OpenVLM-Nav: Training-Free Zero-Shot Object-Goal Navigation via Vision-Language Guidance

Athira Krishnan R
Dept. of Artificial Intelligence
IIT Hyderabad
Telangana, India
ai22resch01001@iith.ac.in

Swapnil Bag
Dept. of Engineering Science
IIT Hyderabad
Telangana, India
es22btech11034@iith.ac.in

Sumohana S Channappayya
Dept. of Electrical Engineering
IIT Hyderabad
Telangana, India
sumohana@ee.iith.ac.in

***Abstract*—We propose OpenVLM-Nav, a training-free framework for zero-shot object-goal navigation using open-source vision–language models. Using CLIP [1], BLIP [2], and Qwen3-VL-2B [3], the agent interprets object descriptions directly from images without task-specific training. Qwen3-VL-2B performs the best, and we further study two extensions: a history module for temporal context and a depth module for geometric cues. Depth provides the most significant gain, improving the Success Rate (SR) from 0.08 to 0.14 and reducing the Distance-to-Goal (DTG) from 7.824 to 7.567. History gives minor, but consistent improvements. These results demonstrate that simple, training-free VLM-based navigation can be enhanced through the incorporation of temporal reasoning and depth information. Related codes will be made public.**

***Index Terms*—Zero-Shot Navigation, Object Goal Navigation, Embodied AI, Habitat.**

I. Introduction

Robotics, an interdisciplinary field, deals with autonomous agents capable of performing repetitive tasks with precision and accuracy. AI has played a crucial role in the wide acceptance of robots/autonomous agents. Embodied AI is a field where AI agents are trained to perceive their environment and make informed decisions. Although simulators such as Gazebo and Unity remain valuable, Habitat [4] is a preferred choice for embodied AI research.

When an agent successfully moves from point A to point B, we refer to it as navigation; similarly, when it does so by sensing its environment, it is referred to as embodied navigation. In object goal navigation, the agent attempts to navigate towards a specific object category. Zero-shot object goal navigation is an arena where models trained on a different downstream capability are utilized for object goal navigation tasks. Few papers [5], [6] have attempted zero-shot object goal navigation, leveraging models trained on embodied navigation, image goal navigation, and pixel goal navigation. In zero-shot navigation, when deploying a model trained on task A to another task B, VLMs are used to bridge the gap between the two. We primarily see researchers using ChatGPT [7], which is a paid solution [6], [8]. Even though they offer better qualitative results, the cost incurred in each iteration makes it less feasible. Training a new VLM from scratch requires compute, data, and time. Through this work, we aim to explore the capabilities of various open-source VLMs trained on open data in relation to ZSON. Thus, we introduce a training-free ZSON framework for embodied navigation, built on generalized open-source VLMs. We also attempt to examine the impact of history and depth on the decision-making capabilities of these VLMs, particularly in the context of embodied navigation.

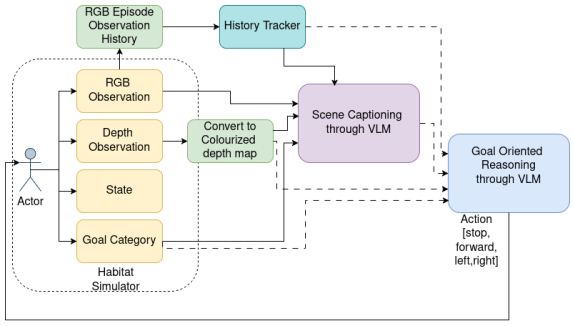

Fig. 1. The proposed methodology is depicted here. The Observations from the actor in the habitat simulator are preprocessed and passed to the Scene Captioning module and the Goal Reasoning module, along with the history and goal category as prompts. The directions returned by the agent are passed to the agent to act. State is used to compute the metrics.

II. Methodology

A. VLM guided Navigation

1) Prompt Engineering: We employ a two-stage prompting pipeline for zero-shot object navigation using a VLM. First, the model generates a structured spatial caption that describes objects, free space, and navigable cues (left/center/right). This caption, together with the raw image and goal category, is used in a second prompt that constrains the model to select a discrete action (FORWARD, LEFT, RIGHT, STOP). Additional prompt rules (directional priors and opening preference) stabilize navigation and improve exploration consistency.

```
prompt = (
        f"You are a navigation agent. Combine the RGB image and the depth map to describe the scene and where a {goal} might be. "
        "Warmer colors in the depth map indicate closer objects. Cooler colors indicate farther objects. "
        "Mention: whether the goal appears visible, which direction (forward/left/right) "
        "has more open space, and any nearby obstacles or doorways. "
        "Keep answer concise: 1-2 sentences, but include clear directional hints like 'open to the left' or 'obstacle ahead'."
    )

prompt = (
        f"You are a navigation controller trying to find and reach a {goal}.\n\n"
        f"{history_str}\n\n"
        f"Current scene description:\n{caption}\n\n"
        f"Your task: Navigate to the {goal}. Based on the scene caption and RGB image, choose ONE action: FORWARD, LEFT, RIGHT, or STOP.\n\n"
        "Rules:\n"
        f"- FORWARD: if there is a clear path or open space ahead that might lead to the {goal}\n"
        f"- LEFT/RIGHT: if there is a door, opening, or clear passage to the respective side that might have the {goal}\n"
        f"- STOP: ONLY if you can clearly see the {goal} directly in front and very close\n\n"
        "Reply with exactly ONE WORD: FORWARD, LEFT, RIGHT, or STOP."
    )
```

Fig. 2. Prompt templates for the VLM: (upper prompt is used to generate scene captions, while the lower prompt is used to predict navigation actions)

2) History and Depth Aware VLM guided Navigation: We incorporate a concise history of the recent actions and scene captions to provide the VLM with a limited form of temporal context. This design enables the model to reason not only over the current frame but also over how the agent has navigated in recent steps. Depth cues, encoded as a heat map, where warmer colours indicate nearby objects and cooler colours represent open spaces, provide the VLM with explicit distance information. This helps it better understand spatial layout, avoid obstacles, and perform safer navigation actions.

III. Results and Discussions

We conducted experiments to assess the applicability of open-source VLMs (without domain-specific knowledge) for zero-shot object-goal navigation. All experiments were performed using Habitat sim [4] on the MP3D [9] dataset.

Results from a few models trained on embodied navigation, which were extended to zero-shot object goal navigation, are taken. Pretrained VLMs, like CLIP, BLIP, and Qwen3-VL-2B, were selected for our experiments. Models were evaluated for 50 episodes on an RTX 3060 GPU. During CLIP-based textual reasoning, the Phi-3 LLM is used to support the process. Scene descriptions generated by the VLM are fed back into the model to predict the next actions. The prompts used in our experiments are shown in Figure 2.

For the best-performing candidate, we extended the inputs to include the depth map along with the history of scene descriptions and actions from the last N(=3) steps. Each of these inputs individually improved Success Rate (SR), Success weighted Path Length (SPL), and Distance to Goal (DTG). We therefore examined whether combining them would yield further performance gains.

A sample successful episode is depicted visually in Figure 3. All resultant metrics are also included in Table I. From the numbers, it is evident that Qwen3 achieves a competitive zero-shot baseline on ZSON. Across settings, adding depth and/or history improves navigation performance. Depth provides the highest performance gains in terms of SR, whereas adding history alone yields moderate improvements, and combining both maintains the high SR. Depth forces the VLM to make an informed trade-off between describing the scene and avoiding obstacles. The low SPL in the case of added depth and history modality is mostly due to confusion when making an informed choice with a lot of prior information.

TABLE I
ZERO-SHOT OBJECT NAVIGATION PERFORMANCE COMPARISON ON THE MP3D DATASET.

Model Type	Model	SR (↑)	SPL (↑)	DTG (↓)
Models trained on embodied navigation task	ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings [5]	0.1467	0.0637	4.68
	RGB PixNav with BLIP [10]	0.0526	0.0312	5.03
	Depth PixNav with BLIP [10]	0.0421	0.0059	4.84
	Ensemble (RGB + Depth) PixNav with BLIP [10]	0.0600	0.00275	7.03
Open-source VLMs (no domain knowledge)	CLIP + Phi-3	0	0	9.484
	BLIP	0	0	7.990
	Qwen3-VL-2B	0.08	0.061	7.824
Effect of depth and history on VLMs	Qwen3-VL-2B with history	0.10	**0.0867**	7.728
	Qwen3-VL-2B with depth	**0.14**	0.0755	**7.567**
	Qwen3-VL-2B with history + depth	0.14	0.0556	7.569

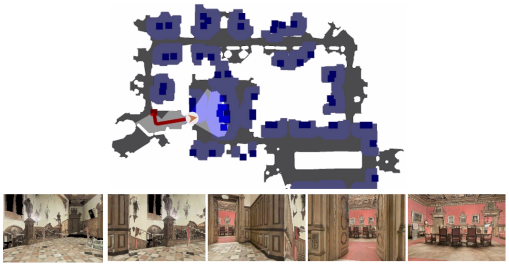

Fig. 3. Top down map view (top) and the observations seen by the agent (bottom) during a successful episode.

IV. Conclusions

Our experiments demonstrate that open-source VLMs trained without domain knowledge can serve as viable agents for zero-shot object goal navigation. With Qwen3 and depth, we got an inference time of ≈200s for an episode with around 80 steps. This shows that the approach is low-cost and lightweight, making it feasible for local inference and edge-computing deployments. Minor modifications to the prompts lead to substantial changes in performance. This highlights the importance of prompt design and suggests a clear direction for future work on prompt tuning.

REFERENCES

[1] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, pp. 12888–12900, PMLR, 2022.

[3] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[4] X. Puig, E. Undersander, A. Szot, M. D. Cote, R. Partsey, J. Yang, R. Desai, A. W. Clegg, M. Hlavac, T. Min, T. Gervet, V. Vondrus, V.-P. Berges, J. Turner, O. Maksymets, Z. Kira, M. Kalakrishnan, J. Malik, D. S. Chaplot, U. Jain, D. Batra, A. Rai, and R. Mottaghi, "Habitat 3.0: A co-habitat for humans, avatars and robots," 2023.

[5] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "Zson: Zero-shot object-goal navigation using multimodal goal embeddings," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[6] 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 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5228–5234, IEEE, 2024.

[7] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[8] X. Zhao, W. Cai, L. Tang, and T. Wang, "Imaginenav: Prompting vision-language models as embodied navigator through scene imagination," *arXiv preprint arXiv:2410.09874*, 2024.

[9] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[10] A. K. R and S. S. Channappayya, "Tepen: Towards an ensemble model for pixel-based embodied navigation," in *Proceedings of the 2025 International Conference on Pattern Recognition and Machine Intelligence (PREMI)*, Lecture Notes in Computer Science, Springer, 2025. In Press.