Think, Remember, Navigate: Zero-Shot Object-Goal Navigation with VLM-Powered Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

While Vision-Language Models (VLMs) are set to transform robotic navigation, existing methods often underutilize their reasoning capabilities. To unlock the full potential of VLMs in robotics, we shift their role from passive observers to active strategists in the navigation process. Our framework outsources high-level planning to a VLM, which leverages its contextual understanding to guide a frontier-based exploration agent. This intelligent guidance is achieved through a trio of techniques: structured chain-of-thought prompting that elicits logical, step-by-step reasoning; dynamic inclusion of the agent's recent action history to prevent getting stuck in loops; and a novel capability that enables the VLM to interpret top-down obstacle maps alongside first-person views, thereby enhancing spatial awareness. When tested on challenging benchmarks like HM3D, Gibson, and MP3D, this method produces exceptionally direct and logical trajectories, marking a substantial improvement in navigation efficiency over existing approaches and charting a path toward more capable embodied agents.

1 Introduction

A primary challenge in robotics is enabling autonomous navigation in unknown environments, a skill crucial for tasks like search and rescue or industrial inspection. Object Goal Navigation (ObjectNav), where an agent must locate a specific object in a new setting, is a particularly demanding version of this problem (3; 18). It requires a combination of advanced spatial awareness and semantic comprehension. Traditional approaches, which often depend on geometric maps and predetermined planning strategies, have difficulty generalizing to novel environments and thus fall short of achieving genuinely intelligent exploration.

The emergence of powerful Vision-Language Models (VLMs) has opened a new frontier, giving robots the potential to interpret their surroundings with a human-like understanding of context. Despite this promise, VLMs have typically been integrated into navigation systems in a limited capacity. Many current methods assign the VLM a passive function, such as scene description or query answering, rather than making it the primary strategist. This underuse is a result of superficial knowledge integration, inflexible prompting techniques, and a lack of memory, all of which lead to inefficient navigation and a restricted role for the VLM.

To overcome these shortcomings, we propose a novel framework that delegates high-level planning to a VLM, using its inherent contextual knowledge to direct navigation. Our approach reimagines the VLM's function, elevating it from a simple reactive component to the main navigator. By combining the VLM's emergent planning skills with frontier-based exploration, dynamic prompting, and multi-view fusion, our framework achieves robust and efficient navigation without relying on conventional planners. Our main contributions include:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- **Chain-of-Thought (CoT) for Navigation:** We employ CoT reasoning within the VLM, which allows it to produce more logical and context-aware instructions by methodically thinking through each step of the navigation task.
- **Dynamic Prompts with Action History:** Our system uses advanced prompts that include the agent's recent actions. This helps to avoid common issues like getting stuck in loops or indecisive movements, leading to more reliable exploration.
- **Top-Down Map Interpretation:** We enhance the VLM's reasoning by enabling it to analyze top-down obstacle maps in conjunction with its first-person view, giving it a better sense of the overall space for long-term planning.

2 Related Work

The challenge of Object Goal Navigation (ObjectNav) has been approached from several angles, beginning with end-to-end learning frameworks that often struggled to generalize beyond their training data (14; 15). Modular pipelines emerged as an alternative, deconstructing the problem into perception, mapping, and planning stages (3). While more robust, these systems could be brittle and prone to cascading errors between components (11). This has motivated a recent surge in zero-shot methodologies that harness the world knowledge of large pre-trained models, thereby avoiding the need for extensive task-specific fine-tuning (18).

The integration of Vision-Language Models (VLMs) first involved using them as powerful feature extractors, with methods like CLIP on Wheels (CoWs) (7) leveraging embeddings to link visual scenes with a target object's name. The role of language models soon became more active, with systems like ESC applying common-sense reasoning to guide exploration (29). More recently, the frontier of ObjectNav has shifted toward offloading high-level strategy entirely to large models. This paradigm includes diverse techniques such as translating maps into text for an LLM to score exploration paths (L3MVN (26)), employing a VLM to directly evaluate the semantic promise of frontiers (VLFM (25)), or even using a VLM to imagine and select optimal future viewpoints (ImagineNav (28)).

Our work advances this paradigm by enhancing the VLM's cognitive role in navigation. We introduce a framework that empowers the VLM with a more profound understanding of the task through a synergistic combination of three key techniques: structured Chain-of-Thought (CoT) prompting to elicit more logical, step-by-step analysis (22); the incorporation of a memory of recent actions to prevent stagnation; and a novel method for providing multimodal spatial context by enabling the VLM to interpret top-down obstacle maps in conjunction with its egocentric view. This holistic approach results in a more effective zero-shot navigator capable of generating more coherent and efficient trajectories.

3 Methodology

Our method's navigation is organized into three phases: initialization, exploration, and goal navigation. Our primary contribution is in the exploration phase, where we use a VLM to bring contextual intelligence to a standard frontier-based exploration framework. As shown in Figure 1, the VLM processes the agent's first-person view, a top-down obstacle map, and a specially designed prompt. It then serves as a high-level guide, suggesting the next move based on its comprehensive understanding of the scene and the goal. Algorithm 1 outlines the comprehensive procedure for the exploration stage, detailing how the VLM guidance, Value Map updates, and frontier prioritization are integrated to navigate the environment.

3.1 Waypoint Generation via Frontier Exploration

The agent is equipped with a 2D action space, including moving forward, turning, and stopping. We use the VER algorithm (23) for low-level motion control to guide the agent toward specific 2D waypoints. These waypoints are identified through a frontier-based exploration method (19), which analyzes depth data to create an obstacle map and find the boundaries between explored and unexplored areas. The midpoints of these frontiers are chosen as potential targets for exploration, promoting a thorough search of the environment.

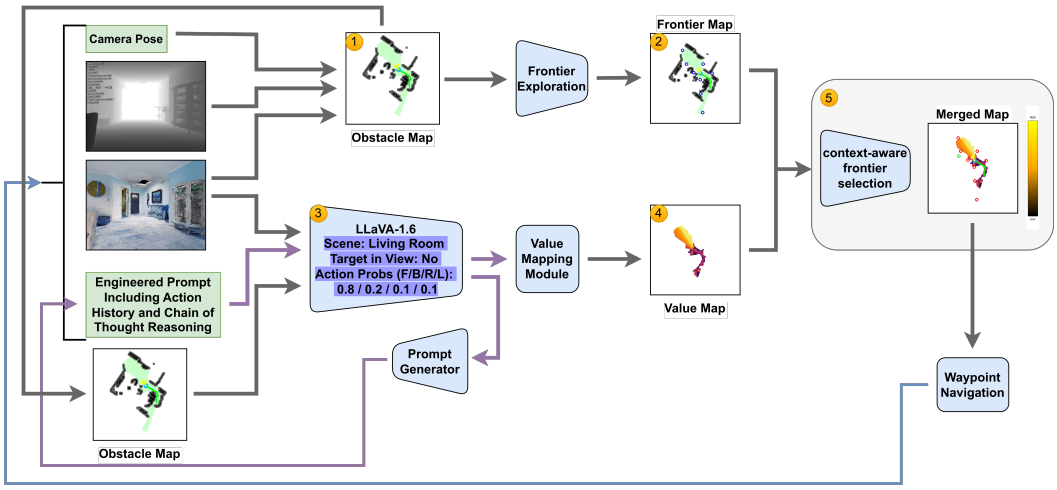
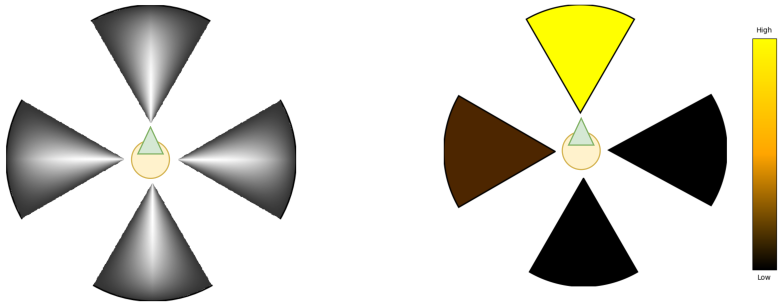


Figure 1: Our System Pipeline. (1) Sensor data is used to create an Obstacle Map. (2) Geometric frontiers are detected on this map. (3) The LLaVA-1.6 VLM analyzes the agent's egocentric view, the map, and a dynamic prompt that includes action history. (4) The VLM produces semantic scores, which are then used to build a Value Map that indicates the relevance of different areas. (5) The Frontier and Value Maps are combined to prioritize waypoints, directing the agent toward the most promising regions.



(a) (b)

Figure 2: Conceptual illustration of the value map components. (a) Robot's field of view (FOV) and the associated viewing uncertainty cone. (b) Example action space visualization with VLM-assigned scores: [Forward: 0.9, Backward: 0, Right: 0, Left: 0.1].

3.2 Value Map Creation using VLM Guidance

Since geometric frontiers do not have semantic meaning, we use the VLM to assign probability scores to actions like moving forward or turning left, based on the current view, action history, and our prompt. These scores are used to build a *value map* that reflects how promising different areas are. The scores are projected onto the local map and are adjusted by a *viewing uncertainty* factor, which accounts for the fact that the VLM's judgments may be less accurate for distant or peripheral regions, as conceptually depicted in Figure 2. The confidence level c for a point at a distance d and angle θ is calculated as:

$$c(d, \theta) = e^{-\lambda d} \cdot \cos^2\left(\frac{\theta}{\theta_{\text{fov}}/2} \cdot \frac{\pi}{2}\right) \tag{1}$$

Here, θ_{fov} is the camera's field of view, and λ is a parameter for distance decay. When the agent observes areas that overlap with previously seen regions, the semantic value $v_{i,j}^{\text{new}}$ and confidence $c_{i,j}^{\text{new}}$ for each pixel (i, j) are updated through a confidence-weighted average. This ensures that more

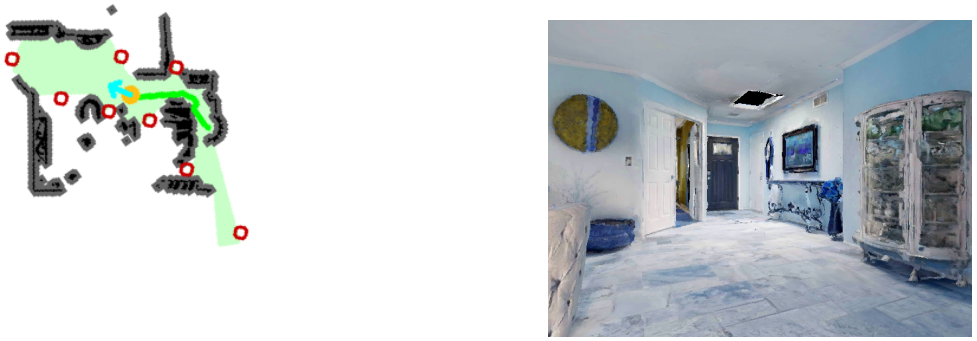
reliable observations have a stronger influence:

$$v_{i,j}^{\text{new}} = \frac{c_{i,j}^{\text{curr}} v_{i,j}^{\text{curr}} + c_{i,j}^{\text{prev}} v_{i,j}^{\text{prev}}}{c_{i,j}^{\text{curr}} + c_{i,j}^{\text{prev}}} \quad \text{and} \quad c_{i,j}^{\text{new}} = \frac{(c_{i,j}^{\text{curr}})^2 + (c_{i,j}^{\text{prev}})^2}{c_{i,j}^{\text{curr}} + c_{i,j}^{\text{prev}}} \tag{2}$$

The resulting value map is then combined with the geometric frontier map, enabling the system to prioritize exploration of the most semantically relevant areas.

3.3 Top-Down View Map Parsing by VLM

Drawing inspiration from human spatial reasoning, our framework equips the VLM with both a first-person egocentric view and a top-down obstacle map (Figure 3). Although VLMs are not typically trained on map data, we provide clear instructions in the prompt, such as: *"The second image is a top-down obstacle map..."*. This merging of first-person and overhead perspectives enhances the VLM's spatial awareness, leading to better-informed navigation choices, a finding supported by our ablation studies.



(a) Top-Down Obstacle Map (b) Egocentric View

Figure 3: Dual visual inputs for the VLM: (a) the top-down map shows the spatial layout with obstacles (in gray) and the agent's heading (arrow), while (b) the egocentric view provides a first-person perspective. This combination improves the VLM's spatial understanding.

Our choice of VLM was guided by a balance of reasoning capability and computational efficiency. After evaluating several open-source models, we selected LLaVA-1.6 (7B) (12) as it provided strong performance on multi-step reasoning tasks while maintaining an inference speed suitable for our navigation loop. During the final goal navigation phase, our system employs a robust hybrid strategy to mitigate potential failures in object detection. If an object is detected with high confidence by standard detectors (e.g., YOLOv7, Grounding-DINO), the system verifies the finding with the VLM and uses a lightweight segmentation model (Mobile-SAM) to delineate the object's boundaries, ensuring a more reliable final approach to the target.

4 Prompt Engineering

Successful interaction with the VLM depends heavily on the design of the prompt. We use several techniques to obtain structured and logical responses that are useful for navigation. CoT prompting improves a model's reasoning by guiding it through a series of intermediate steps to solve a complex problem (22). Instead of just asking for the best move, our prompt (see listing 1) leads the VLM through a logical sequence: (1) *Determine the room type.* (2) *Evaluate if the target object is likely to be in that room.* (3) *Confirm if the target is currently visible.* (4) *Suggest the most sensible action and assign scores.* This methodical approach, which was refined through extensive testing, ensures that the VLM's understanding of the scene is consistent with its recommended actions, reducing the chances of logical errors.

A common issue in navigation is when an agent gets stuck in a loop or becomes indecisive, especially in visually similar environments like long hallways. To address this, we keep a record of the agent's

last 10 actions. This history is included in the VLM prompt with instructions such as: *"Maintain forward progress... and avoid repetitive actions."* If the agent still gets stuck, a backup plan is activated, which temporarily repeats the last successful action to break the cycle. Our ablation study in Section 5.1 confirms that this memory-based strategy is essential for reliable long-term navigation.

5 Experimental Evaluation

We tested our framework in the Habitat simulator (9), using the HM3D, Gibson, and MP3D ObjectNav benchmarks (16; 24; 4). The main metrics for evaluation were Success weighted by Path Length (SPL) and Success Rate (SR) (1). We compared our approach with other top zero-shot methods. As Table 1 shows, our approach delivers exceptional efficiency, achieving the highest SPL on the demanding HM3D and MP3D datasets, and strong performance on Gibson. This indicates that our method's deep integration of VLM reasoning leads to more direct and intentional paths. This high efficiency is a direct result of our CoT and action history techniques, which prevent aimless exploration and keep the agent on a more logical course.

Table 1: Performance Comparison on ObjectNav Datasets. Our approach utilizes the LLaVA-1.6 (7B) model. SR and SPL are reported in percent. Top values are in bold.

Method	**HM3D**		**MP3D**		**Gibson**	
	SR	**SPL**	**SR**	**SPL**	**SR**	**SPL**
CoW (20)	-	6.3	3.7	7.4	-	-
ZSON (21)	25.5	12.6	15.3	4.8	-	-
ESC (29)	39.2	22.3	28.7	14.2	-	-
VLFM (25)	52.5	30.4	**36.4**	17.5	84.0	52.2
L3MVN (26)	50.4	23.1	-	-	76.1	37.7
GAMap (27)	53.1	26.0	-	-	**85.7**	**55.5**
Our Approach	**54.3**	**31.1**	36.0	**17.7**	80.2	53.0

5.1 Ablation Studies

To determine the impact of each part of our system, we conducted a series of ablation studies on a smaller set of benchmark episodes (Table 2). The findings show that every component adds value. Removing the top-down map resulted in a small drop in performance, while taking out the Chain-of-Thought reasoning caused a more significant decrease. Notably, the largest decline in performance happened when the **action history module** was removed, with the Success Rate on HM3D plummeting from 54.3% to 44.0%. This severe drop underscores that providing the agent with a short-term memory of its recent path is a necessity for robust, long-horizon navigation. Without this temporal context, the agent is prone to re-evaluating the same frontiers and becoming trapped in unproductive, oscillatory loops—a common failure mode observed in visually ambiguous environments.

Table 2: Results of the Ablation Study on a 50-episode subset. Each row indicates performance when the corresponding component is removed. The removal of action history has the most negative effect.

Ablation Condition	**HM3D**		**MP3D**		**Gibson**	
	SR (%)	**SPL (%)**	**SR (%)**	**SPL (%)**	**SR (%)**	**SPL (%)**
Full Framework	**54.3**	**31.1**	**36.0**	**17.7**	**80.2**	**53.0**
Without Chain-of-Thought	51.2	29.0	33.9	16.5	75.6	49.4
Without Action History	44.0	23.7	29.2	13.5	65.0	40.4
Without Obstacle Map	53.6	29.6	35.5	16.9	79.2	50.5

5.1.1 Effectiveness of Chain-of-Thought Prompting

To validate our structured reasoning approach, we tested prompts with varying levels of CoT complexity, based on the hypothesis that guiding the VLM through a logical sequence of questions would

yield more coherent and effective navigation decisions. We evaluated four configurations ranging in complexity: a baseline **No CoT** prompt that only requested action scores without any reasoning; a **Basic CoT** prompt asking for the single best action to find the target; an **Intermediate CoT** prompt that added a query for scene identification (e.g., "what room are you in?") before asking for an action; and our **Full NaviGen CoT**, a complete, multi-step prompt that requires the VLM to first identify the scene, then assess the likelihood of finding the target object there, and finally recommend an action based on this analysis.

The results, shown in Table 3, confirm a direct positive correlation between the depth of the CoT prompt and navigation performance. For instance, on HM3D, progressing from a non-CoT prompt to our full framework improved the SR from 51.2% to 54.3% and boosted the efficiency (SPL) from 29.0 to 31.1. This shows that compelling the VLM to "think step-by-step" is a functional requirement for translating visual input into successful, goal-oriented action.

Table 3: Ablation Study on Chain-of-Thought Prompt Complexity. Performance evaluated on a 50-episode subset.

Prompt Variation	**HM3D**		**MP3D**		**Gibson**	
	SR	**SPL**	**SR**	**SPL**	**SR**	**SPL**
No CoT (scores only)	51.2	29.0	32.5	15.1	75.4	49.5
Basic CoT ("Best action?")	52.0	29.5	33.1	15.9	77.0	50.3
Intermediate CoT (Scene ID)	53.3	30.2	34.8	16.8	78.9	51.7
Full CoT (multi-step, 1)	**54.3**	**31.1**	**36.0**	**17.7**	**80.2**	**53.0**

5.2 Qualitative Analysis and Limitations

To investigate the reasons for our framework's efficiency, we conducted a qualitative study of the VLM's decision-making. Figure 4 offers a direct comparison between an agent using our full, multi-step Chain-of-Thought (CoT) prompt and a basic agent with a simpler, non-CoT prompt. This comparison clearly demonstrates how structured thinking leads to better navigation.

The agent without CoT acts impulsively and with little foresight. Its reasoning is basic, connecting a visual stimulus directly to an action without considering the larger situation (e.g., "the TV is on the right... turn right"). This results in aimless wandering and an inability to complete the task on time. In contrast, the agent with our full CoT system follows a more logical and cohesive plan. It methodically analyzes the problem by first identifying the room type ("Bathroom"), then using its general knowledge to determine the probability of finding the target there ("Can a TV be Found Here?: No"), and finally making a sensible decision to search a more likely location ("Recommended Action: Go forward"). This step-by-step approach avoids pointless detours and creates direct and efficient paths. This structured thinking prevents logical inconsistencies in planning. Furthermore, the inclusion of the action history module provides crucial temporal context, allowing the agent to recognize and break out of unproductive loops, as visually demonstrated in Figure 5. This ability to avoid stagnation is critical for successful long-horizon navigation.

Our analysis also revealed limitations in the standard evaluation protocol, as we observed episodes marked as failures even when the agent found a valid object of the target category, simply because it was not the specific instance required by the ground truth. A detailed manual analysis of failed episodes provided further insights, categorizing them into five primary modes: annotation incompleteness (25%), premature episode termination due to loops (24%), in-view target oversight (19%), cross-level navigation deficits (18%), and semantic misclassification (14%). These findings highlight key areas for improvement in both dataset fidelity and the agent's long-horizon exploration and perception capabilities.

6 Conclusion and Future Work

We have presented a new framework for zero-shot object navigation that successfully combines the semantic reasoning of a VLM with methodical frontier-based exploration. Our main contributions—dynamic prompt engineering with CoT reasoning and action history—result in more logical

Figure 4: Qualitative analysis of navigation with and without our full CoT framework. The agent's view and the VLM's reasoning are shown at various timesteps. The **top row (No CoT)** displays an agent with basic reasoning that wanders aimlessly and fails to locate the target. The **bottom row (Full CoT)** illustrates how a structured, step-by-step reasoning process (such as identifying a bathroom, realizing a TV is not there, and choosing to leave) results in a more intelligent exploration strategy and a direct, successful path. This comparison underscores the crucial role of CoT in achieving more effective and intelligent navigation.

and reliable navigation, achieving state-of-the-art efficiency on widely used benchmarks. Despite these encouraging results, our approach has some limitations that point to areas for future research. The high computational demand of large VLMs is a barrier to real-time use, which calls for advancements in model compression and more efficient architectures. Additionally, our prompt engineering is currently a manual process; automating prompt optimization could improve performance. Future work could also improve the VLM's spatial reasoning by adding more detailed semantic information to the top-down map or by training VLMs specifically on map interpretation.

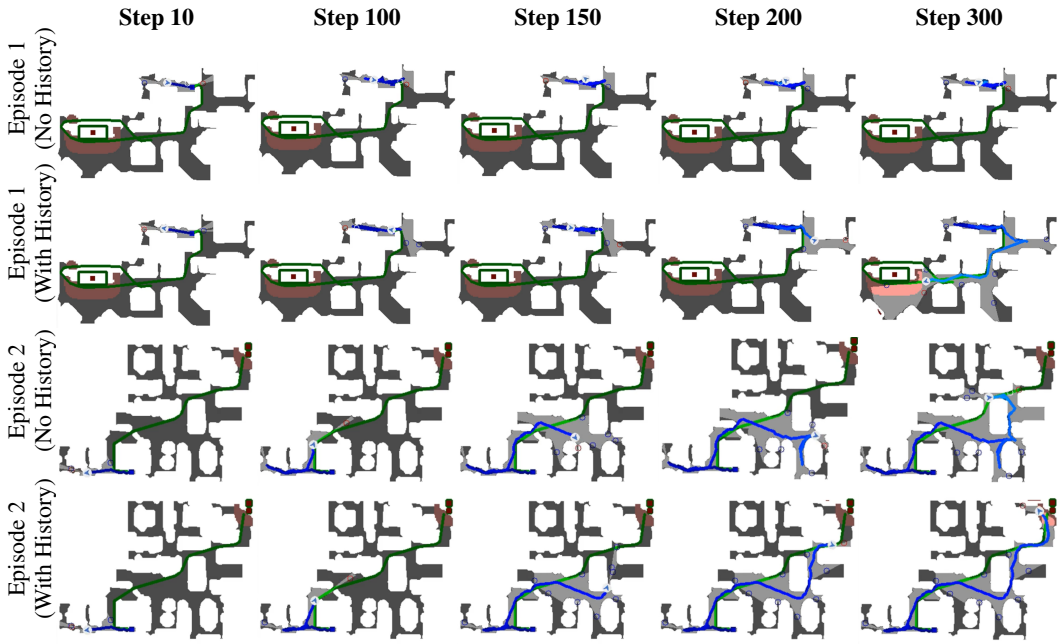


Figure 5: Qualitative comparison across two episodes, demonstrating the critical role of the action history module in resolving common navigation failures. **Episode 1 (Rows 1-2):** The top row shows an agent without action history getting trapped in a corridor. It falls into a persistent decision loop, oscillating between two frontier points and failing to make progress. The second row illustrates how the full framework, using its action history, recognizes the repetitive pattern, breaks the cycle, and successfully finds the target. **Episode 2 (Rows 3-4):** The third row depicts another failure where the agent without temporal memory gets stuck in indecisive movements, ultimately running out of time. In the final row, the action history module prevents this stagnation, enabling the agent to explore efficiently and reach the goal. Each column shows the agent's view at a specific simulation step.

References

[1] P. Anderson, et al. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

[2] W. Cai, et al. (2024). Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *ICRA*.

[3] D. S. Chaplot, et al. (2020). Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*.

[4] A. Chang, et al. (2017). Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*.

[5] V. S. Dorbala, J. F. Mullen, and D. Manocha. (2023). Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*.

[6] V. S. Dorbala, et al. (2022). Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649*.

[7] S. Y. Gadre, et al. (2022). CLIP on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*.

[8] N. Gireesh, et al. (2022). Object goal navigation using data regularized q-learning. In *CASE*.

[9] K. Yadav et al. (2023). Habitat Challenge 2023. https://aihabitat.org/challenge/2023/.

[10] A. Khandelwal, et al. (2022). Simple but Effective: CLIP Embeddings for Embodied AI. In *CVPR*.

[11] G. Kumar, et al. (2021). Gcexp: Goal-conditioned exploration for object goal navigation. In *RO-MAN*.

[12] H. Liu, C. Li, Y. Li, and Y. J. Lee. (2024). Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

[13] A. Majumdar, et al. (2022). Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *NeurIPS*.

[14] A. Mousavian, et al. (2019). Visual representations for semantic target driven navigation. In *ICRA*.

[15] X. Ye and Y. Yang, "Efficient robotic object search via hiem: Hierarchical policy learning with intrinsic-extrinsic modeling," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4425–4432, 2021.

[16] S. K. Ramakrishnan et al. (2021). Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

[17] B. Sun, et al. (2025). FrontierNet: Learning Visual Cues to Explore. *arXiv preprint arXiv:2501.04597*.

[18] J. Sun, J. Wu, Z. Ji, and Y.-K. Lai. (2024). A survey of object goal navigation. *IEEE Transactions on Automation Science and Engineering*.

[19] A. Topiwala, P. Inani, and A. Kathpal. (2018). Frontier Based Exploration for Autonomous Robot. *arXiv preprint arXiv:1806.03581*.

[20] S. Y. Gadre, K. Jiang, M. Wortsman, G. Ilharco, G. Gkioxari, D. Fried, L. Schmidt, A. Farhadi, and M. Rastegari, "CLIP on wheels: Zero-shot object navigation as object localization and exploration," in *Conference on Robot Learning (CoRL)*, 2022, pp. 70–80.

[21] A. Majumdar, Z. Al-Halah, and K. Grauman, "ZSON: A fast and efficient zero-shot open-set object navigator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22744–22753.

[22] J. Wei, et al. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

[23] E. Wijmans, I. Essa, and D. Batra. (2022). Ver: Scaling on-policy rl leads to the emergence of navigation in embodied rearrangement. In *NeurIPS*.

[24] F. Xia, et al. (2018). Gibson Env: Real-World Perception for Embodied Agents. In *CVPR*.

[25] N. Yokoyama, et al. (2024). Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *ICRA*.

[26] B. Yu, H. Kasaei, and M. Cao. (2023). L3mvn: Leveraging large language models for visual target navigation. In *IROS*.

[27] S. Yuan, et al. (2024). GAMap: Zero-Shot Object Goal Navigation with Multi-Scale Geometric-Affordance Guidance. *arXiv preprint arXiv:2410.23978*.

[28] X. Zhao, et al. (2024). ImagineNav: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination. *arXiv preprint arXiv:2410.09874*.

[29] K. Zhou, et al. (2023). Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *ICML*.

A Appendix

This appendix provides the pseudocode for the exploration phase of our approach and the full, optimized VLM prompt that resulted from our iterative refinement process.

Algorithm 1 Object Navigation Exploration Phase

Initialize map $\mathcal{M}_0$, Value map $\mathcal{V}_0(m) \leftarrow 0.5\ \forall m$, Confidence $C_0(m) \leftarrow 0\ \forall m$, pose p_0, action history $\mathcal{H}_a \leftarrow \emptyset$
function Explore($c_{\text{target}}, p_{\text{curr}}, \mathcal{M}_{\text{curr}}, \mathcal{V}_{\text{curr}}, C_{\text{curr}}, \mathcal{H}_{\text{curr}}$)
$p_t \leftarrow p_{\text{curr}}$; $\mathcal{M}_{t-1} \leftarrow \mathcal{M}_{\text{curr}}$; $\mathcal{V}_{t-1} \leftarrow \mathcal{V}_{\text{curr}}$; $C_{t-1} \leftarrow C_{\text{curr}}$; $\mathcal{H}_a \leftarrow \mathcal{H}_{\text{curr}}$
for $t = 0$ **to** MaxExplorationSteps **do**
  Get observation $O_t = (I_t, D_t)$ at pose p_t
  $\mathcal{M}_t \leftarrow \text{UpdateMap}(\mathcal{M}_{t-1}, D_t, p_t)$
  $\mathcal{F}_g \leftarrow \text{FindGeometricFrontiers}(\mathcal{M}_t, p_t)$
  $\pi_{\text{gen},t} \leftarrow \text{GeneratePrompt}(\pi_{\text{template}}, \mathcal{H}_a, c_{\text{target}})$
  $(S_{\text{act},t}, \beta_{\text{obj},t}, R_{\text{vlm},t}) \leftarrow \text{VLMQuery}(I_t, \mathcal{M}_t, \mathcal{H}_a, \pi_{\text{gen},t})$
  $\mathcal{V}_t \leftarrow \mathcal{V}_{t-1}$; $C_t \leftarrow C_{t-1}$
  for each map cell m in current FOV of O_t **do**
    $v_m^{\text{vlm}} \leftarrow \text{ProjectVLMScore}(S_{\text{act},t}, m, p_t, I_t)$
    $c_m^{\text{view}} \leftarrow \text{ViewingConfidence}(m, p_t, \text{FOV}_{\text{params}})$ {Eq. 1}
    $\mathcal{V}_t(m) \leftarrow \frac{c_m^{\text{view}} v_m^{\text{vlm}} + C_{t-1}(m)\mathcal{V}_{t-1}(m)}{c_m^{\text{view}} + C_{t-1}(m) + \epsilon}$ {Eq. 2}
    $C_t(m) \leftarrow \frac{(c_m^{\text{view}})^2 + (C_{t-1}(m))^2}{c_m^{\text{view}} + C_{t-1}(m) + \epsilon}$ {Eq. 2}
  end for
  $\mathcal{F}_p \leftarrow \emptyset$
  for each frontier $f \in \mathcal{F}_g$ **do**
    $v_f \leftarrow \text{QueryValueMap}(\mathcal{V}_t, C_t, f)$
    Add (f, v_f) to $\mathcal{F}_p$
  end for
  Sort $\mathcal{F}_p$ by v_f descending
  $w^* \leftarrow \text{SelectExplorationWaypoint}(\mathcal{F}_p)$
  $(p_{t+1}, a_t) \leftarrow \text{MoveTo}(w^*)$
  $\mathcal{H}_a \leftarrow \text{UpdateActionHistory}(\mathcal{H}_a, a_t)$
  $o_{\text{sensor}} \leftarrow \text{ObjectDetect}(I_t, c_{\text{target}})$
  if $\beta_{\text{obj},t}$ **and** o_{sensor} **then**
    NavigateToTarget(...) {Transition to goal phase}
    break
  end if
end for
end function

```
You are a robot navigating an indoor environment in search of a [
    TARGET_OBJECT].
Once you find it, move near the [TARGET_OBJECT] and stop.
The first image is your current observation and the second image is a
    top downview obstacle map of the environment. The grey areas are
    obstacles and The robots direction is visible with an arrow.
You must think step by step and ensure that all parts of your response
     are consistent.

Here are your recent actions: [ACTION_HISTORY]
Maintain forward progress and avoid getting stuck in a loop.

Here are the tasks:
1. Identify what part of the house we are about to enter (choose from:
     [bedroom, living room, kitchen, corridor, bathroom]).
2. Assess whether a [TARGET_OBJECT] can realistically be found in this
     area, based on common sense and the current observation.
3. Is there a [TARGET_OBJECT] in the current scene?
```

```
4. Determine the most logical next action for the robot (choose from:
    [go forward, go backward, turn right, turn left]).
- The chosen action must prioritize exploring areas likely to contain
    a [TARGET_OBJECT].
- Avoid suggesting actions that contradict previous observations (e.g
    ., don't explore a bathroom if couches aren't found there).
- If you are in a corridor, continue your path and Try to exit the
    corridor and describe where it leads.
5. Provide a probability score for each possible action. Each
    probability score should be a number between 0 and 1.

When providing your response, use this structure:
1. Part of the House: [Your answer]
- Reasoning: [Explain why you think this is the correct part of the
    house based on the observation and map.]
2. Can a [TARGET_OBJECT] Be Found Here?: [Yes/No]
- Reasoning: [Explain why or why not.]
3. Have You Found the [TARGET_OBJECT]?: [Yes/No]
4. Recommended Action: [Your action]
- Reasoning: [Explain why this action is the most logical based on
    steps 1, 2, and the action history.]
5. Probability Scores for Each Action:
- Go forward: [Score]
- Go backward: [Score]
- Turn right: [Score]
- Turn left: [Score]
```

Listing 1: The full, optimized CoT prompt for our approach.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that our framework improves navigation efficiency (SPL) by deeply integrating a VLM using CoT and action history. This is directly supported by the main results in Table 1 and the ablation studies in Table 2.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion section explicitly discusses limitations, including the computational cost of VLMs and the reliance on manual prompt engineering.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is empirical and does not present theoretical results or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the VLM used (LLaVA-1.6 7B), the datasets (HM3D, MP3D, Gibson), the simulation environment (Habitat), and the core components of the method, including the full prompt in the appendix. This provides a clear path for reproduction.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not include a link to the source code at this time. However, it relies entirely on publicly available datasets and models, and the methodology is described in sufficient detail to allow for independent reimplementation.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies that it uses the standard validation splits for the HM3D, MP3D, and Gibson datasets as defined by the Habitat Challenge. As a zero-shot method, no training is performed. Key algorithmic details like the decay parameter λ and the action history length are provided.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Error bars are not reported. The evaluation is performed over large, standard benchmark datasets containing hundreds to thousands of episodes, and the aggregate metrics (SR, SPL) are standard for this domain, providing a robust measure of performance.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The full paper specifies the VLM model (LLaVA-1.6 7B), and the hardware used (NVIDIA RTX A6000), which provides sufficient information to understand the computational requirements.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and models for the task of robotic navigation and does not involve human subjects or sensitive data.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A broader impact statement is included in the conclusion, discussing potential positive applications in assistive robotics and acknowledging potential dual-use concerns.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release new models or datasets. It uses an existing, publicly available VLM. Therefore, safeguards for release are not applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the creators of the datasets (HM3D, MP3D, Gibson), the simulation environment (Habitat), and the VLM (LLaVA) used. These are standard academic assets with permissive licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce any new assets like datasets or models.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: The Vision-Language Model (a type of LLM) is the central and core methodological component of this research. Its usage is described in detail in Section 3 and 4.