VL-TGS: Trajectory Generation and Selection using Vision Language Models in Mapless Outdoor Environments[†]

Daeun Song[1⋆], Jing Liang[2⋆], Xuesu Xiao[1], and Dinesh Manocha[2]

***Abstract*— We present a multi-modal trajectory generation and selection algorithm for real-world mapless outdoor navigation in human-centered environments. Such environments contain rich features like crosswalks, grass, and curbs, which are easily interpretable by humans, but not by mobile robots. We aim to compute suitable trajectories that (1) satisfy the environment-specific traversability constraints and (2) generate human-like paths while navigating on crosswalks, sidewalks, etc. Our formulation uses a Conditional Variational Autoencoder (CVAE) generative model enhanced with traversability constraints to generate multiple candidate trajectories for global navigation. We develop a visual prompting approach and leverage the Visual Language Model's (VLM) zero-shot ability of semantic understanding and logical reasoning to choose the best trajectory given the contextual information about the task. We evaluate our method in various outdoor scenes with wheeled robots and compare the performance with other global navigation algorithms. In practice, we observe an average improvement of 20.81% in satisfying traversability constraints and 28.51% in terms of human-like navigation in four different outdoor navigation scenarios.**

I. Introduction

Social robot navigation in outdoor environments requires a good understanding of the environment to adapt to social norms [2], such as crossing at zebra crossings and staying on pedestrian walkways. Building a large-scale map for such navigation is impractical because outdoor environments are highly dynamic, with frequent changes due to construction, road closures, and shifting pedestrian flow [3], [4]. Thus, a mapless approach enables robots to navigate directly using sensory input [5], [6], allowing them to adapt to environmental changes and social dynamics in real-time without relying on a pre-built map.

Robots must not only recognize physical features, such as walkways, crosswalks, and paved paths, but also interpret their intended use within the environment and navigate accordingly. For example, paved roadways may only be temporarily used when construction blocks the sidewalk, but they can always be used to cross a street when marked with a zebra crossing. This involves identifying areas designated for pedestrian movement, detecting obstacles or temporary changes, and understanding how these elements influence viable paths. Achieving this requires contextual reasoning to understand and adapt to the implicit rules and expectations of human-centered environments [7].

To build such contextual understanding of the environment, many existing methods [8], [9] rely on segmentation or classification [10], [11]. However, these require extensive training with ground truth data and are limited to labeled datasets. This limitation hinders their generalizability to unknown scenes. Recent advances in Large Language Models (LLMs) and Vision Language Models (VLMs) have demonstrated strong zero-shot capabilities across a wide range of tasks, including logical reasoning [12], [13] and visual understanding [14], [15]. VLMs, in particular, have the ability to process and understand both visual and textual information, enabling them to perform a wide range of multi-modal tasks including making decisions for outdoor navigation.

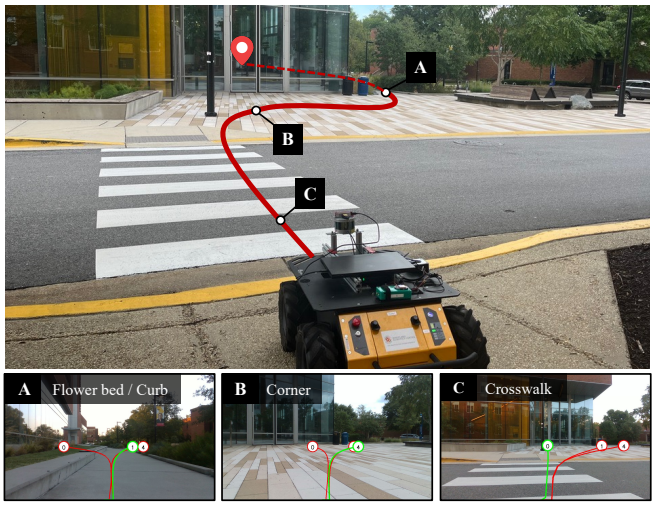

Fig. 1. Trajectories generated and selected using VL-TGS. The example path includes three different types of scenarios: (A) flower bed and curb, (B) corner, and (C) crosswalk. On the top, the map pin icon marks the goal behind the building, with the red solid or dashed line highlighting the robot's path. On the bottom, candidate trajectories are marked in red lines with numbers. The green path corresponds to the final trajectory.

Main Results: We present VL-TGS, a novel multi-modal approach for trajectory generation and selection in mapless outdoor navigation (Fig. 1). Our method combines LiDAR-based geometric information with RGB image data for comprehensive traversability analysis and scene understanding. Using a CVAE-based approach, we first generate multiple candidate trajectories based on the LiDAR scene perception. A VLM is then employed for trajectory selection based on the environmental context understanding through RGB image data. While VLMs lack the capability to produce precise spatial outputs, they can effectively utilize visual annotations to guide the selection process among a discrete set of coarse options [16]–[18]. By incorporating VLMs, our approach enables human-like decision-making to select

⋆Equal contribution.

[†]Extended abstract of the original work published in RA-L [1].

[1]George Mason University. [2]University of Maryland, College Park.

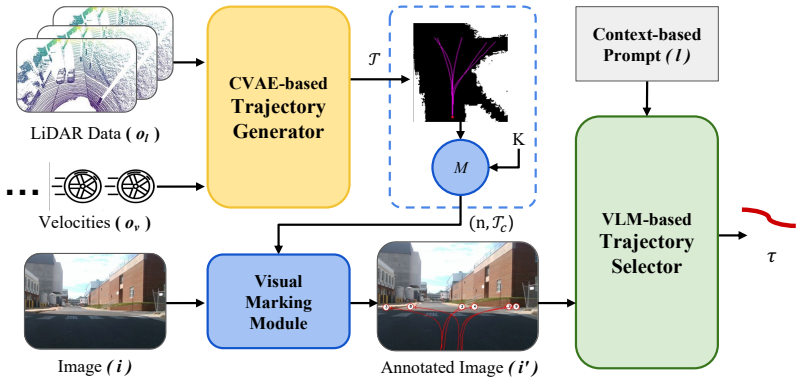

Fig. 2. **Architecture:** Our approach consists of two stages: CVAE-based trajectory generation and VLM-based trajectory selection. In the first stage, our attention-based CVAE takes consecutive frames of LiDAR point clouds and robot velocities as input, generating multiple diverse trajectories. These trajectories are sorted and visually marked with lines and numbers in the robot-view RGB image. In the second stage, our VLM-based trajectory selection module identifies the best trajectory number based on semantic feasibility, ensuring it lies on the sidewalk, avoids structures, crosses at zebra crossings, and adheres to other contextual rules.

optimal trajectories from the candidates, ensuring they align with geometric traversability constraints while addressing the contextual demands of global navigation.

II. Approach

A. Overview

Our approach computes a trajectory in a mapless environment for global navigation. Mapless global navigation requires a robot to reach a distant target beyond its immediate surroundings without relying on a pre-built map. To achieve this, we utilize multi-modal sensor data, combining both geometric and RGB visual information, to iteratively generate local trajectories that guide the robot towards the goal. Our approach follows a two-stage pipeline, as illustrated in Fig. 2. In the first stage, we generate multiple candidate trajectories, each spanning a fixed length (e.g., 10m) that satisfy the geometric traversability constraints. Then, we select the best trajectory based on human-like decision-making. Given a target goal $g \in \mathcal{O}_g$, we use a GPS sensor to provide the relative position between the target and the current location. Our goal is to compute a trajectory, τ, that aims to provide the best path to the goal, and that satisfies the traversability constraints of the scenario, $\tau = \text{VL-TGS}(\ell, \mathbf{i}, \mathbf{o}, g)$, where $\mathbf{o} = \{\mathbf{o}_l, \mathbf{o}_v, \mathbf{i}\}$ represents the robot's observations. $\mathbf{o}_l \in \mathcal{O}_l$ represents LiDAR observations, $\mathbf{o}_v \in \mathcal{O}_v$ indicates the robot's velocity, and $\mathbf{i} \in \mathcal{I}$ represents the RGB images from the camera. $\ell \in \mathcal{L}$ represents the language instructions to the Vision-Language Models (VLMs) for acquiring traversable trajectories.

We use Conditional Variational Autoencoder (CVAE) [5] to process the geometric information, $\mathbf{o}_l \in \mathcal{O}_l$, from the LiDAR sensor and the consecutive velocities, $\mathbf{o}_v \in \mathcal{O}_v$, from the robot's odometer. We efficiently generate a set of trajectories lying in geometrically traversable areas, $\mathcal{T} = \text{CVAE}(\mathbf{o}_l, \mathbf{o}_v)$. These generated trajectories cannot handle geometrically similar but color-semantically different situations, such as crosswalks as shown in Fig. 1 (C). We use VLMs to provide scene understanding from the RGB images.

However, the generated real-world waypoints from CVAE and the image observations are in two different modalities. To fuse these, we overlay the trajectories onto the images. VLMs are then used to assess whether the trajectories align with the contextual constraints of the environment. We assume that VLMs can infer common-sense reasoning from the images. We place these numbers at the end of each trajectory, starting from 0. The numbers indicate the order of distances to the goal, with the lowest number corresponding to the trajectory with the shortest distance. Thus, we map the real-world trajectories to image pixel-level objects by

$$(\mathbf{n}, \mathcal{T}_c) = M(\text{CVAE}(\mathbf{o}_l, \mathbf{o}_v), K), \tag{1}$$

where K denotes the conversion matrix from the real-world LiDAR frame to the image plane, $\mathcal{T}_c$ denotes the converted trajectories, and $\mathbf{n} \in \mathcal{N}$ are the numbers corresponding to each trajectory.

Given the language instruction ℓ, the image $\mathbf{i}$ with the converted trajectories $\mathcal{T}_c$, and numbers $\mathbf{n} \in \mathcal{N}$, our VLM selects one traversable trajectory based on the color-semantic understanding of the scenarios:

$$\tau = \text{VLM}(\ell, \mathbf{i}, \mathcal{T}_c, \mathbf{n}). \tag{2}$$

We choose the trajectory with the highest probability as the human-like trajectories, $\max P(\tau|\ell, \mathbf{i}, \mathcal{T}_c, \mathbf{n})$. Therefore, the problem is defined as:

$$\max \ P(\tau|\ell, \mathbf{i}, \mathcal{T}_c, \mathbf{n}). \tag{3}$$

B. Geometry-based Trajectory Generation

The trajectory set, $\mathcal{T}$, is generated by a CVAE to generate trajectories with associated confidences. For each observation $\{\mathbf{o}_l, \mathbf{o}_v\}$, we calculate the condition value $\mathbf{c} = f_e(\mathbf{o}_l, \mathbf{o}_v)$ for the CVAE decoder, where $f_e(\cdot)$ denotes the perception encoder. The embedding vector is then calculated from $\mathbf{c}$ as $\mathbf{z} = f_z(\mathbf{c})$, with $f_z(\cdot)$ representing a neural network.

To generate a sufficient number of candidates for the robot's navigation, we need to create multiple diverse trajectories that cover all traversable areas in front of the robot. Since the decoder is designed to generate a single trajectory from one embedding vector, producing a variety of diverse trajectories requires the use of representative

and varied embedding vectors. We project the embedding vector $\mathbf{z}$ onto orthogonal axes by linear transformations, each projected vector corresponding to one traversable area. Then we generate trajectories based on the condition $\mathbf{c}$:

$$\mathbf{z}_k = A_k(\mathbf{c})\mathbf{z} + b_k(\mathbf{c}) = h_{\psi_k}(\mathbf{z}),$$

where h_{ψ_k} denotes the linear transformation of $\mathbf{z}$. Using each embedding vector $\mathbf{z}_k$, the decoder generates a trajectory τ_k, as $p\left(\tau_k|\mathbf{z}_k, \mathbf{c}, \bar{\mathcal{Z}}_k\right)$. $\tau_k \in \mathcal{T}$ represents generated trajectories. $\mathbf{z}_k$ and $\bar{\mathcal{Z}}_k$ are the embedding vectors of the current trajectory and the set of other trajectory embeddings, respectively. The training of the trajectory generator is the same as MTG [5], where we use traversability loss, CVAE lower bound, and diversity loss to train the model.

C. VLM-based Trajectory Selection

While the generated trajectories $\mathcal{T}_n$ effectively cover the traversable areas in front of the robot [5], the deep-learning-based generative model cannot guarantee the consistent generation of traversable trajectories. To address this, we sample consecutive $t = 2$ time steps, introducing redundancy to increase the likelihood that at least one of the generated trajectories will be traversable. Given the collected trajectories in $\mathcal{T}$, we convert them to the image plane with numbers, where we sort the trajectories in terms of heuristic, which is the distance between the last waypoint of the trajectory and the goal, as shown in Eq. 1.

We then project the trajectories $\mathcal{T}$ from the robot's frame to the image plane by transformation matrices K, $\mathcal{T}_c = P_c(\mathcal{T}, K)$. Following the trajectory generation sequence, we annotate the trajectories with numbers, $\mathbf{n}$. Finally, we use the VLM to select the best trajectory in terms of satisfying traversability and social compliance. The annotated trajectories $(\mathbf{n}, \mathcal{T}_c)$ and the current observation image $\mathbf{i}$ are input into the VLM with the prompt instruction ℓ. The VLM selects the best trajectory, τ, in terms of traversability, social compliance, and traveling distance to the goal.

Given the selected trajectory τ, our motion planner generates the corresponding robot action $\mathbf{a}$ to follow it. The VLM is re-prompted each time it returns a response. Although our VLM-based trajectory selector operates at a relatively low frequency, *i.e.,* , every 2 to 4 seconds, the trajectory generator efficiently produces 10m trajectories, ensuring the latency remains manageable.

III. Exprimental Results

A. Implementation Details

Our approach is tested on a Clearpath Husky equipped with a Velodyne VLP16 LiDAR, a Realsense D435i camera, and a laptop with an Intel i7 CPU and an Nvidia GeForce RTX 2080 GPU. We use CVAE [5] with an attention mechanism to generate multiple trajectories (approximately 10m each) and use GPT-4V [22] as the VLM to select the best traversable trajectory.

The training dataset [23] for our CVAE-based trajectory generation model contains three parts: 1) LiDAR point cloud and robot velocities, 2) binary traversability maps, 3) randomly generated diverse targets with the shortest ground truth trajectories to the targets. The binary traversability map is constructed from LiDAR points and is used only for training and evaluation and not used during inference.

To validate VL-TGS, we present qualitative and quantitative results compared with MTG [5], ViNT [19], NoMaD [20], PIVOT [17], and CoNVOI [21]. We evaluate the performance in four benchmark scenarios:

- **Flower bed:** A robot navigating a paved area next to a flower bed. The robot must stay on the paved path and avoid entering the flower bed.
- **Curb:** A robot navigating on a sidewalk, which is distinctly separated from the roadway by a curb. The robot must stay on a sidewalk or select a traversable trajectory to go around the curb.
- **Crosswalk:** A robot crossing the street. The robot must stay on the crosswalk when crossing the street.
- **Behind the corner:** When the target is behind an obstruction, and there is a large open space ahead, the straight path may lead to an obstacle. The robot must choose a trajectory to navigate around the corner.

B. Qualitative Results

Fig. 3 shows the resulting robot trajectories corresponding to six different approaches in four different scenarios. The upper row shows the trajectories generated by all the comparison methods including ours and the lower row shows the results of VL-TGS with the candidate trajectories (gray) and the selected one (red).

As MTG relies solely on LiDAR's geometric data, it is unable to deal with traversability differences in flower beds, curbs, and crosswalks, where structure alone provides little distinction. The performances of ViNT and NoMaD heavily depend on the quality of pre-built topological maps. While they perform well when following straight paths with distinct visual features, such as a crosswalk, they often struggle in environments with turns or significant scene variations. While PIVOT selects the most semantically feasible trajectory from the given candidates, it does not explicitly detect geometric information and its random trajectory generation disregards both geometric and semantic information, potentially resulting in no viable options for the VLM to choose from. Compared to other methods, CoNVOI generally produces trajectories that are both geometrically and semantically feasible. However, its zigzag motion results in non-smooth robot movements. As shown in the bottom row of each scenario in Fig. 3, our approach produces diverse trajectories and selects the best one that is traversable and contextually appropriate.

C. Quantitative Results

To further validate VL-TGS, we evaluate the methods using two different metrics:

- **Traversability:** The ratio of the generated trajectory lying on a traversable area. The binary traversability map, initially generated using LiDAR and then manually refined, is used for evaluation. This metric is

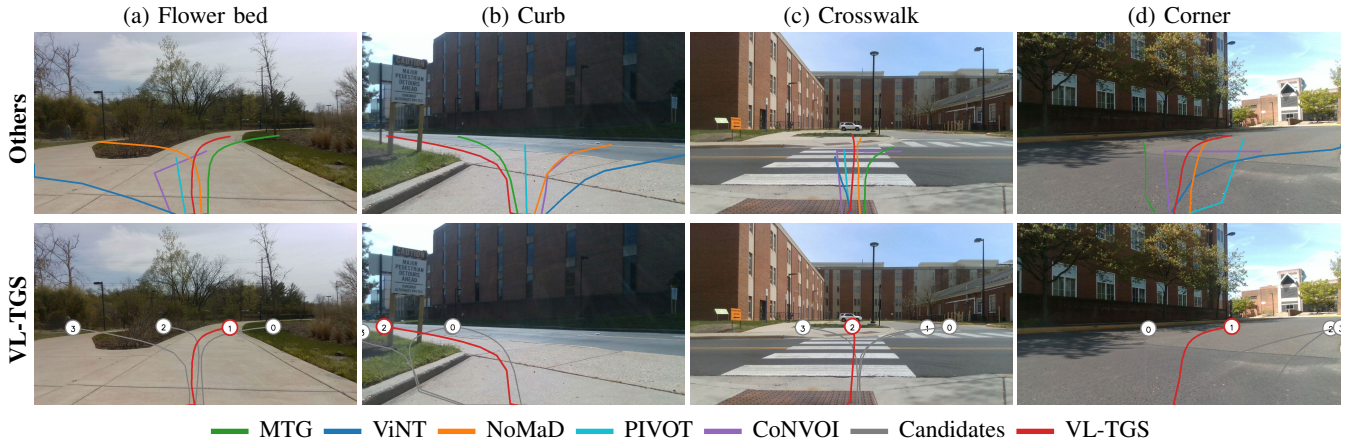

Fig. 3. **Qualitative Results:** The top row shows the generated trajectories using all the methods, MTG [5] in green, ViNT [19] in blue, NoMaD [20] in orange, PIVOT [17] in cyan, CoNVOI [21] in purple, and VL-TGS in red. The bottom row shows the candidate trajectories in gray marked with numbers and the selected trajectory in red using VL-TGS. VL-TGS can generate and select a trajectory that is both geometrically and semantically feasible.

calculated as

$$tr(\mathcal{A}, \hat{\boldsymbol{\tau}}) = \sum_{m=1}^{M} c(\mathcal{A}, \mathbf{w}_m), \quad \mathbf{w}_m \in \hat{\boldsymbol{\tau}}, \tag{4}$$

where $c(\cdot, \cdot)$ tells if the waypoint $\mathbf{w}_m$ is in the traversable area $\mathcal{A}$.

- **Fréchet Distance w.r.t. Human Tele-operation:** Fréchet Distance [24] is one of the measures of similarity between two curves. We measure the similarity between the trajectories generated by the methods and human-like trajectories, which are collected by human tele-operating the robot. A lower distance indicates a higher degree of similarity.

Table I reports the results averaged over 20 different frames, with five repetitions for each frame, scenario, and method. In the Input column, L indicates LiDAR point cloud and I indicates RGB images. While MTG, ViNT, NoMad, and PIVOT rely on a single sensory input, CoNVOI and VL-TGS utilize both LiDAR point clouds and RGB images. The results demonstrate that VL-TGS outperforms other state-of-the-art approaches in most of the cases. Specifically, we achieve at least 3.35% and at most 47.74% improvement in terms of average traversability, and at least 19.62% and at most 40.99% improvement in terms of average Fréchet distance.

TABLE I

QUANTITATIVE RESULTS: COMPARISONS WITH OTHER METHODS

Metric	Method	Input	Scenario			
			Flower bed	Curb	Crosswalk	Corner
Traversability (%) ↑	MTG	L	58.19	67.12	61.82	44.71
	ViNT	I	63.62	78.37	84.78	44.95
	NoMaD	I	75.64	83.13	79.24	77.38
	PIVOT	I	64.75	79.58	76.78	68.66
	CoNVOI	I+L	81.10	75.68	86.24	**88.46**
	VL-TGS	I+L	**87.22**	**89.93**	**87.44**	78.00
Fréchet Distance (m) ↓	MTG	L	6.61	8.40	10.42	9.93
	ViNT	I	10.43	10.78	8.94	12.71
	NoMaD	I	7.65	8.71	11.87	9.62
	PIVOT	I	8.41	7.86	10.53	9.48
	CoNVOI	I+L	11.64	12.24	11.33	12.36
	VL-TGS	I+L	**5.27**	**7.93**	**6.38**	**8.49**

We observe that MTG produces very low results in terms of traversability. This is not only because our benchmark scenarios were selected based on scenarios that are difficult to detect with LiDAR, but also because MTG often fails to consider traversability while focusing on optimality to the goal. In terms of Fréchet distance, MTG and VL-TGS produce good results because they output smooth trajectories similar to a human-operated trajectory we compare against. In contrast, CoNVOI generates a linear trajectory that differs significantly from typical human-operated trajectories, resulting in a lower similarity. CoNVOI generates short trajectories using only two waypoints, reducing the likelihood of waypoints landing in non-traversable areas and leading to a high traversability result. However, in practice, intermediate points may still fall into non-traversable regions. Both ViNT and NoMaD are image-based navigation approaches, but NoMaD generally outperforms ViNT. While both perform well in straight-line following scenarios (e.g., crosswalks), they tend to go off-course when robots are taking turns or the scenarios are dynamic. Additionally, since some of our flower bed and curb scenarios included smooth turns, their variance is notably high. As PIVOT generates random straight-line candidates, its performance is inconsistent, exhibiting high variation in results. The result demonstrates that VL-TGS generates human-like trajectories in human-centered environments while ensuring good traversability.

IV. CONCLUSION

We propose VL-TGS, a novel multi-modal Trajectory Generation and Selection approach for mapless outdoor navigation. VL-TGS integrates a CVAE-based trajectory generation method with a VLM-based trajectory selection process to compute geometrically and semantically feasible, human-like trajectories in human-centered outdoor environments. Our approach achieves a 20.81% improvement in traversability and a 28.51% improvement in similarity to human-operated trajectories on average.

REFERENCES

[1] D. Song, J. Liang, X. Xiao, and D. Manocha, "Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5791–5798, 2025.

[2] R. Möller *et al.*, "A survey on human-aware robot navigation," *Robotics and Autonomous Systems*, vol. 145, p. 103837, 2021.

[3] L. Wijayathunga, A. Rassau, and D. Chai, "Challenges and solutions for autonomous ground robot scene understanding and navigation in unstructured outdoor environments: A review," *Applied Sciences*, vol. 13, no. 17, p. 9877, 2023.

[4] I. Jeong, Y. Jang, J. Park, and Y. K. Cho, "Motion planning of mobile robots for autonomous navigation on uneven ground surfaces," *Journal of Computing in Civil Engineering*, vol. 35, no. 3, p. 04021001, 2021.

[5] J. Liang *et al.*, "Mtg: Mapless trajectory generator with traversability coverage for outdoor navigation," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 2396–2402.

[6] J. Liang, A. Payandeh, D. Song, X. Xiao, and D. Manocha, "Dtg: Diffusion-based trajectory generation for mapless global navigation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 5340–5347.

[7] K. Charalampous, I. Kostavelis, and A. Gasteratos, "Recent trends in social aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 93, pp. 85–104, 2017.

[8] K. Weerakoon *et al.*, "Graspe: Graph based multimodal fusion for robot navigation in outdoor environments," *IEEE Robotics and Automation Letters*, 2023.

[9] J. Zhang *et al.*, "Trans4trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 173–19 186, 2022.

[10] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[11] P. V. Borges *et al.*, "A survey on terrain traversability analysis for autonomous ground vehicles: Methods, sensors, and challenges." *Field Robotics*, vol. 2, no. 1, pp. 1567–1627, 2022.

[12] J. Austin *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.

[13] H. Ha and S. Song, "Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models," *arXiv preprint arXiv:2207.11514*, 2022.

[14] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[15] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[16] A. Shtedritski *et al.*, "What does clip know about a red circle? visual prompt engineering for vlms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 987–11 997.

[17] S. Nasiriany *et al.*, "Pivot: Iterative visual prompting elicits actionable knowledge for vlms," *arXiv preprint arXiv:2402.07872*, 2024.

[18] J. Yang *et al.*, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.

[19] 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.

[20] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 63–70.

[21] A. J. Sathyamoorthy, K. Weerakoon, M. Elnoor, A. Zore, B. Ichter, F. Xia, J. Tan, W. Yu, and D. Manocha, "Convoi: Context-aware navigation using vision language models in outdoor and indoor environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13 837–13 844.

[22] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[23] J. Liang *et al.*, "Gnd: Global navigation dataset with multi-modal perception and multi-category traversability in outdoor campus environments," *IEEE International Conference on Robotics and Automation*, 2025.

[24] H. Alt and M. Godau, "Computing the fréchet distance between two polygonal curves," *International Journal of Computational Geometry & Applications*, vol. 5, no. 01n02, pp. 75–91, 1995.