OpenFrontier: General Navigation with Visual-Language Grounded Frontiers

Boyang Sun[1], Cesar Cadena[1], Marc Pollefeys[1,2], Hermann Blum[1,3]

***Abstract*— Open-world navigation requires robots to make decisions in complex, dynamic environments and adapt to flexible task requirements. Traditional approaches often rely on hand-crafted goal metrics and struggle to generalize beyond specific tasks. Recent advances in vision-language-action (VLA) models enable end-to-end policies conditioned on natural language, but they typically require interactive training or large-scale data collection with a mobile agent. We frame navigation as a discrete sub-goal identification problem and extend our previous work, *FrontierNet*—a learning-based exploration system that detects and localizes frontiers directly from visual cues. We integrate FrontierNet with pre-trained vision-language models (VLMs) through a set-of-mark prompting strategy, enabling direct zero-shot, general-purpose navigation from natural language instructions. FrontierNet achieves state-of-the-art performance in autonomous exploration, and when combined with a VLM, demonstrates zero-shot adaptation across a variety of semantic tasks, such as object search—without requiring any additional training or map updating.**

I. Introduction

Navigation is a core capability for autonomous robots, critical for performing diverse tasks. Conventional approaches typically rely on modular pipelines that decompose the problem into perception, reasoning, and control. While such systems offer flexibility and interoperability, they often require significant manual effort for system design, integration, and hand-crafted task-specific metrics. Recent research has explored learning-based alternatives, including reinforcement learning (RL) from trial-and-error interaction [1]–[5], and large-scale vision-language-action models trained on curated navigation datasets [6]–[10]. These methods have demonstrated strong performance in in-distribution environments. However, their performance often degrades in open-world scenarios, which vary significantly from the training distribution. Moreover, adapting or retraining such models for new settings remains costly and time-consuming.

In this work, we propose a lightweight, reactive, and robust alternative for general-purpose navigation in open-world settings. Our system, *OpenFrontier*, is designed to be easy to deploy, flexible to various situations, and capable of accepting natural language task specifications—without requiring re-training. We observe that navigation can be framed as a sparse goal identification problem, which does not inherently require dense 3D scene representations or continuous action prediction. Building on this insight, we leverage both the structure of conventional modular systems and the generalization capabilities of pre-trained vision-language models (VLMs). At the core of OpenFrontier is *FrontierNet* [11], our previously proposed model that detects and registers 3D frontier candidates directly from posed RGB images, without relying on accurate dense 3D mapping. Unlike traditional exploration pipelines, FrontierNet uses only visual input to predict frontier pixels and estimate their potential information gain. This allows the system to capture rich visual cues that are useful for both spatial reasoning and semantic interpretation. To extend FrontierNet for general-purpose navigation, we introduce a novel mechanism for querying VLMs using the predicted frontiers. We employ a *set-of-marks* prompting strategy [12], [13], where the proposed frontiers act as spatial markers, and the VLM reweights them based on an open-vocabulary task prompt. This process effectively grounds semantic goals in the scene by identifying the most relevant frontiers to pursue. We validate our approach in both simulation and real-world settings. FrontierNet achieves state-of-the-art performance on autonomous exploration benchmarks. When combined with a VLM, OpenFrontier exhibits strong zero-shot generalization across diverse semantic navigation tasks, without any additional training. In summary, our contributions are:

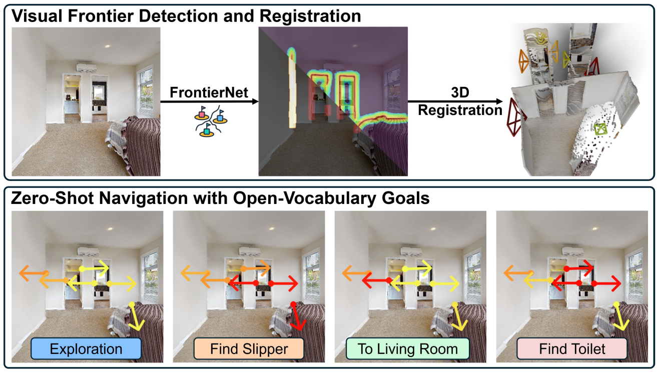

Visual Frontier Detection and Registration
FrontierNet
3D Registration
Zero-Shot Navigation with Open-Vocabulary Goals
Exploration
Find Slipper
To Living Room
Find Toilet

Fig. 1: **Top**: FrontierNet processes an RGB image (left) to propose frontier pixels and their information gain (middle), registering candidate goal viewpoints with varying priorities in 3D (right). **Bottom:** The proposed frontiers are used to query a vision-language model via set-of-mark prompting. Information gain (high-low) dynamically adapts to the open-vocabulary query, guiding the robot toward the target.

- **FrontierNet**: a learning-based model for efficient autonomous exploration that predicts frontier proposals and their associated information gain directly from visual input.
- **VLM-based extension** of FrontierNet that enables zero-shot general-purpose navigation by grounding open-vocabulary task prompts onto predicted frontiers.
- **Extensive evaluations** in simulation and real-world settings that validate both the exploration performance of FrontierNet and the zero-shot semantic navigation capabilities of its VLM-based extension.

[1]ETH Zurich, [2] Microsoft, [3] University of Bonn

Details about FrontierNet: https://boysun045.github.io/FrontierNet-Project/

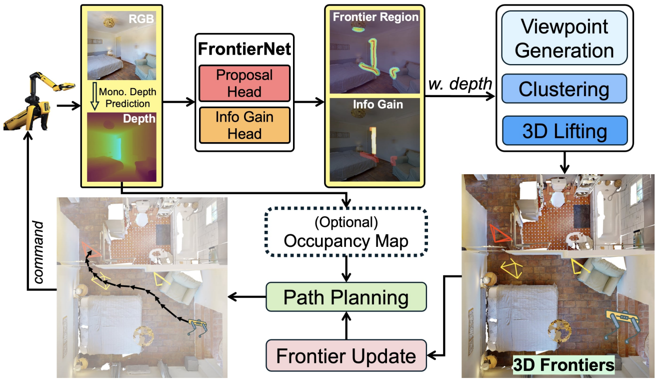


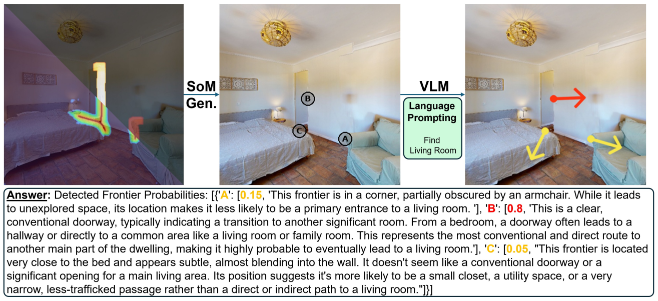
Fig. 2: **FrontierNet Overview.** Our system processes posed RGB images with a depth prediction model [14] to generate estimated depth. FrontierNet uses visual input to predict 2D frontier regions and their info gain, which are transformed into sparse 3D frontiers with different gains (colored frustums). These frontiers are tracked, and the planning module selects the next best goal and plans a path using the occupancy map.

II. Related Work

Autonomous exploration has been widely studied, with most methods falling into frontier-based or sampling-based approaches that operate on 3D representations such as occupancy grids [15], [16], signed distance fields [17], or point clouds [18]. Recent works incorporate learning-based scene understanding to guide viewpoint selection [19], [20], or explore emerging 3D representations like neural fields [21], [22] and 3D Gaussians [23]. Others introduce semantics via object-level maps [24], or frame exploration as a decision-making problem using reinforcement learning from RGB inputs [2]. More recent efforts leverage vision foundation models and large language models for interactive, human-like exploration [25]–[27]. These works show the value of visual cues for exploration, but typically rely on dense maps or auxiliary models. In contrast, we first demonstrate appearance cues alone can be used to directly detect and evaluate frontier regions without requiring full 3D reconstruction.

Navigation builds on exploration, shifting from uncovering space to reaching semantic targets such as objects [3], [28], places [29], or dynamic entities [30]. Recent approaches follow three main directions: 1) scene-centric methods [26], [31]–[33] that improve map building, representation, or interpretation; 2) reinforcement learning-based methods [5], [34], [35] that learn navigation via interaction; and 3) end-to-end policies [6]–[9] that map raw inputs to actions through imitation learning. All of these directions show promising progress, but also face limitations. They often rely on resource-intensive components, either in building and maintaining map representations, or in training large models. Moreover, they lack flexibility and are difficult to adapt to new setups without retraining policys or updating the map.

In this work, we treat navigation as a special case of exploration and use generic frontier goals to drive task-conditioned behavior. We introduce a lightweight framework that detects frontiers from individual images and maintains modularity with downstream planners and controllers. This allows us to avoid costly mapping or training, while enabling strong generalization and fast adaptation across tasks.



Answer: Detected Frontier Probabilities: [{'A': [0.15, 'This frontier is in a corner, partially obscured by an armchair. While it leads to unexplored space, its location makes it less likely to be a primary entrance to a living room. '], 'B': [0.8, 'This is a clear, conventional doorway, typically indicating a transition to another significant room. From a bedroom, a doorway often leads to a hallway or directly to a common area like a living room or family room. This represents the most conventional and direct route to another main part of the dwelling, making it highly probable to eventually lead to a living room.'], 'C': [0.05, "This frontier is located very close to the bed and appears subtle, almost blending into the wall. It doesn't seem like a conventional doorway or a significant opening for a main living area. Its position suggests it's more likely to be a small closet, a utility space, or a very narrow, less-trafficked passage rather than a direct or indirect path to a living room."]}]

Fig. 3: **VLM-based Extension** of FrontierNet. OpenFrontier takes the 2D frontier clusters produced by FrontierNet and queries a VLM using set-of-mark prompting. This re-weights the information gain of each frontier based on a given language instruction, grounding vision-language understanding into detected frontiers. In practice, we merge these relevance weights with the information gain predicted by FrontierNet, allowing the system to balance task-driven exploitation with exploration.

III. Method

A. System Overview

OpenFrontier consists of two main components: 1) FrontierNet, a learning-based model that detects frontier regions from visual input and registers them in 3D space. 2) VLM-based extension that leverages the frontiers proposed by FrontierNet to support natural language queries for general-purpose navigation. An overview of the FrontierNet pipeline is shown in Fig. 2. It performs joint frontier prediction and information gain estimation from posed RGB input, followed by 3D anchoring and goal selection via planning. During exploration, the system maintains an updating mechanism to track the status of frontiers over time. A path planner selects the next frontier goal and generates a trajectory accordingly. The VLM-based extension, illustrated in Fig. 3, shows how OpenFrontier integrates the frontier proposals from FrontierNet with a pre-trained VLM. This enables open-vocabulary goal specification and task-aware frontier re-weighting via a set-of-marks prompting mechanism. For full details, please refer to the FrontierNet [11] paper.

B. Learning to Propose Frontiers from Visual Appearance

Following Yamauchi's formulation [15], we define frontier as region of free space that directly borders unexplored space. Commonly, frontiers are therefore proposed from 3D voxel maps. Instead, we consider ***frontier pixels*** as the 2D projection of 3D frontier voxels within a camera's observed space and train a model that locates these pixels directly on image plane. Conventional frontier definitions treat every frontier as equally valuable. Recent studies [19], [20], [36] address this limitation by introducing quantitative metrics, often called information gain, that rank frontiers according to their expected exploratory benefit. We define the additional observable volume previously unknown from a frontier as its information gain (***info gain***) and train our model to also predict it from the visual input. This prediction depends only on individual images, assuming no prior exploration.

To unify the proposal of frontier pixels with the prediction of info gain, we employ a two-head UNet-like structure, and frame the task as an image-to-image prediction. It utilizes both the color image and its corresponding monocular depth prior as input and jointly predicts the frontier pixels and info gain. We design one prediction head that learns to estimate the distance of each image pixel to the nearest frontier boundary. Inspired by recent advances in structured edge and line detection [37], [38], our model outputs a soft distance field over the image, where lower values indicate proximity to likely frontiers. This allows the system to reason about spatial layout and boundary structure directly from image. In parallel, we predict an information gain map over the image on another head, representing how much new environment is likely to be revealed. Rather than regressing exact values, which can be noisy and sensitive to input variance [39], we discretize the gain into semantic bins and formulate the task as a classification problem. This improves stability and allows the model to prioritize meaningful frontiers.

C. Anchoring Frontier in 3D

To generate actionable navigation targets, we introduce an anchoring stage that lifts frontier predictions into 3D space as candidate viewpoints. Starting from pixel-wise frontier region and info gain maps predicted by FrontierNet, we identify sparse 2D frontier regions and infer the viewing direction of each pixel. We estimate their directions toward occluded or unknown regions using local geometry provided by monocular depth prior. We then cluster the frontier pixels based on spatial proximity, viewing direction, and estimated info gain. Each cluster yields a representative 2D frontier, comprising its pixel location, averaged viewing angle, and aggregated info gain. To lift these into 3D, we estimate depth at each frontier by sampling foreground and background depth values along the local gradient direction. This produces a set of sparse 3D frontiers, each associated with a 3D position, orientation, and task-relevant score. This anchoring procedure provides the planner with semantically meaningful and spatially grounded navigation targets. By relying only image input, it avoids the need for dense 3D reconstructions and remains robust to partial or noisy depth predictions.

D. VLM-based Information Gain Re-weighting

The original information gain predicted by FrontierNet measures how much unknown space each frontier could uncover—an effective metric for exploration. However, this geometric definition does not generalize to tasks like object search or language-conditioned navigation. Manually designing new metrics for each task is infeasible, and training end-to-end navigation models requires large-scale data collection or simulation. Instead, we use the predicted frontiers as spatial queries to a pre-trained vision-language model, enabling task-aware frontier re-weighting without retraining. As shown in Fig. 3, we cluster frontiers on each image and visualize them as markers overlaid on the RGB view. This guides the VLM's attention toward candidate regions. Using a set-of-marks prompting scheme, we query the VLM (Gemini-2.5) to assess the likelihood that reaching each frontier would accomplish a given natural language goal. The VLM leverages its pre-trained priors and surrounding visual context to return goal-conditioned beliefs over frontiers. We then merge the original info gain with these VLM-predicted scores, allowing the planner to balance exploration and exploitation. A stop condition can similarly be implemented by querying whether the goal has been achieved.

E. Exploration Planning

We design an exploration planner that selects the next navigation target based on a utility score balancing information gain and travel distance. For each predicted 3D frontier, its utility reflects the ratio between its estimated info gain and the effort required to reach it from the robot's current pose. This encourages the planner to prefer informative frontiers that are also efficient to reach.

During navigation, we maintain a tree-based structure that links robot poses and observed frontiers. Each node corresponds to either a past robot pose or a visible frontier. Whenever a frontier is registered, it is anchored to the pose from which it was observed, creating a local visibility edge. This structure ensures that all frontiers are physically reachable and have at least one valid line of sight from the robot's prior trajectory. To reach a selected frontier that lies beyond the current 3D map boundary, the planner samples intermediate waypoints along its visibility edge to find a reachable location within the known map. It then plans toward this intermediate point and progressively updates the map as it moves, eventually enabling a path to the original frontier goal. This planning strategy is especially robust in long-range navigation or under uncertainty from monocular depth, as it relies primarily on visibility and visual anchoring. It also supports operation in resource-limited scenarios by running in a completely map-free mode.

IV. Experiment And Result

A. Experimental Setup

We evaluate FrontierNet in 10 diverse indoor scenes from the HM3D dataset, covering a wide range of layouts, scales, and geometric complexity. Importantly, none of the evaluation scenes were used to train either FrontierNet or the monocular depth model. We simulate exploration using a virtual RGBD camera and construct occupancy maps using Octomap. Two types of depth input are considered: perfect rendered depth and predicted depth from Metric3D [14].

We benchmark against several exploration baselines, including the classic frontier-based method [15], the learning-based SEER [19], and the sampling-based NBVP [40], using official or re-implemented versions of each. We introduce **Vox@**k**(%)**—the fraction of scene volume explored after k% of the maximum allowed steps as the main metric. We report this metric at $k = 50$ and 100, averaged across multiple randomized runs per scene. Success rate is defined by achieving over 40% coverage before step limits are reached.

B. Results

For autonomous exploration, Table I summarizes the results across 10 validation scenes. FrontierNet achieves the best overall performance across all metrics. Notably, it attains the highest early-stage coverage (Vox@50) in all scenes, highlighting its ability to effectively prioritize informative regions during exploration. Complete result can be found in [11]. Figure 4 shows qualitative examples. We use the

		824	827	876	880	804	807	812	834	854	879	
		10/79/21	8/65/19	14/148/8	11/70/16	10/111/11	14/256/12	8/67/16	10/90/13	6/72/5.0	15/126/28	Mean
Vox@50	Classic	29.1±4.8	37.6±8.0	31.9±7.2	26.1±7.2	39.4±0.0	24.2±0.0	27.7±6.7	37.5±5.6	43.1±4.4	39.2±5.6	33.6
	NBVP	46.2±5.7	46.1±5.9	44.5±5.1	31.0±1.3	46.6±4.6	35.3±2.7	49.4±6.6	44.1±3.0	52.3±2.6	45.5±4.8	44.1
	SEER	47.0±4.4	46.6±6.2	30.4±9.9	57.1±2.2	40.4±7.7	32.2±6.0	41.0±5.5	22.8±5.2	43.5±3.1	44.8±3.9	40.6
	Ours	**58.0±4.8**	**61.9±3.9**	**58.2±4.2**	**61.9±7.5**	53.9±**4.2**	**50.7±4.5**	60.3±**8.1**	53.7±**5.0**	**72.1±9.8**	55.5±**5.7**	**58.6**
Vox@100	Classic	47.6±1.6	61.2±8.6	45.0±8.2	61.3±5.2	53.7±0.0	45.2±0.0	68.6±10.9	48.3±5.0	54.1±3.7	50.5±5.4	53.6
	NBVP	65.0±5.6	**78.5±4.9**	60.8±9.3	49.8±1.6	**69.7±4.8**	49.9±2.1	**83.4±3.5**	70.0±8.8	80.1±20.3	**62.6±5.6**	67.0
	SEER	60.6±6.7	60.1±5.6	50.5±8.8	60.3±6.1	62.3±3.2	51.7±5.6	60.8±8.3	45.1±4.9	51.0±3.4	48.1±3.0	55.1
	Ours	71.2±**6.0**	72.6±8.9	72.0±**8.5**	68.4±**10.8**	62.2±8.9	**59.8±6.1**	82.2±10.1	**70.3±10.1**	**98.3±13.2**	58.8±6.5	**71.5**
Suc.	Classic	33.3	86.7	38.0	40.0	6.3	5.6	37.5	31.3	90.0	20.0	38.9
	NBVP	**100.0**	**100.0**	**90.0**	50.0	**100.0**	65.0	**100.0**	**100.0**	60.0	**100.0**	86.5
	SEER	88.9	55.6	61.1	66.7	55.6	33.3	55.6	33.3	80.0	77.8	60.8
	Ours	**100.0**	81.3	83.3	**100.0**	80.0	**80.0**	**100.0**	86.7	**100.0**	75.0	**88.6**

TABLE I: Comparison of mapping efficiency (Vox@k%) and success rate (Suc.) with baseline methods. All methods use ground-truth depth from the simulator. SEER is our re-implementation of the original frontier-proposal technique, paired with our planner and evaluated under identical test conditions.. The 3-digit numbers in the first row are scene IDs. The three parameters below each scene ID are the retrieved relevant scene parameters from HM3D metadata (num_rooms, navigable_area, and navigation_complexity)

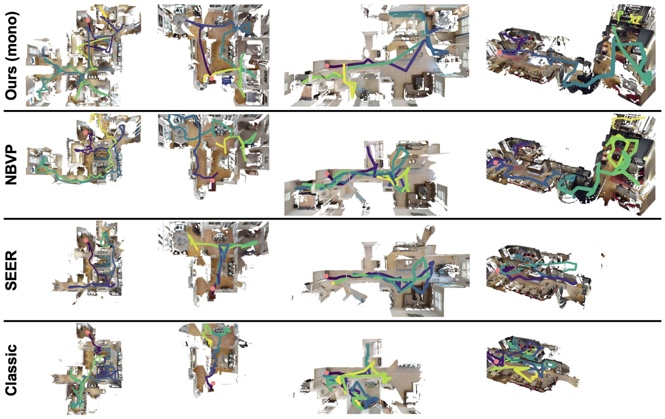


Fig. 4: **Qualitative Exploration Comparison** of FrontierNet compared to three baseline methods across four different scenes (left to right: 876, 824. 880, 854). Starting location is marked as red point. Notably, our approach successfully handles multi-floor environments (scene 854), a challenge for traditional frontier-based methods. All 3D meshes in this visualization are generated by TSDF integration using ground-truth depth images just for fair and clearer comparison.

same path planner, frontier assignment, and update logic for our method, our implementation of the Classic method, and SEER. FrontierNet's superior results therefore arise solely from its own strengths: it detects frontiers more reliably and estimates information gain more accurately. These improvements show that leveraging the visual cues leads to more effective exploration.

For open-vocabulary navigation, Figure 5 illustrates trajectories generated by OpenFrontier. The robot starts from the same initial pose and is given different navigation targets in natural language. Without any retraining or fine-tuning, our system achieves strong zero-shot performance. In all six test cases, the agent successfully navigates to the correct semantic goal, showcasing the framework's generality, task flexibility, and ease of adaptation across tasks—all enabled by integrating visual frontiers with a vision-language model. A video of the exploration process is available at *YouTube*.

C. Real-world Validation

We deploy our system on a Boston Dynamics Spot robot with a front RGB camera (640×480 @ 3 Hz). Running in real time on a laptop (i9, 3080Ti), FrontierNet achieves ∼5 Hz inference. As shown in Fig. 6, the robot explores a large indoor space without human input. Despite training only in simulation, the system generalizes well to real-world scenes.

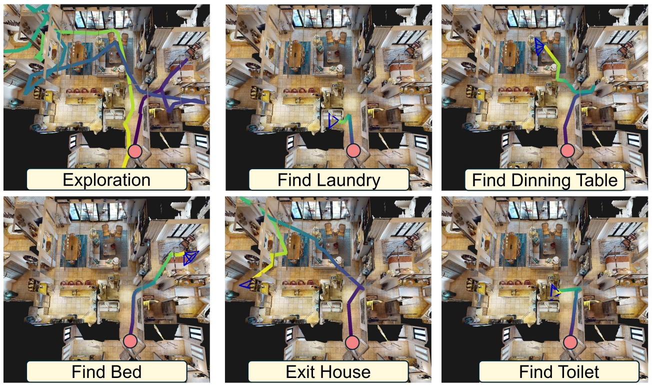


Fig. 5: **Qualitative Navigation Result** of our OpenFrontier in scene 876. Starting locations are all the same and marked as red points.

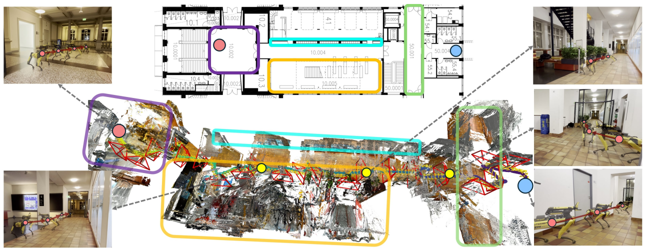


Fig. 6: **Real-world Validation Result.** Exploration process of a quadrupedal robot in a real-world environment. Top: Floor plan. Bottom: Reconstructed map and exploration path from TSDF integration using monocular depth prediction. Colored boxes indicate key correspondences between the map and floor plan.

V. Conclusion

In this work, we present a general-purpose navigation system that supports natural language instructions through a novel integration of frontier-based spatial reasoning and vision-language models. We treat navigation as a special case of exploration and use frontiers as grounding targets to query a VLM for task-aware prioritization via set-of-marks prompting. Our proposed model, FrontierNet, detects and registers frontiers directly from visual input without relying on dense mapping. Built on top of it, *OpenFrontier* demonstrates strong zero-shot adaptability across a range of navigation tasks, with minimal assumptions about the environment or training data. Experimental results highlight the effectiveness of each component, and ongoing work aims to further improve the system's reactivity and robustness in real-world scenarios.

References

[1] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023.

[2] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov, "Active Neural Localization," *ICLR*, 2018.

[3] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *In Neural Information Processing Systems*, 2020.

[4] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," in *ICLR*, 2020.

[5] K.-H. Zeng, Z. Zhang, K. Ehsani, R. Hendrix, J. Salvador, A. Herrasti, R. Girshick, A. Kembhavi, and L. Weihs, "Poliformer: Scaling on-policy rl with transformers results in masterful navigators," *arXiv*, 2024.

[6] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A foundation model for visual navigation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2306.14846

[7] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration," *arXiv pre-print*, 2023. [Online]. Available: https://arxiv.org/abs/2310.07896

[8] A.-C. Cheng, Y. Ji, Z. Yang, X. Zou, J. Kautz, E. Biyik, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," in *RSS*, 2025.

[9] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *Robotics: Science and Systems*, 2024.

[10] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," *arXiv preprint arXiv:2412.06224*, 2024.

[11] B. Sun, H. Chen, S. Leutenegger, C. Cadena, M. Pollefeys, and H. Blum, "Frontiernet: Learning visual cues to explore," *IEEE Robotics and Automation Letters*, vol. 10, no. 7, pp. 6576–6583, 2025.

[12] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.

[13] R. Shah, A. Yu, Y. Zhu, Y. Zhu, and R. Martín-Martín, "Bumble: Unifying reasoning and acting with vision-language models for building-wide mobile manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 13 337–13 345.

[14] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *International Symposium on Computational Intelligence in Robotics and Automation CIRA*, 1997.

[16] M. Selin, M. Tiger, D. Duberg, F. Heintz, and P. Jensfelt, "Efficient autonomous exploration planning of large-scale 3-d environments," *Robotics and Automation Letters*, 2019.

[17] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon path planning for 3d exploration and surface inspection," *Autonomous Robots*, vol. 42, pp. 291–306, 2018.

[18] C. Cao, H. Zhu, H. Choset, and J. Zhang, "Tare: A hierarchical framework for efficiently exploring complex 3d environments." in *Robotics: Science and Systems*, vol. 5, 2021.

[19] Y. Tao, Y. Wu, B. Li, F. Cladera, A. Zhou, D. Thakur, and V. Kumar, "Seer: Safe efficient exploration for aerial robots using learning to predict information gain," in *ICRA*, 2023.

[20] L. Schmid, M. N. Cheema, V. Reijgwart, R. Siegwart, F. Tombari, and C. Cadena, "Sc-explorer: Incremental 3d scene completion for safe and efficient exploration mapping and planning," *arXiv preprint arXiv:2208.08307*, 2022.

[21] Z. Yan, H. Yang, and H. Zha, "Active neural mapping," in *ICCV*, 2023.

[22] S. Lee, C. Le, W. Jiahao, A. Liniger, S. Kumar, and F. Yu, "Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields," *IEEE Robotics and Automation Letters*, 2022.

[23] W. Jiang, B. Lei, and K. Daniilidis, "Fisherrf: Active view selection and mapping with radiance fields using fisher information," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–440.

[24] S. Papatheodorou, N. Funk, D. Tzoumanikas, C. Choi, B. Xu, and S. Leutenegger, "Finding things in the unknown: Semantic object-centric exploration with an mav," in *ICRA*, 2023.

[25] J. Chen, G. Li, S. Kumar, B. Ghanem, and F. Yu, "How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers," in *Proceedings of Robotics: Science and Systems(RSS)*, 2023.

[26] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *International Conference on Robotics and Automation (ICRA)*, 2024.

[27] K. Qu, J. Tan, T. Zhang, F. Xia, C. Cadena, and M. Hutter, "Ippon: Common sense guided informative path planning for object goal navigation," *arXiv preprint arXiv:2410.19697*, 2024.

[28] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[29] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, S. Y. Min, K. Shah, C. Paxton, S. Gupta, D. Batra *et al.*, "Goat: Go to any thing," *arXiv preprint arXiv:2311.06430*, 2023.

[30] C. Scheidemann, L. Werner, V. Reijgwart, A. Cramariuc, J. Chomarat, J.-R. Chiu, R. Siegwart, and M. Hutter, "Obstacle-avoidant leader following with a quadruped robot," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1407–1413.

[31] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[32] W. Gao, B. Ai, J. Loo, D. Hsu *et al.*, "Intentionnet: Map-lite visual navigation at the kilometre scale," *arXiv preprint arXiv:2407.03122*, 2024.

[33] J. Loo, Z. Wu, and D. Hsu, "Open scene graphs for open-world object-goal navigation," *arXiv preprint arXiv:2508.04678*, 2025.

[34] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1gX8C4YPr

[35] W. Xie, H. Jiang, Y. Zhu, J. Qian, and J. Xie, "Naviformer: A spatio-temporal context-aware transformer for object navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 708–14 716.

[36] A. Dai, S. Papatheodorou, N. Funk, D. Tzoumanikas, and S. Leutenegger, "Fast frontier-based information-driven autonomous exploration with an mav," in *ICRA*, 2020.

[37] R. Pautrat, D. Barath, V. Larsson, M. R. Oswald, and M. Pollefeys, "Deeplsd: Line segment detection and refinement with deep image gradients," in *CVPR*, 2023.

[38] N. Xue, T. Wu, S. Bai, F. Wang, G.-S. Xia, L. Zhang, and P. H. Torr, "Holistically-attracted wireframe parsing," in *CVPR*, 2020.

[39] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.

[40] L. Schmid, M. Pantic, R. Khanna, L. Ott, R. Siegwart, and J. Nieto, "An efficient sampling-based method for online informative path planning in unknown environments," *IEEE Robotics and Automation Letters*, 2020.