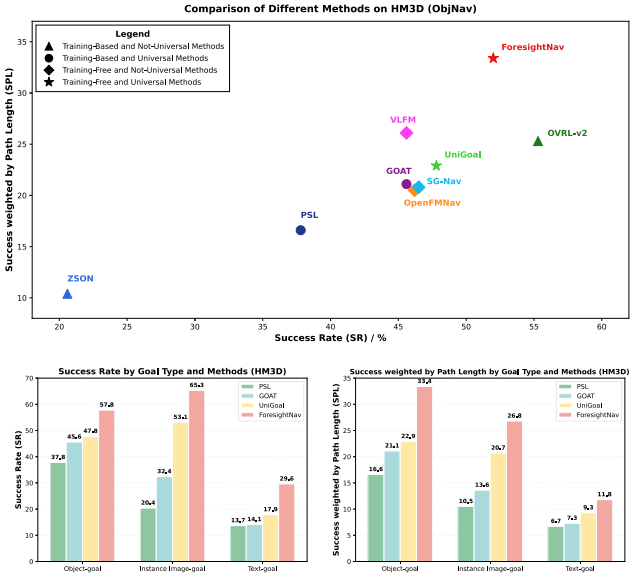


FIGURE 1. Comparison of representative methods on the HM3D benchmark. ForesightNav is evaluated against representative baselines across ObjectNav, InstanceNav, and TextNav. The results indicate improvements over UniGoal, particularly in SPL.

Recent learning-based methods have progressed through training policies or goal representations for multiple goal types. GOAT [7] demonstrated universal navigation across category, image, and language goals, but relied on experience accumulation and was not purely training-free. PSL [8] uses CLIP [9] embeddings to unify different goal representations, but still requires reinforcement learning to build its policy network. Such training-based policies may suffer from limited transferability when deployed in unseen layouts, novel object distributions, or in real-world robotic platforms.

To reduce task-specific training, recent zero-shot navigation methods leverage large language models (LLMs), vision-language models (VLMs), and open-vocabulary perception [10], [11], [12]. The existing approaches often follow two complementary paradigms. Semantic-map-based methods, such as value-map navigation [28], provide broad spatial guidance by estimating the goal relevance over explored regions, but they may fail to distinguish a specific target instance from semantically similar objects. In contrast, graph-matching methods [12] construct structured goal and scene representations for fine-grained target verification, but become unreliable when the observed scene graph is sparse during early exploration. Moreover, both paradigms may rely on greedy frontier selection, which can lead to redundant movements and path oscillations.

These limitations reveal a gap between broad semantic exploration and precise target localization in universal zero-shot navigation. Semantic value maps can provide coarse spatial guidance over explored regions, but they may be ambiguous when the goal refers to a specific instance or a language-described target. Graph matching can support fine-grained verification, but it becomes less reliable when the observed scene graph is sparse during early exploration. Therefore, an effective training-free navigator should not rely on only one of these signals; instead, it should decide when to expand the map, when to follow semantic cues, and when to activate graph-based verification.

To address the challenges in universal zero-shot goal-oriented navigation, where semantic cues may be insufficient for reliable target localization and graph-matching evidence is often incomplete during early exploration, we aim to improve target discovery by enabling an agent to adaptively balance broad exploration and fine-grained target verification. To this end, we propose ForesightNav, a training-free universal zero-shot goal-oriented navigation framework that coordinates three complementary sources of information: geometric coverage, semantic relevance, and graph-matching evidence. Specifically, ForesightNav maintains a context-aware semantic value map for large-scale exploration and an online scene graph for target verification. According to semantic-map dispersion and graph-matching confidence, the planner switches among Geometric Exploration (GE), Semantic Exploration (SE), Partial-Match Exploration (PME), and Target Verification (TV). High-value frontiers are further clustered and globally ordered using TSP-based frontier ordering to reduce redundant paths.

The main contributions of this paper are summarized as follows:

- We formulate target discovery in universal zero-shot navigation as an adaptive coordination problem between semantic exploration and graph-based verification.
- We propose ForesightNav, a training-free navigation framework that combines context-aware semantic value mapping and graph-based target verification.
- We introduce an adaptive semantic-geometric exploration strategy together with TSP-based frontier ordering to reduce redundant exploration.
- We conduct extensive experiments on MP3D, HM3D, and RoboTHOR across ObjectNav, InstanceNav, and TextNav. Compared with UniGoal, ForesightNav improves ObjectNav SR by 10.7, 10.0, and 18.6 percentage points on MP3D, HM3D, and RoboTHOR, respectively, supporting adaptive semantic-geometric coordination as a practical planning principle for universal zero-shot goal-oriented navigation.

The remainder of this paper is organized as follows. Section II reviews related studies on goal-oriented navigation, vision-language navigation, zero-shot multimodal navigation, scene graphs, and semantic planning. Section III formulates the universal zero-shot goal-oriented navigation problem, including multimodal goal definitions, agent observations, map representations, and evaluation objectives. Section IV presents the proposed ForesightNav framework, including the context-aware semantic value map, adaptive semantic-geometric exploration strategy, graph-based target verification, and TSP-based frontier ordering. Section V reports quantitative comparisons, qualitative analysis, and ablation studies on MP3D, HM3D, and RoboTHOR.

II. RELATED WORK

A. GOAL-ORIENTED AND VISION-LANGUAGE NAVIGATION

Goal-oriented navigation requires an embodied agent to reach a target specified by a goal modality (object category, image, or text) rather than step-by-step route instructions. Representative tasks include ObjectNav [2], [3], [4], InstanceNav [5], [6], and TextNav [8]. Instruction-following VLN tasks, such as R2R [1] and REVERIE [16], address a different setting in which agents follow route-like language instructions. These tasks have driven progress through recurrent [17], [18], [19] and Transformer-based architectures [20], [21], [22], [23], but most methods rely on large-scale supervised training and may not generalize to unseen layouts or real-world robotic platforms.

B. ZERO-SHOT MULTIMODAL NAVIGATION

Zero-shot multimodal navigation aims to handle unseen environments and goals without task-specific policy training. Recent methods have leveraged open-vocabulary perception, LLM-based commonsense reasoning, or VLM-based semantic scoring to infer promising exploration regions. For zero-shot ObjectNav, CoW [26] provides a CLIP-based benchmark and baseline, while ESC [4], OpenFMNav [27], and VLFM [28] use commonsense reasoning or vision-language foundation models to guide exploration. In addition, L3MVN [25] leverages large language models to score frontier candidates for visual target navigation, and PixNav [24] bridges zero-shot object navigation with foundation models via pixel-guided navigation. For InstanceNav, Mod-IIN [6] combines frontier-based exploration with keypoint matching for target recognition.

Universal zero-shot navigation methods aim to support multiple goal modalities within a unified framework. UniGoal [12] represents both goals and scenes as graphs and performs graph matching for ObjectNav, InstanceNav, and TextNav. However, existing zero-shot methods still face a trade-off between broad exploration and precise localization: semantic cues support efficient searches but may be ambiguous, whereas graph-based matching enables verification but is unreliable when observations are sparse.

C. SCENE GRAPHS AND SEMANTIC PLANNING

Scene graphs provide structured representations of objects, attributes, and spatial relations, which makes them useful for fine-grained target reasoning. OVSG [29] constructs open-vocabulary scene graphs for entity grounding, while SG-Nav [3] builds an online hierarchical scene graph and uses LLM prompting for zero-shot ObjectNav. UniGoal [12] further formulated multimodal goals as graphs and matched them with an observed scene graph.

Despite their precision, graph-based methods can be inefficient when the observed scene graph contains few informative nodes or relationships. In semantically sparse regions, unreliable partial matches can lead to inefficient exploration or repeated visits. In contrast, semantic value maps provide broader spatial guidance but may lack fine-grained target verification. ForesightNav addresses this gap by combining a context-aware semantic value map, graph-based verification, and adaptive semantic-geometric planning.

III. PROBLEM FORMULATION

We consider universal zero-shot goal-oriented navigation in an unknown indoor environment. At the beginning of an episode, the agent receives a multimodal goal

$$g \in \mathcal{G} = \mathcal{G}_{\text{obj}} \cup \mathcal{G}_{\text{img}} \cup \mathcal{G}_{\text{txt}}, \tag{1}$$

where $\mathcal{G}_{\text{obj}}$, $\mathcal{G}_{\text{img}}$, and $\mathcal{G}_{\text{txt}}$ denote object-category goals, instance-image goals, and text-description goals, respectively.

At each time step t, the agent observes RGB-D and pose information,

$$o_t = (I_t, D_t, \mathbf{p}_t), \tag{2}$$

where I_t is the RGB observation, D_t is the depth observation, and $\mathbf{p}_t$ is the estimated agent pose. The agent maintains an occupancy map M_t^{occ}, semantic value map M_t^{sem}, confidence map C_t, and online scene graph $G_t^{\text{scene}} = (V_t^{\text{scene}}, E_t^{\text{scene}})$.

The goal is to generate a sequence of actions

$$\mathcal{A}_{1:T} = \{a_1, a_2, \ldots, a_T\} \tag{3}$$

that navigates the agent to the target within a step budget $T_{\max}$. ForesightNav uses a training-free decision policy composed of frozen foundation models, semantic mapping, graph matching, and frontier planning:

$$a_t = \Pi\left(g, o_{1:t}, M_t^{\text{occ}}, M_t^{\text{sem}}, C_t, G_t^{\text{scene}}\right). \tag{4}$$

No task-specific policy parameters were optimized during navigation. This objective is evaluated using standard metrics, including the Success Rate (SR) and Success weighted by the Path Length (SPL).

IV. METHOD

ForesightNav addresses the trade-off between efficient exploration and precise target localization in universal zero-shot goal-oriented navigation. As shown in Fig. 2, the framework consists of three closely coupled modules: a context-aware semantic value map, an adaptive exploration strategy, and graph-based target verification. The context-aware semantic value map estimates goal relevance over the explored space by incorporating the current observation, recent navigation history, and VLM-based semantic scores. This map provides coarse spatial guidance when the target is not directly visible or when the observed scene graph is still sparse.

The adaptive exploration strategy determines how the agent selects the next long-term goal. It uses semantic-frontier dispersion and graph-matching confidence to switch among Geometric Exploration, Semantic Exploration, Partial-Match Exploration, and Target Verification. When semantic evidence is weak, the agent expands the map

geometrically. When semantic scores become concentrated, it follows semantically guided frontiers. During Semantic Exploration, high-value frontiers are clustered and ordered using a TSP-based strategy to reduce unnecessary backtracking during long-range exploration.

Graph-based target verification provides fine-grained confirmation when sufficient object and relation evidence has been observed. A partial graph match guides the agent toward potentially relevant regions, while a high-confidence match activates target verification. In this way, ForesightNav does not rely on semantic mapping or graph matching alone. Instead, it progressively coordinates semantic guidance, geometric coverage, and graph-based verification according to the reliability of available evidence. The overall navigation loop is summarized in Algorithm 1.

Algorithm 1 ForesightNav Navigation Loop

Require: Multimodal goal g, maximum steps $T_{\max}$
Ensure: Navigation action sequence $\mathcal{A}_{1:T}$
Initialize M_0^{occ}, M_0^{sem}, C_0, G_0^{scene}, and goal graph G^{goal}
for $t = 1$ to $T_{\max}$ **do**
  Receive observation $o_t = (I_t, D_t, \mathbf{p}_t)$
  Update occupancy map M_t^{occ}
  Update semantic value map M_t^{sem}
  Update scene graph G_t^{scene}
  Extract frontiers F_t and compute scores $\mathcal{S}_F$
  if semantic cues are sufficient **then**
    Compute graph-matching score $S_t = \text{Match}(G_t^{\text{scene}}, G^{\text{goal}})$
  else
    Set $S_t = 0$
  end if
  Select long-term goal $\mathbf{f}_t^{\text{goal}}$ using Algorithm 2
  Execute local policy toward $\mathbf{f}_t^{\text{goal}}$ and obtain action a_t
  Append a_t to $\mathcal{A}_{1:T}$
  if target is verified **then**
    break
  end if
end for
return $\mathcal{A}_{1:T}$

A. CONTEXT-AWARE SEMANTIC VALUE MAP

Graph matching can be unreliable when the explored scene contains few goal-relevant entities. To provide broad guidance under semantic sparsity, ForesightNav maintains a semantic value map M_t^{sem} that estimates goal relevance over explored space. The map is updated from context-aware VLM scores and is used to identify promising frontiers before reliable graph matching becomes available.

1) CONTEXT-INJECTED VLM QUERIES

At each timestep, the VLM receives the current RGB observation I_t, the goal g, and a structured context prompt

$$h_t = \{s_t^{\text{hist}}, \mathbf{p}_t, \mathcal{O}_t^{\text{recent}}, \mathcal{Q}_t^{\text{recent}}\}, \tag{5}$$

Algorithm 2 Adaptive Semantic-Geometric Exploration

Require: Frontiers F_t, semantic scores $\mathcal{S}_F$, graph score S_t, thresholds τ_{partial}, τ_{full}, τ_r, τ_σ, β
Ensure: Long-term goal $\mathbf{f}_t^{\text{goal}}$
Compute score ratio R and standard deviation σ_s
if $S_t \geq \tau_{\text{full}}$ **then**
  $\mathbf{f}_t^{\text{goal}} \leftarrow \text{TargetVerification}(G_t^{\text{scene}}, G^{\text{goal}})$
else if $S_t \geq \tau_{\text{partial}}$ **then**
  $\mathbf{f}_t^{\text{goal}} \leftarrow \text{PartialMatchExploration}(G_t^{\text{scene}}, G^{\text{goal}}, M_t^{\text{sem}})$
else if $R > \tau_r$ and $\sigma_s > \tau_\sigma$ **then**
  $\mathbf{f}_t^{\text{goal}} \leftarrow \text{SemanticExplorationTSP}(F_t, \mathcal{S}_F)$
else
  $\mathbf{f}_t^{\text{goal}} \leftarrow \text{GeometricExploration}(F_t, M_t^{\text{occ}})$
end if
return f_t^{goal}

where s_t^{hist} is a long-term exploration summary, p_t is the current pose, $\mathcal{O}_t^{recent}$ denotes recently observed objects, and $\mathcal{Q}_t^{recent}$ denotes recent VLM interactions. The VLM outputs a semantic relevance score and a confidence score:

$$(s_t, c_t) = \text{VLM}(I_t, g, h_t), \quad s_t, c_t \in [0, 1]. \tag{6}$$

The scored visual region was projected onto a global map using depth observation D_t and pose p_t.

2) MAP UPDATE MECHANISM

Let Ω_t be a set of grid cells projected from the scored visual region. For each $(x, y) \in \Omega_t$, the semantic value is updated using confidence-weighted accumulation:

$$M_t^{\text{sem}}(x, y) = \frac{C_{t-1}(x, y) M_{t-1}^{\text{sem}}(x, y) + c_t(x, y) s_t(x, y)}{C_{t-1}(x, y) + c_t(x, y) + \epsilon}, \tag{7}$$

where ϵ is a small constant for numerical stability. For all $(x, y) \in \Omega_t$, we set $s_t(x, y) = s_t$ and $c_t(x, y) = c_t$, i.e., the per-frame VLM relevance and confidence scores from Eq. (6) are uniformly broadcast over the cells projected from the current scored region. For cells outside Ω_t, $M_t^{sem}(x, y)$ and $C_t(x, y)$ are inherited from the previous step. The confidence map is updated as follows:

$$C_t(x, y) = C_{t-1}(x, y) + c_t(x, y). \tag{8}$$

where $C_t(x, y)$ denotes the cumulative confidence of grid cell (x, y) up to time step t. This formula ensures that high-confidence observations have a greater impact, while gradually building a stable semantic understanding of the environment.

B. ADAPTIVE EXPLORATION STRATEGY

The adaptive strategy determines whether an agent should expand the map geometrically, investigate high-value semantic frontiers, or verify a potential target through graph matching. The decision is based on two signals: dispersion of semantic frontier scores and graph-matching confidence. The procedure is summarized in Algorithm 2.

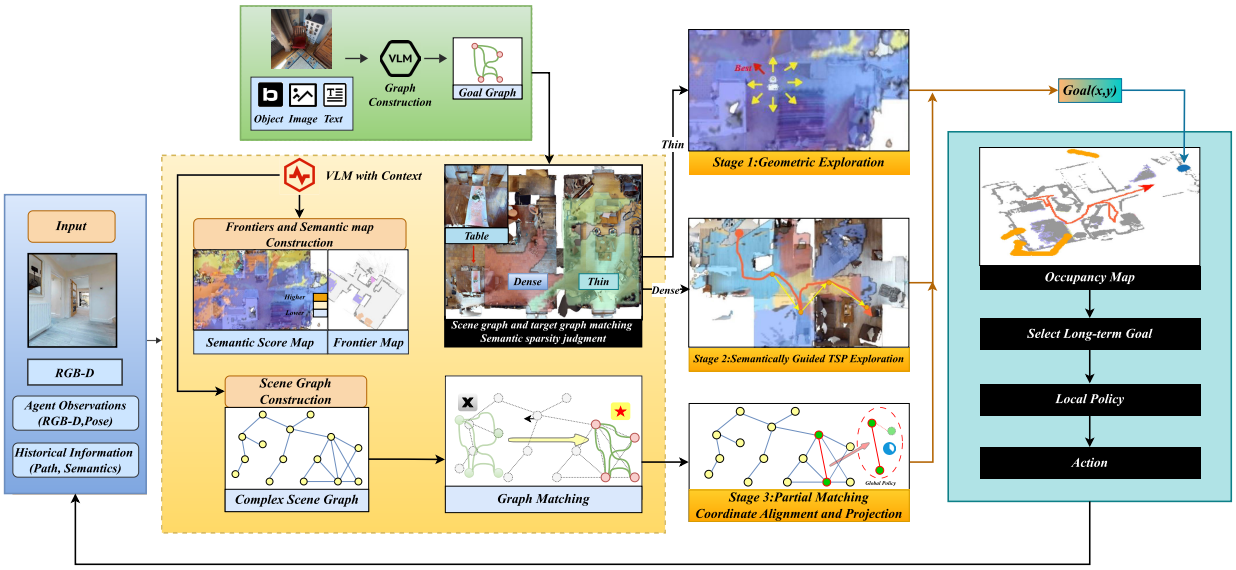


FIGURE 2. Overview of ForesightNav. A multimodal goal, specified as an ObjectNav, InstanceNav, or TextNav target, is converted into a goal graph and processed together with RGB-D observations. The framework updates a context-aware semantic value map for large-scale guidance and an online scene graph for fine-grained verification. The adaptive planner switches among Geometric Exploration, Semantic Exploration with TSP-based frontier ordering, Partial-Match Exploration, and Target Verification according to semantic-map dispersion and graph-matching confidence.

1) EXPLORATION MODE DECISION

Given frontier scores $\mathcal{S}_F = \{s(f) \mid f \in F_t\}$, we compute the score ratio and standard deviation:

$$R = \frac{\max(\mathcal{S}_F)}{\text{mean}(\mathcal{S}_F) + \epsilon}, \quad \sigma_s = \text{std}(\mathcal{S}_F). \tag{9}$$

A high R and σ_s indicate that a small number of frontiers are more goal-relevant than others. The exploration mode was selected as follows:

$$m_t = \begin{cases} \text{TV}, & S_t \geq \tau_{\text{full}}, \\ \text{PME}, & \tau_{\text{partial}} \leq S_t < \tau_{\text{full}}, \\ \text{SE}, & S_t < \tau_{\text{partial}} \wedge R > \tau_r \wedge \sigma_s > \tau_\sigma, \\ \text{GE}, & \text{otherwise}. \end{cases} \tag{10}$$

where TV, PME, SE, and GE denote the Target Verification, Partial-Match Exploration, Semantic Exploration, and Geometric Exploration, respectively.

If the cues are weak or ambiguous (low R and σ_s), the system defaults on Geometric Exploration to efficiently expand the map. If strong, concentrated cues are found, it switches to Semantic Exploration.

2) GEOMETRIC EXPLORATION

When semantic cues are weak or uniformly distributed, the agent prioritizes the map coverage by selecting an informative frontier while penalizing long travel distances and recently visited regions. This mode prevents unreliable semantic or graph-based reasoning from dominating the early exploration.

3) SEMANTIC EXPLORATION WITH TSP-BASED FRONTIER ORDERING

When semantic scores become concentrated, the agent investigates the high-value frontiers. First, we select semantic frontiers by

$$F_t^+ = \{\mathbf{f} \in F_t \mid s(\mathbf{f}) \geq \mu_s + \beta\sigma_s\}, \tag{11}$$

where μ_s and σ_s are the mean and standard deviation of frontier scores, respectively, and β is a sensitivity coefficient that controls the strictness of the frontier selection. To remove spatial redundancy, DBSCAN clusters F_t^+ and selects the highest-scoring frontier from each cluster, yielding representative frontiers

$$\hat{F}_t = \{\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_K\}. \tag{12}$$

We then compute their visiting order by solving a TSP-style optimization:

$$\rho^* = \arg\min_{\rho} \left[\text{dist}(\mathbf{p}_t, \hat{\mathbf{f}}_{\rho_1}) + \sum_{i=1}^{K-1} \text{dist}(\hat{\mathbf{f}}_{\rho_i}, \hat{\mathbf{f}}_{\rho_{i+1}}) \right], \tag{13}$$

where ρ is a permutation of the frontier indices. The next long-term goal is selected as

$$f_t^{goal} = \hat{f}_{\rho_1^*}. \tag{14}$$

In practice, clustering maintains K small, making the TSP solver efficient for online planning.

C. GRAPH-BASED TARGET VERIFICATION

A semantic value map provides coarse spatial guidance, but precise localization requires fine-grained target verification.

TABLE 1. Comparison with representative methods across goal-oriented navigation tasks and datasets. The ↑ symbol indicates that higher values are better. Bold and underlined values indicate the best and second-best results, respectively. IEVE† uses Mask R-CNN backbone (InternImage variant: SR=66.0, SPL=23.7).

Method	Zero-Shot	ObjNav						InsNav		TextNav	
		MP3D		HM3D		RoboTHOR		HM3D		HM3D	
		SR(%)↑	SPL↑	SR(%)↑	SPL↑	SR(%)↑	SPL↑	SR(%)↑	SPL↑	SR(%)↑	SPL↑
SemEXP (NeurIPS,2020) [2]	×	32.3	13.2	—	—	—	—	—	—	—	—
ZSON (NeurIPS,2022) [30]	×	13.5	4.0	20.6	10.4	—	—	—	—	—	—
OVRL-v2 [31]	×	—	—	55.3	25.3	—	—	—	—	—	—
OVRL-v2-IIN [31]	×	—	—	—	—	—	—	21.0	11.0	—	—
IEVE† (CVPR,2024) [33]	×	—	—	—	—	—	—	<u>62.5</u>	**28.0**	<u>28.7</u>	<u>11.4</u>
PSL (ECCV,2024) [8]	×	—	—	37.8	16.6	—	—	20.4	10.5	13.7	6.7
GOAT (CVPR,2023) [7]	×	—	—	45.6	21.1	—	—	32.4	13.6	14.1	7.3
Mod-IIN (ICCV,2023) [6]	×	—	—	—	—	—	—	49.8	19.7	23.7	10.5
OpenFMNav (ECCV,2024) [27]	✓	32.6	13.2	46.2	20.5	39.3	20.5	44.9	17.1	16.9	7.7
VLFM (ICRA,2024) [28]	✓	32.1	13.3	45.6	26.1	37.6	20.4	44.4	16.9	16.5	7.5
SG-Nav (NeurIPS,2024) [3]	✓	36.3	14.0	46.5	20.8	43.0	21.4	45.6	17.4	17.4	8.0
TriHelper (IROS,2024) [34]	✓	38.1	13.5	51.9	23.6	46.2	20.8	50.6	19.3	21.4	8.9
ImagineNav (ICLR,2025) [35]	✓	35.7	12.8	49.2	22.2	43.6	19.6	48.9	18.6	19.7	8.5
InstructNav (CoRL,2024) [36]	✓	39.0	11.3	53.0	19.6	47.3	17.4	51.2	18.0	22.6	8.3
Think-Rem-Nav (NeurIPS-W,2025) [37]	✓	33.4	16.4	50.3	28.9	44.6	25.3	49.9	19.6	20.9	9.2
WMNav (IROS,2025) [38]	✓	41.8	15.9	53.1	29.0	47.4	25.4	50.9	19.9	22.1	9.6
ApexNav (RA-L,2025) [39]	✓	36.3	16.5	54.1	30.6	48.4	26.8	51.8	20.2	23.0	9.9
STRIVE (ICRA,2026) [40]	✓	<u>46.2</u>	<u>20.5</u>	<u>56.6</u>	<u>31.8</u>	<u>60.9</u>	**33.6**	61.3	25.5	28.0	11.1
GAMap (NeurIPS,2024) [41]	✓	34.6	13.9	49.3	24.3	43.7	21.3	49.4	19.0	20.2	8.6
UniGoal (CVPR,2025) [12]	✓	36.8	14.3	47.8	22.9	43.9	21.5	53.1	20.7	17.9	9.3
ForesightNav (Ours)	✓	**47.5**	**21.2**	**57.8**	**33.4**	**62.5**	<u>32.9</u>	**65.3**	<u>26.8</u>	**29.6**	**11.8**

Therefore, ForesightNav maintains a scene graph G_t^{scene} in parallel with the semantic map. The multimodal goal is converted into a goal graph G^{goal}, and a graph-matching score

$$S_t = \text{Match}(G_t^{\text{scene}}, G^{\text{goal}}), \quad S_t \in [0, 1], \tag{15}$$

measures the consistency between the observed scene and the target goal.

Depending on S_t, the adaptive planner uses graph matching in three regimes:

- **No reliable match** ($S_t < \tau_{\text{partial}}$): The observed scene graph is insufficient for target verification, and the agent relies on geometric or semantic frontier exploration.
- **Partial match** ($\tau_{\text{partial}} \leq S_t < \tau_{\text{full}}$): Potential target-related entities are observed, and the agent performs local semantic exploration around the matched region.
- **High-confidence match** ($S_t \geq \tau_{\text{full}}$): The agent activates target verification and navigates toward the matched target region.

Our graph construction and matching followed the unified graphical representation of UniGoal [12]. In contrast to purely graph-centered exploration, ForesightNav uses graph matching as an adaptive verification signal and couples it with semantic value mapping and global frontier ordering. This design avoids over-reliance on unreliable graph matches during early exploration while preserving fine-grained target localization when sufficient semantic evidence is available.

V. EXPERIMENTS

We evaluate ForesightNav through quantitative comparisons, foundation-model analysis, ablations, and qualitative visualizations. The experiments focus on the following three questions:

Q1: How does ForesightNav compare with the representative task-specific, universal, supervised, and training-free baselines?

Q2: How robust is ForesightNav across different foundation-model backbones, and does its benefit persist under both strong and lightweight models?

Q3: How much does each core component contribute to navigation performance, and what qualitative behaviors explain these gains?

A. EXPERIMENTAL SETUP

1) PLATFORM AND MODELS

All the experiments were conducted using the Habitat simulator with RGB-D observations and pose information. `Gemma-3-12B-It` [32] was used as the frozen LLM/VLM backbone. Task-specific policy training was not conducted.

2) DATASETS AND TASKS

We evaluate three goal modalities. ObjectNav is tested on MP3D [13], HM3D [14], and RoboTHOR [15] following SG-Nav [3]. InstanceNav is tested on HM3D following Mod-IIN [6], and TextNav is tested on HM3D following PSL [8].

TABLE 2. Impact of different foundation models on navigation performance. We evaluate our method (ForesightNav) against a zero-shot baseline (InstructNav) using a wide range of state-of-the-art closed-source and open-source models on the HM3D dataset. * denotes the default model configuration used in our main experiments.

Method	Type	Foundation Model	ObjNav		InsNav		TextNav	
			SR(%)↑	SPL↑	SR(%)↑	SPL↑	SR(%)↑	SPL↑
InstructNav [36]	Closed	Claude-3.7-Sonnet	54.7	20.2	53.1	19.0	24.9	9.2
		GPT-4.1	54.3	19.9	52.6	18.6	24.4	8.9
		Gemini-2.0-flash	54.0	19.7	52.5	18.3	24.0	8.6
		GPT-4o	53.0	19.6	51.2	18.0	22.6	8.3
		Claude-3.5-Sonnet	52.2	19.0	49.8	17.4	21.2	7.7
		GPT-4o-mini	42.3	15.5	39.5	13.6	11.8	4.2
		Gemini-1.5-Pro	39.0	14.1	36.2	12.2	9.9	3.6
	Open	gemma-3-27b-it	52.6	19.3	50.3	17.7	21.6	8.0
		Qwen2.5-VL-72B-Ins	51.5	18.6	48.9	16.9	20.2	7.3
		Llama-3.2-90B-Vision-Ins	50.9	18.3	48.4	16.7	19.7	7.1
		Gemma-3-12B-It	53.0	19.6	46.5	15.5	18.0	6.4
		Qwen2.5-VL-7B-Ins	45.6	16.2	42.3	14.3	14.6	5.2
		Llama-3.2-11B-Vision-Ins	40.9	14.9	38.1	13.0	10.8	3.9
		Qwen2.5-VL-3B-Ins	32.9	11.8	30.1	9.9	8.0	2.8
Ours: ForesightNav								
ForesightNav	Closed	Claude-3.7-Sonnet	**60.0**	**35.0**	**68.2**	**28.9**	**33.4**	**13.9**
		GPT-4.1	59.7	34.8	67.7	28.7	32.7	13.6
		Gemini-2.0-flash	59.4	34.6	67.2	28.4	32.1	13.3
		GPT-4o	59.0	34.3	66.9	28.0	32.0	13.0
		Claude-3.5-Sonnet	58.3	33.8	66.0	27.3	30.6	12.2
		GPT-4o-mini	53.1	30.1	59.7	23.5	22.1	8.3
		Gemini-1.5-Pro	51.2	28.7	57.3	22.1	19.7	7.3
	Open	gemma-3-27b-it	58.8	34.0	66.6	27.7	31.5	12.7
		Qwen2.5-VL-72B-Ins	57.5	33.1	64.9	26.5	29.0	11.3
		Llama-3.2-90B-Vision-Ins	57.2	32.7	64.4	26.1	28.4	11.1
		Gemma-3-12B-It*	**57.8**	**33.4**	**65.3**	**26.8**	**29.6**	**11.8**
		Qwen2.5-VL-7B-Ins	55.0	31.5	61.6	24.9	24.9	9.6
		Llama-3.2-11B-Vision-Ins	52.2	29.6	58.8	23.0	21.2	8.0
		Qwen2.5-VL-3B-Ins	47.5	26.8	53.1	19.7	17.4	6.1

3) METRICS

We report the Success Rate (SR) and Success weighted by Path Length (SPL). The SR measures the percentage of episodes in which the agent reaches within 1 meter of the target under the step limit. The SPL further penalizes inefficient paths in successful episodes. Higher values were better for both metrics.

4) PARAMETERS

We set $\tau_r = 1.10$ and $\tau_\sigma = 0.015$ for adaptive exploration. The graph-matching thresholds for mode switching are set to $\tau_{\text{partial}} = 0.30$ and $\tau_{\text{full}} = 0.65$ (Eq. 10), and the frontier sensitivity coefficient β in Eq. 11 is set to 1.0. DBSCAN uses `eps`=30 pixels, approximately 1.5 m, and `min_samples`=1. The short-term VLM memory length was set to 5.

All quantitative comparisons are based on 10 independent runs with different random seeds for the VLM sampling module. Statistical significance between ForesightNav and STRIVE is assessed using Welch's t-test. Detailed statistical results are provided in Section V-G.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

To answer **Q1**, we compare ForesightNav with a diverse set of state-of-the-art methods, including task-specific zero-shot approaches, supervised universal frameworks, and recent training-free universal baselines. These baselines differ in training requirements, supported goal modalities, and planning mechanisms, which allows us to examine not only absolute navigation performance but also the trade-off among task generality, training-free deployment, and path efficiency. Fig. 1 provides a comparative overview of the representative methods on the HM3D benchmark under different goal modalities. The results suggest that ForesightNav can improve path efficiency relative to UniGoal, as reflected by SPL, while maintaining competitive performance across ObjectNav, InstanceNav, and TextNav. The quantitative results are summarized in Table 1 and Fig. 3.

On ObjectNav, ForesightNav achieves an SR of 47.5% and an SPL of 21.2 on MP3D, an SR of 57.8% and an SPL of 33.4 on HM3D, and an SR of 62.5% and an SPL of 32.9 on RoboTHOR. Compared with UniGoal [12], a recent training-free universal baseline, ForesightNav improves SR by 10.7, 10.0, and 18.6 percentage points on MP3D, HM3D, and RoboTHOR ObjectNav, respectively. The corresponding SPL improvements are 6.9, 10.5, and 11.4, respectively. These results indicate that the proposed adaptive semantic-geometric planner provides more effective exploration guidance than relying primarily on graph matching, especially when early scene observations are sparse or incomplete.

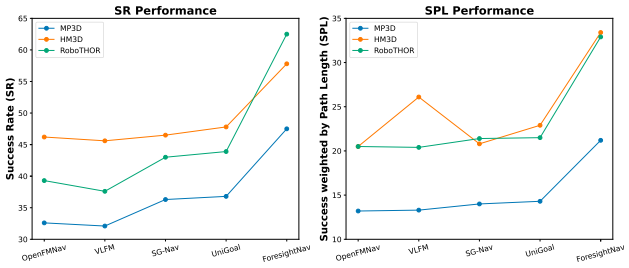


FIGURE 3. ObjectNav performance on MP3D, HM3D, and RoboTHOR. ForesightNav improves over prior universal baselines on both SR and SPL, with particularly clear gains in path efficiency.

Compared with other zero-shot ObjectNav methods, such as OpenFMNav [27], VLFM [28], and SG-Nav [3], ForesightNav also maintains favorable performance across the three ObjectNav datasets. For example, on HM3D ObjectNav, ForesightNav achieves an SR of 57.8% and an SPL of 33.4. In comparison, the SR and SPL are 46.2% and 20.5 for OpenFMNav, 45.6% and 26.1 for VLFM, and 46.5% and 20.8 for SG-Nav, respectively. This suggests that combining context-aware semantic value mapping with adaptive graph-based verification can provide more stable guidance than using semantic frontier maps or scene-graph reasoning alone.

The comparison with STRIVE [40] and IEVE [33] further clarifies the behavior of ForesightNav under different evaluation metrics. On RoboTHOR ObjectNav, STRIVE achieves a higher SPL of 33.6, compared with 32.9 for ForesightNav. This result suggests that STRIVE can generate more compact successful trajectories on this ObjectNav benchmark, possibly because its reasoning and planning process is more directly optimized for path compactness in category-level object search. On HM3D InstanceNav, IEVE records a higher SPL of 28.0, compared with 26.8 for ForesightNav. InstanceNav requires precise target-instance localization, and an instance-oriented verification pipeline can shorten successful trajectories once reliable instance evidence is detected. In contrast, ForesightNav uses a unified adaptive semantic-geometric planning process rather than a task-specific path optimizer. Therefore, these SPL differences reflect the effect of task specialization on successful-episode path compactness, while ForesightNav is designed to maintain general applicability across ObjectNav, InstanceNav, and TextNav without task-specific training or modality-dependent optimization.

For TextNav on HM3D, ForesightNav achieves an SR of 29.6% and an SPL of 11.8. In comparison, UniGoal achieves an SR of 17.9% and an SPL of 9.3, while IEVE achieves an SR of 28.7% and an SPL of 11.4. This result suggests that the proposed framework remains applicable when the target is specified by language rather than by an object category or instance image. Overall, the table-level comparison shows that ForesightNav provides a practical balance among generality, training-free deployment, target-reaching reliability, and navigation efficiency, highlighting the benefit of foresight-driven planning in universal zero-shot goal-oriented navigation.

C. IMPACT OF FOUNDATION MODELS

To answer **Q2**, we evaluate the sensitivity of ForesightNav to different foundation-model backbones, as shown in Table 2. The comparison is conducted against InstructNav [36] under the same backbone settings, covering both closed-source and open-source models. This setting allows us to examine whether the performance gain mainly comes from stronger foundation models or from the proposed planning framework itself.

Across the tested backbones, ForesightNav improves both SR and SPL over InstructNav when the two methods use the same foundation model. For example, with Claude-3.7-Sonnet, ForesightNav improves ObjectNav SR from 54.7% to 60.0% and SPL from 20.2 to 35.0. On the same backbone, InstanceNav SR increases from 53.1% to 68.2%, and TextNav SR increases from 24.9% to 33.4%. These results suggest that the proposed semantic mapping and adaptive planning components provide navigation benefits beyond the representation capability of the foundation model itself.

The improvement is also observed with open-source backbones. With the default Gemma-3-12B-It backbone, ForesightNav improves ObjectNav from 53.0% SR and 19.6 SPL to 57.8% SR and 33.4 SPL. For InstanceNav, the SR increases from 46.5% to 65.3%, and for TextNav, the SR increases from 18.0% to 29.6%. These gains indicate that ForesightNav remains effective when deployed with an open-source backbone, rather than relying only on closed-source foundation models.

The results with lightweight models further support this observation. When using Qwen2.5-VL-3B-Ins, ForesightNav improves ObjectNav SR from 32.9% to 47.5% and SPL from 11.8 to 26.8. It also improves InstanceNav SR from 30.1% to 53.1% and TextNav SR from 8.0% to 17.4%. Although the absolute performance remains lower than that of larger backbones, the improvement over InstructNav is still clear, suggesting that the proposed planning mechanism can compensate, to some extent, for weaker perception or reasoning capability.

Several trends can be observed from Table 2. First, stronger foundation models generally lead to higher absolute performance, as shown by the closed-source models such as Claude-3.7-Sonnet and GPT-4.1. Second, the benefit of ForesightNav varies across goal modalities: ObjectNav shows particularly evident SPL improvement, while InstanceNav shows large SR improvement. Third, open-source models achieve competitive results within the proposed framework. For instance, Gemma-3-27B-It reaches 58.8% SR and 34.0 SPL on ObjectNav, which is close to the performance of stronger closed-source backbones.

We use Gemma-3-12B-It as the default backbone because it provides a practical balance between navigation

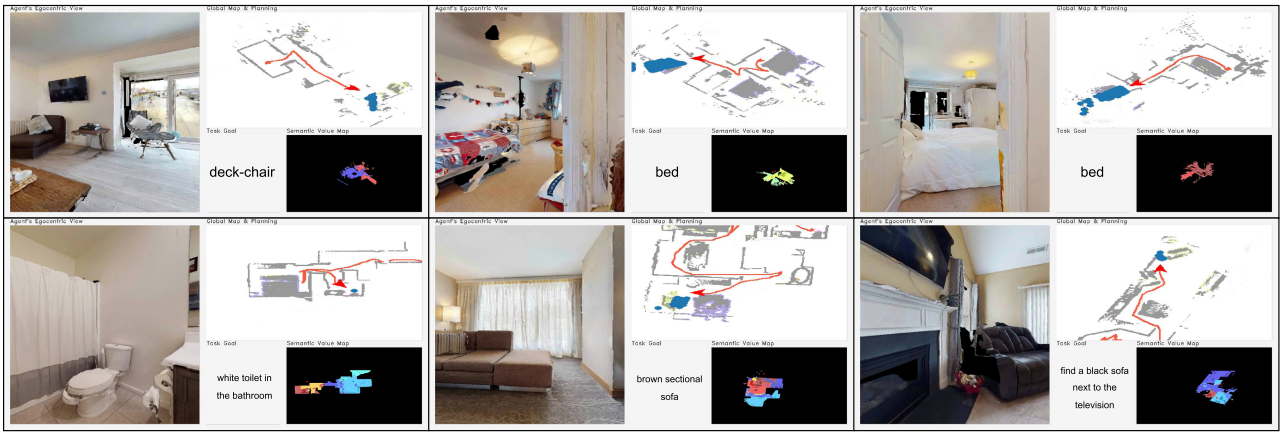


FIGURE 4. Qualitative navigation examples for InstanceNav and TextNav. Each example shows the explored trajectory and task goal in representative scenes, illustrating how ForesightNav uses semantic guidance to approach target regions.

performance and model scale. Compared with Gemma-3-27B-It, Gemma-3-12B-It achieves slightly lower but close performance on HM3D, with 57.8% versus 58.8% SR on ObjectNav and 65.3% versus 66.6% SR on InstanceNav. Considering online navigation requires both accuracy and computational efficiency, Gemma-3-12B-It is a practical default choice. Overall, Table 2 indicates that ForesightNav is not tied to a single foundation model; instead, its adaptive planning design can complement different backbone capabilities in a stable manner.

D. QUALITATIVE ANALYSIS

Figures 4 and 5 provide qualitative examples of how ForesightNav links semantic reasoning with frontier-based exploration. Fig. 4 shows representative InstanceNav and TextNav cases. Each example includes the agent's egocentric observation, the explored map and trajectory, the task goal, and the associated semantic value map. The goals cover both instance-level targets, such as ''deck-chair'' and ''bed'', and language-specified targets, such as ''white toilet in the bathroom'', ''brown sectional sofa'', and ''find a black sofa next to the television''. These examples illustrate that ForesightNav uses a unified planning process for different goal specifications, rather than relying on separate navigation policies for each task.



FIGURE 5. Qualitative examples on ObjectNav. The trajectories illustrate how semantic guidance and global frontier ordering support efficient exploration.

The semantic value maps in Fig. 4 help explain the role of context-aware semantic mapping. As more observations are accumulated, goal-relevant regions become more evident in the explored area. These regions provide coarse semantic guidance for subsequent frontier selection before the target can be verified with high confidence. This process is consistent with the motivation of ForesightNav: the agent first uses semantic evidence to guide exploration and then relies on graph-based verification when more reliable target-related observations are available. For text-specified goals, the semantic value map also reflects contextual information from the description, such as object category, room type, or nearby objects, which helps constrain the search direction.

Fig. 5 presents representative ObjectNav trajectories in larger scenes. The orange curves denote the executed paths, and the background maps show the explored areas. These examples are used to illustrate the effect of global frontier ordering on the overall exploration pattern. Instead of selecting frontiers only according to local or greedy scores, ForesightNav considers high-value frontiers in a more global manner. This helps the agent organize long-range exploration and reduces unnecessary revisits when multiple candidate regions exist.

Overall, the qualitative results complement the quantitative comparisons by showing how the main components of ForesightNav work together during navigation. The semantic value map provides broad goal-directed guidance, graph-based verification supports fine-grained target confirmation, and TSP-based frontier ordering helps structure the exploration route. Fig. 4 mainly explains how semantic evidence is accumulated for instance- and text-specified goals, while Fig. 5 shows how ObjectNav trajectories are organized in larger scenes. These visualizations make the objective of ForesightNav clearer: to improve universal

zero-shot goal-oriented navigation by adaptively combining semantic guidance, geometric exploration, and target verification.

E. ABLATION STUDY

To answer **Q3**, we conducted ablations of HM3D ObjectNav and InstanceNav to quantify the contribution of each module. Table 3 presents the results.

TABLE 3. Ablation study on HM3D ObjectNav and InstanceNav.

Method	ObjNav-HM3D		InsNav-HM3D	
	SR(%)↑	SPL↑	SR(%)↑	SPL↑
ForesightNav (Full)	**57.8**	**33.4**	**65.3**	**26.8**
w/o Adaptive Strategy	54.2	30.5	59.5	23.9
w/o TSP Ordering	54.9	31.3	61.7	25.3
w/o Context-VLM	56.1	32.4	62.9	25.8

The full ForesightNav model achieves an SR/SPL of 57.8%/33.4 on HM3D ObjectNav and 65.3%/26.8 on HM3D InstanceNav. Removing the adaptive strategy causes the largest overall degradation, reducing ObjectNav SR/SPL to 54.2%/30.5 and InstanceNav SR/SPL to 59.5%/23.9. This indicates that fixed exploration is less effective when semantic cues are sparse or when graph-matching evidence is still incomplete. The larger SR drop on InstanceNav suggests that adaptive switching is particularly important for instance-level goals, where premature verification or insufficient local exploration can affect target discovery.

Replacing TSP-based frontier ordering with greedy frontier selection also reduces performance, with ObjectNav SPL decreasing from 33.4 to 31.3 and InstanceNav SPL decreasing from 26.8 to 25.3. This result is consistent with the role of global frontier ordering: when multiple promising frontiers are available, considering them jointly helps organize the exploration route and reduces unnecessary backtracking. The SR decrease also suggests that more organized frontier selection can indirectly support target discovery by avoiding inefficient revisits.

Removing the context-aware VLM component leads to a smaller but consistent performance drop. ObjectNav SR/SPL decreases from 57.8%/33.4 to 56.1%/32.4, and InstanceNav SR/SPL decreases from 65.3%/26.8 to 62.9%/25.8. This suggests that navigation history and recent observations improve semantic scoring, especially when the current observation alone is insufficient to infer the goal relevance of unexplored regions.

Overall, the ablation results show that the three components play complementary roles. The adaptive strategy mainly improves decision robustness under uncertain semantic and graph evidence, TSP-based ordering improves the organization of long-range exploration, and context-aware VLM scoring provides more reliable semantic guidance. Their combination leads to the final performance of ForesightNav on both category-level and instance-level navigation tasks.

F. THRESHOLD SENSITIVITY ANALYSIS

To assess the robustness of ForesightNav to its handcrafted thresholds, we conduct a systematic sensitivity analysis on HM3D ObjectNav. We adopt a one-at-a-time (OAT) protocol: each threshold is independently swept over five uniformly spaced values centered at its default while the other three are fixed; SR and SPL trends are reported in Fig. 6. Because τ_{partial} and τ_{full} are coupled through the structural constraint $\tau_{\text{full}} > \tau_{\text{partial}}$, we additionally conduct a 5×5 joint grid search over these two parameters (Fig. 7). The two analyses are independent runs; minor SR differences at shared parameter points (≤ 2 pp) reflect normal simulation variance and do not alter any qualitative conclusion.

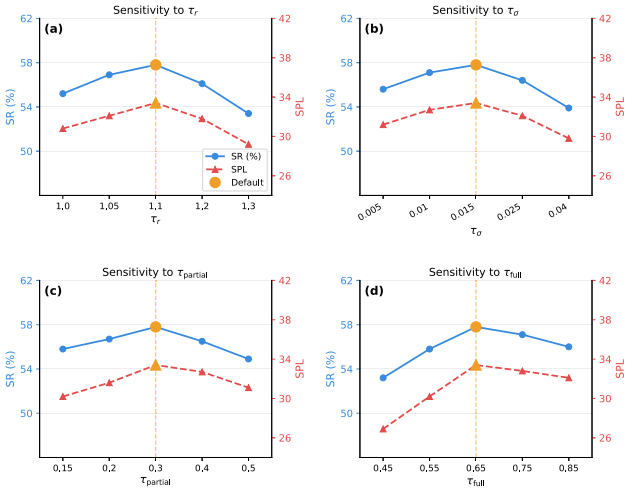


FIGURE 6. Threshold sensitivity analysis on HM3D ObjectNav (OAT). Each panel independently varies one threshold across five values while the other three are fixed at their defaults. Solid blue (left axis): SR (%); dashed red (right axis): SPL. Vertical dashed lines mark the default value. (a) τ_r. (b) τ_σ. (c) τ_{partial}. (d) τ_{full}.

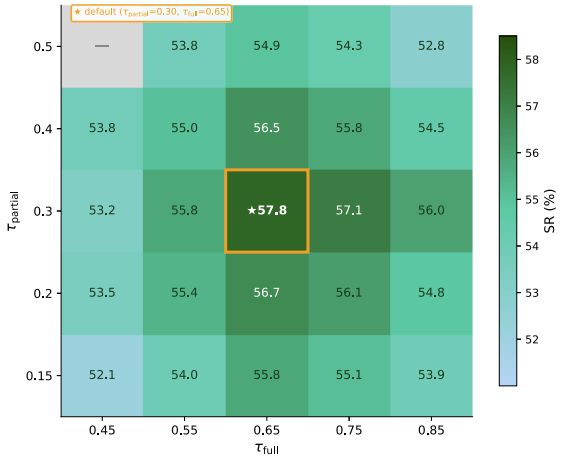


FIGURE 7. Joint sensitivity of τ_{partial} and τ_{full} on HM3D ObjectNav SR (%). Gray cells indicate invalid configurations ($\tau_{\text{full}} \leq \tau_{\text{partial}}$). The highlighted cell (★) denotes the default (τ_{partial}=0.30, τ_{full}=0.65).

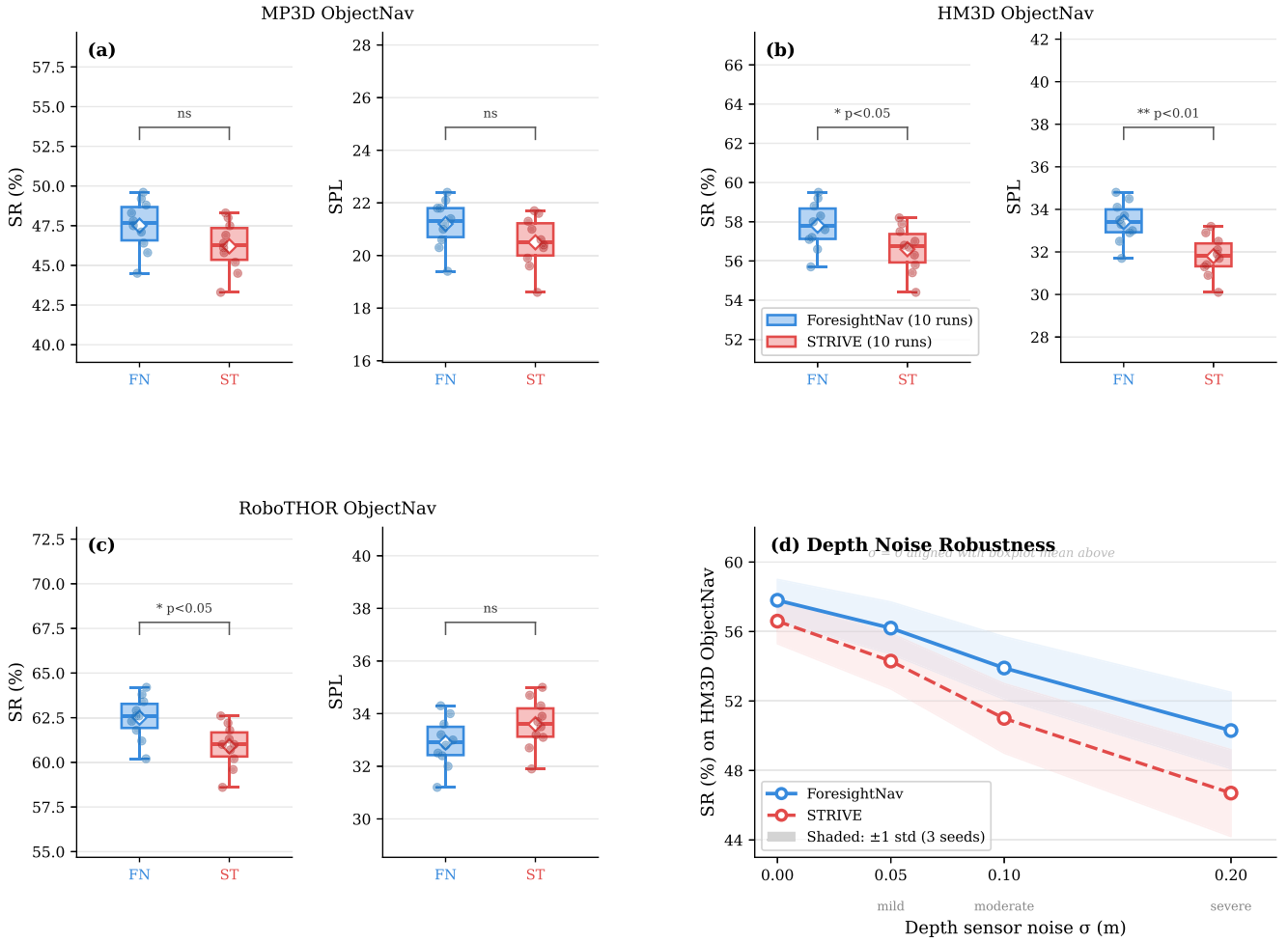


FIGURE 8. **Statistical stability and depth-noise robustness of ForesightNav. (a)–(c) Distributions of SR (%) and SPL over 10 independent runs for ForesightNav (blue) and STRIVE (red) on MP3D, HM3D, and RoboTHOR ObjectNav, respectively. Boxes: IQR; center lines: median; diamonds: mean; whiskers: full range; scatter points: individual run results. Significance from Welch's t-test: $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$. (d) SR under increasing Gaussian depth-sensor noise (σ m) on HM3D ObjectNav (3 seeds per noise level). Lines: mean; shaded bands: ±1 std. The performance gap between ForesightNav and STRIVE widens from 1.2 to 3.6 percentage points as noise increases, attributable to the adaptive fallback to Geometric Exploration under degraded semantic signals.**

All four thresholds yield peak performance at their defaults and degrade gracefully under perturbation. The dominant finding is the strong asymmetric sensitivity of τ_{full}: reducing τ_{full} from 0.65 to 0.45 decreases SR by 4.6 pp and SPL by 6.5, because an overly low threshold activates Target Verification before the scene graph is sufficiently populated, generating false-positive stops and degrading path efficiency in successful episodes. Conversely, raising τ_{full} to 0.85 causes only a 1.8-pp SR decrease and a 1.3-SPL decrease, as the agent merely extends its Partial-Match Exploration phase before locating the target. The lower-side-to-upper-side drop ratio of approximately 2.6 : 1 (4.6 pp vs. 1.8 pp) directly supports $\tau_{\text{full}} = 0.65$ as a precision-favoring threshold. For τ_{partial}, deviations in either direction yield moderate and comparable drops of 2–3 pp (2.0 pp SR and 3.2 SPL at 0.15; 2.9 pp SR and 2.3 SPL at 0.50), indicating broad robustness to this parameter with no strong directional bias. The τ_r and τ_σ curves are more symmetric, with the largest drops at $\tau_r = 1.30$ (4.4 pp SR, 4.2 SPL) and $\tau_\sigma = 0.040$ (3.9 pp SR, 3.6 SPL).

The joint heatmap (Fig. 7) confirms that the default configuration ($\tau_{\text{partial}} = 0.30$, $\tau_{\text{full}} = 0.65$) achieves the highest SR (57.8%) among all evaluated combinations. Along the τ_{full} axis (with τ_{partial} fixed at 0.30), SR increases monotonically from 53.2% at τ_{full}=0.45 to the peak of 57.8%, then stabilizes (56.0% at τ_{full}=0.85). Along the τ_{partial} axis (with τ_{full} fixed at 0.65), SR declines toward both extremes (55.8% at τ_{partial}=0.15; 54.9% at τ_{partial}=0.50). Performance degrades smoothly from this optimum, with the steepest drops concentrated in the minimum-τ_{full} column, consistent with the OAT finding that τ_{full} is the dominant parameter. Near-diagonal configurations—where τ_{full} only marginally exceeds τ_{partial}, leaving an extremely narrow PME window—also underperform, reinforcing the necessity of a

sufficient separation between these two thresholds. The broad high-performance plateau surrounding the default indicates that ForesightNav does not require fine threshold tuning within a reasonable operating range.

G. STATISTICAL STABILITY AND ROBUSTNESS ANALYSIS

To quantify the statistical stability of ForesightNav, we conduct 10 independent runs with different VLM sampling random seeds for both ForesightNav and STRIVE—the overall best-performing baseline—on the MP3D, HM3D, and RoboTHOR ObjectNav datasets. The Habitat evaluation set is fixed and deterministic across runs; run-to-run variability arises solely from VLM temperature sampling, providing a faithful measure of performance stability. Statistical significance is assessed via Welch's t-test at $\alpha = 0.05$. To evaluate robustness to sensor perturbations, we additionally inject Gaussian depth noise ($\sigma \in \{0,\ 0.05,\ 0.10,\ 0.20\}$ m) into depth observations on HM3D ObjectNav, with 3 seeds per noise level.

Fig. 8(a)–(c) present the run-level SR and SPL distributions of both methods across three datasets. ForesightNav consistently achieves higher SR than STRIVE: on MP3D, a positive trend is observed ($47.5 \pm 1.6\%$ vs. $46.2 \pm 1.6\%$, $p = 0.083$); on HM3D, the improvement reaches statistical significance ($57.8 \pm 1.2\%$ vs. $56.6 \pm 1.2\%$, $p = 0.037$); and on RoboTHOR, the advantage is highly significant ($62.5 \pm 1.2\%$ vs. $60.9 \pm 1.2\%$, $p = 0.008$). For SPL, the improvement on HM3D is the most pronounced (33.4 ± 0.9 vs. 31.8 ± 0.9, $p = 0.001$), confirming the practical path-efficiency gains from TSP-based frontier ordering. On RoboTHOR, STRIVE achieves a marginally higher SPL of 33.6 ± 0.9 compared with 32.9 ± 0.9 for ForesightNav; this difference does not reach statistical significance ($p = 0.114$), consistent with the qualitative discussion in Section V-B. The SR standard deviation across 10 runs remains below 1.7 percentage points for all methods, demonstrating stable performance against VLM sampling randomness.

Fig. 8(d) presents the depth-noise robustness results on HM3D ObjectNav. As σ increases from 0 to 0.20 m, the SR of ForesightNav decreases from 57.8% to 50.3% (a drop of 7.5 percentage points), compared with a larger drop of 9.9 percentage points for STRIVE (from 56.6% to 46.7%); the performance gap consequently widens from 1.2 to 3.6 percentage points, demonstrating that ForesightNav degrades more gracefully under increasing sensor noise. This advantage arises from the adaptive planning mechanism: when depth noise corrupts the semantic score distribution (reducing σ_s), the planner automatically falls back to Geometric Exploration, which relies only on the occupancy map and is insensitive to degraded semantic signals; when a noisy scene graph yields a low graph-matching score $S_t < \tau_{\text{partial}}$, neither Partial-Match Exploration nor Target Verification is triggered, preventing false-positive stopping decisions. Moreover, depth noise induces map-projection errors that are mechanistically equivalent to pose-estimation drift—in both cases, projected semantic scores and scene-graph node coordinates are spatially offset from their true positions—so these results also provide indirect evidence of robustness to pose uncertainty. Beyond sensor perturbations, Table 2 evaluates ForesightNav across 14 backbone models spanning a wide capability range from lightweight 3B-parameter models to state-of-the-art closed-source models, constituting a systematic robustness analysis under varying levels of VLM perceptual reliability; ForesightNav consistently outperforms InstructNav on the same backbone across all configurations, indicating that the adaptive planning framework compensates for weaker backbone capability. Furthermore, the consistent performance across the three benchmarks in Table 1, which represent markedly different levels of scene complexity (large-scale and intricate for MP3D, moderately diverse for HM3D, and compact and densely furnished for RoboTHOR), further demonstrates robustness to scene-complexity variation without task-specific tuning.

VI. CONCLUSION

This paper presented ForesightNav, a training-free framework for universal zero-shot goal-oriented navigation. The core idea is to adaptively combine semantic guidance, geometric exploration, and graph-based target verification within a unified planning process. Specifically, ForesightNav maintains a context-aware semantic value map for broad goal-directed exploration and an online scene graph for fine-grained target confirmation. According to semantic-score dispersion and graph-matching confidence, the planner switches among Geometric Exploration, Semantic Exploration, Partial-Match Exploration, and Target Verification. TSP-based frontier ordering is further used to organize high-value frontiers and reduce unnecessary backtracking. Experiments on MP3D, HM3D, and RoboTHOR demonstrate that ForesightNav provides favorable performance across different goal modalities. The comparison with UniGoal shows clear improvements in ObjectNav, while the HM3D results further indicate that the same training-free planning process remains applicable to instance- and language-specified goals. These findings suggest that adaptive semantic-geometric planning can improve target-reaching reliability while maintaining competitive navigation efficiency.

The foundation-model analysis shows that ForesightNav is not tied to a single backbone, as improvements over InstructNav are observed across closed-source, open-source, and lightweight models. The ablation study further confirms the contributions of adaptive switching, context-aware semantic mapping, and TSP-based frontier ordering. Qualitative examples also illustrate how semantic value maps guide exploration and how global frontier ordering helps organize trajectories in larger scenes. Several limitations remain. The semantic value map still depends on the reliability of VLM-based scoring, and inaccurate semantic responses may affect frontier selection. The current framework also assumes reasonably reliable depth and pose information,

which may be challenging in noisier real-world deployments. Future work will investigate uncertainty-aware semantic mapping, more robust perception, and dynamic-scene reasoning to further improve the applicability of ForesightNav in complex real-world environments.

ACKNOWLEDGMENT

(Yu Liu and Chengjie Fang are co-first authors.)

REFERENCES

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3674–3683, doi: 10.1109/CVPR.2018.00387.

[2] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 4247–4258.

[3] H. Yin, X. Xu, Z. Wu, J. Zhou, and J. Lu, "SG-nav: Online 3D scene graph prompting for LLM-based zero-shot object navigation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 5285–5307.

[4] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "ESC: Exploration with soft commonsense constraints for zero-shot object navigation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 42829–42842.

[5] J. Krantz, S. Lee, J. Malik, D. Batra, and D. S. Chaplot, "Instance-specific image goal navigation: Training embodied agents to find object instances," 2022, *arXiv:2211.15876*.

[6] J. Krantz, T. Gervet, K. Yadav, A. Wang, C. Paxton, R. Mottaghi, D. Batra, J. Malik, S. Lee, and D. S. Chaplot, "Navigating to objects specified by images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10882–10891.

[7] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, S. Y. Min, K. Shah, C. Paxton, S. Gupta, D. Batra, R. Mottaghi, J. Malik, and D. S. Chaplot, "GOAT: GO to any thing," 2023, *arXiv:2311.06430*.

[8] X. Sun, L. Liu, H. Zhi, R. Qiu, and J. Liang, "Prioritized semantic learning for zero-shot instance navigation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 161–178, doi: 10.1007/978-3-031-73254-6_10.

[9] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[10] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[12] H. Yin, X. Xu, L. Zhao, Z. Wang, J. Zhou, and J. Lu, "UniGoal: Towards universal zero-shot goal-oriented navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 19057–19066.

[13] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 667–676, doi: 10.1109/3DV.2017.00081.

[14] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI," 2021, *arXiv:2109.08238*.

[15] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "RoboTHOR: An open simulation-to-real embodied AI platform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3161–3171, doi: 10.1109/CVPR42600.2020.00323.

[16] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. Van Den Hengel, "REVERIE: Remote embodied visual referring expression in real indoor environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9979–9988, doi: 10.1109/CVPR42600.2020.01000.

[17] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–11.

[18] A. B. Vasudevan, D. Dai, and L. Van Gool, "Talk2Nav: Long-range vision-and-language navigation with dual attention and spatial memory," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 246–266, Jan. 2021, doi: 10.1007/s11263-020-01374-3.

[19] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6622–6631, doi: 10.1109/CVPR.2019.00679.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "VLN BERT: A recurrent vision-and-language BERT for navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1643–1653, doi: 10.1109/CVPR46437.2021.00169.

[22] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 5834–5847.

[23] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16516–16526, doi: 10.1109/CVPR52688.2022.01604.

[24] W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, and H. Dong, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 5228–5234.

[25] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 3554–3560.

[26] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "CoWs on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23171–23181.

[27] Y. Kuang, H. Lin, and M. Jiang, "OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models," in *Proc. Findings Assoc. Comput. Linguistics, NAACL*, 2024, pp. 338–351, doi: 10.18653/v1/2024.findings-naacl.24.

[28] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "VLFM: Vision-language frontier maps for zero-shot semantic navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 42–48.

[29] H. Chang, K. Boyalakuntla, S. Lu, S. Cai, E. P. Jing, S. Keskar, S. Geng, A. Abbas, L. Zhou, K. Bekris, and A. Boularias, "Context-aware entity grounding with open-vocabulary 3D scene graphs," in *Proc. Conf. Robot Learn. (CoRL)*, 2023, pp. 1950–1974.

[30] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "ZSON: Zero-shot object-goal navigation using multimodal goal embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 32340–32352.

[31] K. Yadav, A. Majumdar, R. Ramrakhya, N. Yokoyama, A. Baevski, Z. Kira, O. Maksymets, and D. Batra, "OVRL-v2: A simple state-of-art baseline for ImageNav and ObjectNav," 2023, *arXiv:2303.07798*.

[32] A. Kamath et al., "Gemma 3 technical report," 2025, *arXiv:2503.19786*.

[33] X. Lei, M. Wang, W. Zhou, L. Li, and H. Li, "Instance-aware exploration-verification-exploitation for instance ImageGoal navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 16329–16339.

[34] L. Zhang, Q. Zhang, H. Wang, E. Xiao, Z. Jiang, H. Chen, and R. Xu, "TriHelper: Zero-shot object navigation with dynamic assistance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 10035–10042, doi: 10.1109/IROS58592.2024.10802670.

[35] X. Zhao, W. Cai, L. Tang, and T. Wang, "ImagineNav: Prompting vision-language models as embodied navigator through scene imagination," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 1–11. [Online]. Available: https://openreview.net/forum?id=vQFw9ryKyK

[36] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "InstructNav: Zero-shot system for generic instruction navigation in unexplored environment," in *Proc. Conf. Robot Learn. (CoRL)*, 2024, pp. 1–7.

[37] M. Habibpour and F. Afghah, "Think, remember, navigate: Zero-shot object-goal navigation with VLM-powered reasoning," in *Proc. NeurIPS Workshop SPACE Vis., Lang., Embodied AI (SpaVLE)*, 2025, pp. 1–11.
[38] D. Nie, X. Guo, Y. Duan, R. Zhang, and L. Chen, "WMNav: Integrating vision-language models into world models for object goal navigation," 2025, *arXiv:2503.02247*.
[39] M. Zhang, Y. Du, C. Wu, J. Zhou, Z. Qi, J. Ma, and B. Zhou, "ApexNAV: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion," *IEEE Robot. Autom. Lett.*, vol. 10, no. 11, pp. 11530–11537, Nov. 2025, doi: 10.1109/LRA.2025.3606388.
[40] H. Zhu, Z. Li, Z. Liu, W. Wang, J. Zhang, J. Francis, and J. Oh, "STRIVE: Structured representation integrating VLM reasoning for efficient object navigation," 2025, *arXiv:2505.06729*.
[41] S. Yuan, H. Huang, Y. Hao, C. Wen, A. Tzes, and Y. Fang, "GAMap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 39386–39408.



YU LIU received the B.S. degree from Southwest Petroleum University. She is currently pursuing the M.S. degree in information and communication engineering with Beijing Institute of Graphic Communication, Beijing, China. Her main research interests include artificial intelligence and computer vision.



CHENGJIE FANG is currently pursuing the graduate degree majoring in artificial intelligence with the University of Chinese Academy of Sciences, Beijing, China. His primary research interests include embodied intelligence and marine science.



LIKUN LU is currently an Associate Professor with the School of Information Engineering, Beijing Institute of Graphic Communication. He was the Director of the Laboratory for Digital Asset Evaluation of Publications under the National Technical Committee for Standardization of Press Distribution and the Deputy Director of the Key Technology Application Research and Service Laboratory for the Press and Publication Industry of the National Press and Publication Administration. He is also the Leader of Beijing Academic Innovation Team on Digital Printing and Inspection. His research interests include artificial intelligence, computer vision, and inkjet digital printing technology.