Comparison of Specialized and Prompt-Driven Vision Models for Off-Road Traversability Estimation on RELLIS-3D

Mrinank Sivakumar
Research Project B
University of Adelaide
Adelaide, Australia
mrinank.sivakumar@student.adelaide.edu.au

Dr. Feras Dayoub
Research Supervisor
Australian Institute of Machine Learning
Adelaide, Australia
feras.dayoub@adelaide.edu.au

***Abstract*—Reliable traversability estimation is essential for autonomous navigation. Especially in unstructured off-road environments, where visual ambiguity, irregular terrain, and weak structural cues pose significant challenges for perception systems. This project investigates two contrasting approaches for binary traversability prediction using monocular RGB imagery from the RELLIS-3D dataset. These approaches are a specialized, supervised model, Follow-the-Footprints (FTF), and a prompt-driven segmentation pipeline constructed from vision foundation models, Grounded SAM (GSAM). FTF directly learns traversability from labelled data, whereas GSAM leverages pre-trained open-world detection and segmentation models guided by natural-language prompts. Quantitative evaluation shows that both methods achieve comparable overall F1 and IoU scores, but with distinct error profiles. FTF exhibits very high recall and low precision, while GSAM demonstrates high precision but substantial under-segmentation. Qualitative analysis further reveals that FTF produces consistent but sometimes overly lenient masks. However, GSAM produces conservative predictions shaped by prompt phrasing. These findings highlight trade-offs between supervised, task-specific learning and flexible, prompt-based vision. This suggests that while foundation-model pipelines offer promising adaptability, specialized models currently remain more dependable for dense off-road traversability estimation.**

***Index Terms*—Traversability estimation, off-road robotics, semantic segmentation, vision foundation models, Grounded SAM, Follow-the-Footprints, autonomous navigation, RELLIS-3D.**

I. Introduction

A. Context

Autonomous navigation in unstructured, off-road environments continues to be a core challenge in mobile robotics. Unlike urban settings, off-road environments exhibit high visual variability, weak structural cues, and terrain boundaries that blend rather than distinctly transition. Factors such as vegetation, loose materials, shadows, and inconsistent lighting further complicate terrain interpretation. As such, robots must reliably infer safe regions from complex visual data, making traversability estimation a fundamental component of robust off-road autonomy.

Vision-based methods are preferred for this task, as monocular RGB cameras are lightweight, inexpensive, and widely deployable. Deep learning approaches, such as convolutional neural networks (CNNs), have shown strong capabilities in extracting semantic information. However, applying them to these chaotic environments remains difficult. The struggle to maintain pixel-level accuracy despite these uncertain factors is what motivates the exploration of both specialized models trained for traversability and more flexible general-purpose segmentation systems.

B. Problem Statement

This project explores binary traversability estimation using monocular RGB images from the RELLIS-3D dataset. The goal is to classify each pixel as traversable or non-traversable and ultimately enable a robot to plan safe paths through natural environments. Traditional approaches to this problem implement task-specific supervised learning, such as the Follow-the-Footprints (FTF) model used in this study. FTF directly learns a mapping from image pixels to traversability labels but requires curated training data and retraining whenever task requirements change.

An alternative perspective comes from segmentation pipelines based off large vision foundation models. In this project, a Grounded-SAM (GSAM) pipeline is constructed using Grounding DINO for prompt-driven region detection and the Segment Anything Model (SAM) for mask refinement. This approach does not learn traversability directly, but rather generates masks based on object categories associated with safe or unsafe terrain. Whether such a flexible, prompt-based system can approximate or rival specialized, supervised models in off-road settings remains unclear. This comparison forms the central focus of the study.

C. Research Questions

To guide the investigation, the project addresses the following research questions:

1) **How effectively can a prompt-driven segmentation pipeline (GSAM) generate binary traversability masks compared to a specialized supervised model (FTF) on the RELLIS-3D dataset?**

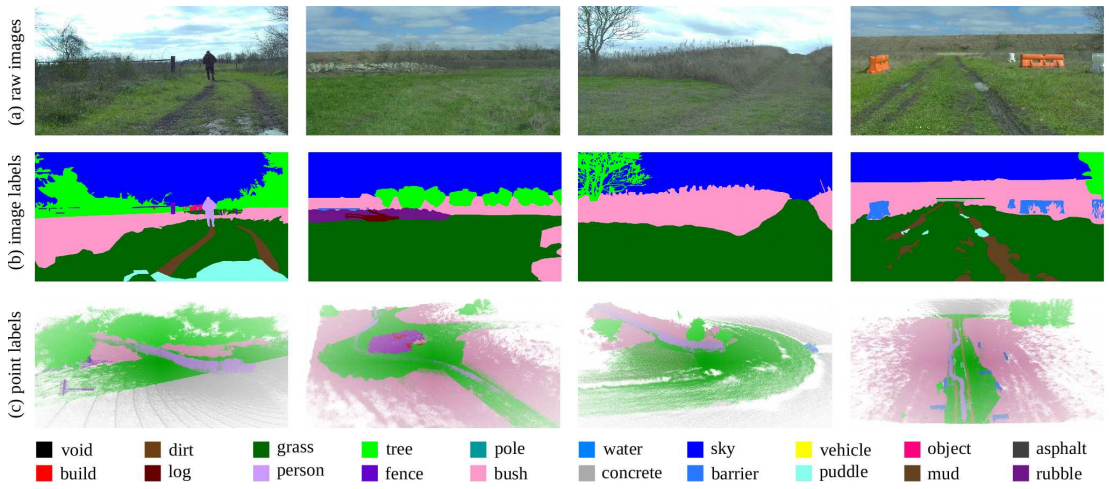

Fig. 1. An example from the dataset showing its dense semantic segmentation labels.

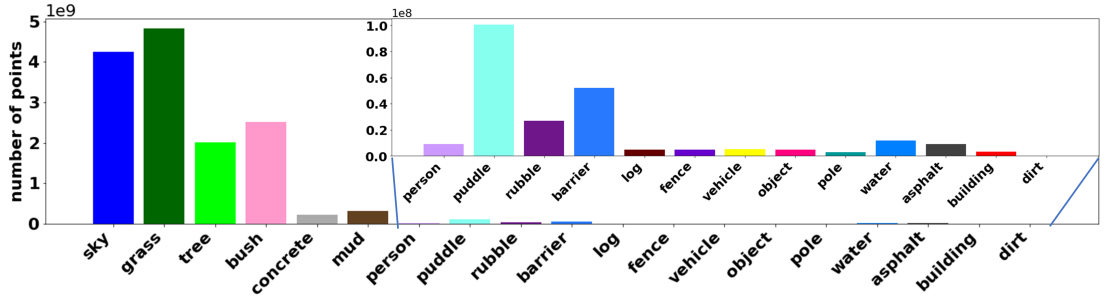

Fig. 2. A visualization of the image distribution showing class imbalance in image annotations.

2) **What differences emerge in the qualitative behaviour of FTF and GSAM when interpreting complex off-road scenes?**
3) **What do these performance and behavioural differences imply for the suitability of prompt-based vision models in autonomous off-road robotics?**

D. Dataset Overview

As mentioned earlier, all experiments are conducted using the RELLIS-3D dataset, which is a multimodal off-road perception dataset. It was collected from a ground vehicle that operated across natural and semi-structured environments. RELLIS-3D provides high-resolution RGB images with dense semantic segmentation labels covering terrain types such as dirt, grass, gravel, mud, vegetation, rocks, fences, and buildings [Fig. 1]. For this study, these semantic classes are mapped into a binary traversability format: flat, low-risk terrain is labelled traversable, while obstacles and hazardous regions are labelled non-traversable.

A notable property of RELLIS-3D is its class imbalance. Traversable terrain often dominates the frame, while non-traversable objects appear in smaller, visually diverse forms [Fig. 2]. Combined with variations in illumination, shadows, and terrain texture, these characteristics create a realistic and challenging benchmark for evaluating traversability prediction models.

E. Overview of Methods & High-Level Results

Two contrasting methods were chosen to be evaluated. The first being FTF, a specialized convolutional encoder–decoder network trained on binary traversability masks. The second is the GSAM pipeline, which constructs traversability masks by detecting prompt-defined regions with Grounding DINO and refining them with SAM. This comparison effectively pits supervised, task-specific learning against a flexible, prompt-driven segmentation strategy.

High-level results reveal distinct performance profiles. FTF achieves higher overall accuracy and very strong recall, indicating reliable detection of traversable regions. Meanwhile, GSAM attains higher precision, reflecting conservative but accurate positive predictions, though it occasionally under-segments terrain. Qualitatively, FTF produces smooth and coherent traversability maps, whereas GSAM's outputs reflect the influence of its prompt selection and the generalization behaviour of foundation models. These findings motivate a deeper analysis in the following sections.

II. Literature Review

Traversability estimation has been recognized as a core problem in mobile robotics for several decades. Early work

relied on geometric cues from range sensors and stereo vision, using hand-crafted features and heuristic thresholds to classify terrain as safe or unsafe. Manduchi et al. combined obstacle detection with terrain classification for autonomous off-road navigation, and showed that purely geometric reasoning struggles in cluttered, visually diverse environments where vegetation, rocks, and ruts violate planar assumptions [18]. Angelova et al. demonstrated that visual appearance can be used to predict slip and terrain properties, highlighting the importance of learning from image data rather than relying solely on geometry [1]. A recent survey by Sevastopoulos and Konstantopoulos traces this evolution from geometry-based to learning-based and appearance-based methods. It emphasizes the challenges posed by ambiguous textures, occlusions, and varying surface conditions in unstructured outdoor scenes [35].

With the emergence of deep learning, semantic segmentation and learned feature extractors have become central to terrain estimation. Valada et al. proposed self-supervised model adaptation for multimodal semantic segmentation. This would enable models to adapt to new environments using unlabelled data and mitigate domain shift between training and deployment [37]. Architectures such as Mask2Former further unify instance, panoptic, and semantic segmentation under a transformer-based framework [5], illustrating a broader trend towards flexible, general dense prediction models. Despite these advances, off-road environments remain difficult due to class imbalance, subtle transitions between traversable and non-traversable terrain, and small but critical obstacles that are easy to miss.

Recent research has focused on traversability-specific learning frameworks. Follow-the-Footprints learns traversability by combining geometric and visual cues in a self-supervised manner. It uses the vehicle's own motion as a supervisory signal to infer which regions led to successful traversal [12]. RoadRunner extends this idea by training traversability estimators for fast off-road driving using large-scale, real-world data and multi-sensor inputs [8]. Wild Visual Navigation leverages pre-trained models and online self-supervision to rapidly adapt traversability predictions to new terrain [19]. These works show that specialized models can learn robust traversability representations when supported by appropriate data and training regimes. However, they typically require task-specific engineering and curated datasets.

High-quality datasets have been crucial for benchmarking these methods. RELLIS-3D provides high-resolution RGB imagery, LiDAR, and dense semantic labels collected from a ground vehicle operating in natural and semi-structured environments [13]. Its class set includes multiple ground, vegetation, obstacle, and man-made categories, with strong class imbalance and substantial variability in lighting and layout. Jiang et al. highlight that RELLIS-3D poses significant challenges for both semantic segmentation and off-road perception, making it an appropriate benchmark for traversability estimation [13]. In this project, the semantic labels are collapsed into a binary traversable/non-traversable representation, aligning the dataset with the traversability task while retaining its difficulty.

In parallel, vision foundation models (VFMs) and vision–language models (VLMs) have emerged as powerful general-purpose tools for visual understanding. CLIP learns image–text alignment from large-scale web data and supports zero-shot recognition through natural-language prompts [32]. GLIP goes one step further and applies grounded language–image pre-training for open-set detection [15]. Meanwhile, Liang et al. extend CLIP-style representations to dense prediction through Mask-adapted CLIP for open-vocabulary semantic segmentation [16]. The Segment Anything Model (SAM) introduces promptable segmentation at scale [14], and Grounding DINO aligns textual queries with object bounding boxes for open-set object detection [17]. Ren et al. combine these components into Grounded SAM, a framework capable of detecting and segmenting user-defined concepts specified by natural-language prompts [33]. Together, these works suggest that prompt-driven perception using foundation models may reduce the need for task-specific training in some domains.

The influence of foundation models is beginning to extend to off-road robotics. AnyTraverse integrates a VLM with a human operator to build traversability maps in off-road environments, demonstrating how open-vocabulary reasoning can support navigation [34]. However, such systems often operate at a higher semantic level, rely on human input, or use VFMs as supplementary components rather than as the primary mechanism for dense traversability prediction. They also rarely provide direct, quantitative comparisons against specialized traversability models trained on the same dataset.

Additional techniques from the broader computer vision literature also inform traversability models. Depth completion networks such as GuideNet use guided convolutions to fuse RGB and depth, improving dense predictions [36]. Scribble-supervised segmentation methods, such as the uncertainty reduction framework of Pan et al., show that strong segmentation performance can be achieved from sparse annotations via uncertainty modelling and self-supervision in feature space [22]. These works underline a trend towards combining strong priors from pre-trained models with weak or self-supervision when labels are limited.

Despite this rich body of work, there is limited experimental evidence on how a prompt-driven, segmentation pipeline constructed from foundation models compares directly with a task-specific, supervised traversability model. VFM-based systems typically focus on semantic understanding or interactive mapping rather than thorough pixel-wise traversability benchmarking [34]. Similarly, the trade-off between recall-oriented specialized models, which aim to minimize missed safe terrain [8], [12], and precision-oriented prompt-driven pipelines has not been systematically explored. This project addresses that gap by comparing Follow-the-Footprints and a Grounded-SAM-based pipeline under a shared binary traversability formulation, using RELLIS-3D as a common evaluation dataset.

III. Project Review

While the Literature Review outlined the broader academic landscape, this section focuses on concrete implementations that relate most directly to this project.

A. Specialized Traversability Models

Recent traversability frameworks learn terrain safety directly from data, often through self-supervision. FTF is one such model and serves as the specialized baseline in this study. It generates traversability maps by using the vehicle's future path as a supervisory signal, learning which regions were actually traversed [12]. Architecturally, FTF adopts an encoder–decoder design and originally merges both RGB and geometric inputs to estimate a continuous traversability cost.

Other specialized systems follow similar principles. RoadRunner trains high-speed off-road traversability estimators using large-scale, multi-sensor datasets [8]. Meanwhile, Wild Visual Navigation adapts traversability predictions online through self-supervision, allowing rapid adjustment to unfamiliar terrain [19]. All of these approaches treat traversability as a primary output, shaping their data collection and model structure around that goal.

In this project, FTF is adapted into a simplified RGB-only binary segmentation model to ensure a fair architectural comparison with GSAM. This restricts FTF relative to its original design, but isolates the effect of supervised visual learning on the traversability task.

B. Prompt-Driven Segmentation and Foundation-Model Pipelines

Parallel work has explored whether general-purpose foundation models can support off-road perception. Grounded SAM combines Grounding DINO for text-driven detection with SAM for high-resolution mask generation [14], [17], [33]. Although not intended for traversability estimation, this framework provides a template for constructing prompt-based segmentation pipelines.

AnyTraverse applies vision–language reasoning to off-road environments by integrating a VLM with human-in-the-loop concept selection [34]. While effective for open-vocabulary scene understanding, it does not provide a direct comparison to specialized traversability models and relies heavily on operator input.

In this project, GSAM is adapted for fully automated traversability prediction. Prompt categories such as "dirt", "trail", and "grass" guide Grounding DINO detections, which SAM then refines into binary masks. Unlike FTF, GSAM performs no task-specific training, relying solely on pre-trained priors and prompt design. This raises the central question of how closely such a system can approximate a supervised traversability model.

C. Hybrid and Weakly Supervised Components

Several supporting techniques from the literature inform the design of both specialized and prompt-driven systems. GuideNet demonstrates how guided convolutions can fuse RGB and depth to improve dense predictions [36], while Pan et al. show that scribble-supervised segmentation can reduce annotation cost using uncertainty modelling and feature-space self-supervision [22]. Although not used directly here, these methods highlight broader trends toward combining strong priors with weak or self-supervision—principles relevant to both FTF and GSAM.

D. Positioning of the Present Project

Overall, the existing work separates into two directions. First, specialized models such as FTF, RoadRunner, and Wild Visual Navigation, which are explicitly trained for traversability [8], [12], [19]. Then, foundation-model-based pipelines such as Grounded SAM and AnyTraverse, which emphasize open-world semantic understanding [33], [34]. However, these two lines of research rarely intersect, and prompt-driven models have not been rigorously evaluated as drop-in alternatives to supervised traversability networks.

This project bridges that gap by directly comparing FTF and GSAM on the same dataset and under a shared binary traversability definition. The aim is not to propose a new architecture, but to provide a controlled, systematic evaluation of how supervised and prompt-driven approaches differ in accuracy, qualitative behaviour, and practical suitability for autonomous off-road robotics.

IV. Methodology

A. Experimental Setup and Environment

All experiments were implemented in Python using PyTorch for model development [30]. NumPy, SciPy, and scikit-learn handled numerical operations and evaluation utilities [9], [23], [38], while OpenCV and Pillow were used for image loading and preprocessing [3], [6]. Configuration management relied on PyYAML and `easydict` [31], while standard Python modules, such as `argparse`, `os`, and `pathlib`) were used for command-line tools and filesystem access.

A dedicated Miniconda environment ensured dependency isolation [2]. All experiments were executed in WSL2 with GPU acceleration, following Microsoft's documentation for CUDA-enabled workflows [20]. The complete codebase and experiment configurations will eventually be uploaded to a GitHub repository to support reproducibility [4].

B. Data Preparation and Binary Labelling

From the RELLIS-3D dataset, only RGB images and semantic labels were used [13]. The original label set was collapsed into a binary traversability format where terrain such as dirt, gravel, and short or managed grass was labelled as traversable, while vegetation, obstacles, structures, and ambiguous regions were labelled non-traversable. This reflects a conservative off-road driving strategy.

Images were resized to a uniform resolution and normalized using standard ImageNet statistics. Binary masks were resized with nearest-neighbour interpolation. Augmentation was intentionally minimal to ensure that differences in performance reflected the models rather than augmentation strategies.

C. FTF Baseline

The FTF model was implemented following its published architecture and reference code [12]. The original architecture fuses RGB and geometric inputs, but in this project the model was adapted to operate on RGB alone to enable a controlled comparison with GSAM. The model was trained as a binary segmentation network using a combination of binary cross-entropy and IoU/Dice-style losses, with ImageNet-initialized encoders and randomly initialized decoders. Model selection was performed on a validation split based on F1 or IoU.

Restricting FTF to RGB-only inputs simplifies the comparison, but removes one of its core advantages. This limitation is noted when interpreting the results.

D. GSAM Pipeline

The GSAM pipeline combines Grounding DINO for prompt-based detection and SAM for segmentation [14], [17], [33]. No training was performed. Instead, GSAM relies entirely on pre-trained foundation models and natural-language prompts.

A list of terrain-related prompts (e.g., "road", "dirt", "gravel", "trail", "short grass") guided Grounding DINO detections. SAM then produced pixel-accurate masks for each detection, which were merged into a single traversable region via logical union. Because GSAM never accesses ground-truth labels, its performance depends solely on the semantics encoded in its prompts and pre-trained weights.

E. Evaluation Protocol

Both models were evaluated on the same RELLIS-3D image set using pixel-wise binary labels. Traversable terrain was treated as the positive class. Accuracy, precision, recall, F1 score, and IoU were computed. These metrics jointly capture both risk-sensitive properties, such as avoiding false positives on hazardous regions, and coverage of safe terrain.

Qualitative evaluation was also performed by visualizing predictions alongside input images and ground-truth labels. Representative examples were selected to illustrate typical success cases and characteristic failure cases for each method.

F. Ethics Statement

This project uses only publicly released datasets and open-source software. No human participants or sensitive data were involved. All models and libraries, including FTF, Grounding DINO, SAM, and supporting codebases, were used under their respective licences [12], [14], [17], [33]. Experiments were conducted entirely offline, and no model produced outputs that informed real-world robotic behaviour.

Because traversability estimation relates to safety-critical autonomy, particular care is taken in reporting limitations and failure modes. The environmental impact of training was kept modest by training only the FTF baseline and relying on pre-trained foundation models for GSAM.

V. Results

This section presents the quantitative and qualitative results obtained from evaluating FTF and GSAM on the RELLIS-3D image set. Metrics were computed at the pixel level using the binary traversability labels described earlier.

A. Quantitative Performance

Table I summarizes the classification metrics for both models. FTF achieves higher overall accuracy (0.703) than GSAM (0.679), driven by its strong recall of traversable regions (0.959). This indicates that FTF rarely misses areas that are genuinely safe to traverse, but it does so at the cost of low precision (0.164), with many traversable predictions falling on pixels that should be non-traversable.

GSAM shows the opposite trend. Its precision is substantially higher (0.766), meaning that predicted traversable pixels are usually correct, but its recall is much lower (0.173), reflecting conservative masks that overlook large portions of usable terrain. The F1 scores (0.280 vs. 0.283) and IoU values (0.163 vs. 0.165) are nearly identical, suggesting that while the models behave very differently, their overall segmentation performance is comparable.

TABLE I
Traversability Metrics on RELLIS-3D

Metric	FTF	GSAM
Accuracy	0.703	0.679
Precision	0.164	0.766
Recall	0.959	0.173
F1 Score	0.280	0.283
IoU	0.163	0.165

A normalized confusion matrix for GSAM is shown in Figure 3. The diagonal dominance in the non-traversable class reflects strong specificity but weak sensitivity. FTF exhibits the opposite pattern (high sensitivity, weaker specificity), consistent with its recall-oriented behaviour.

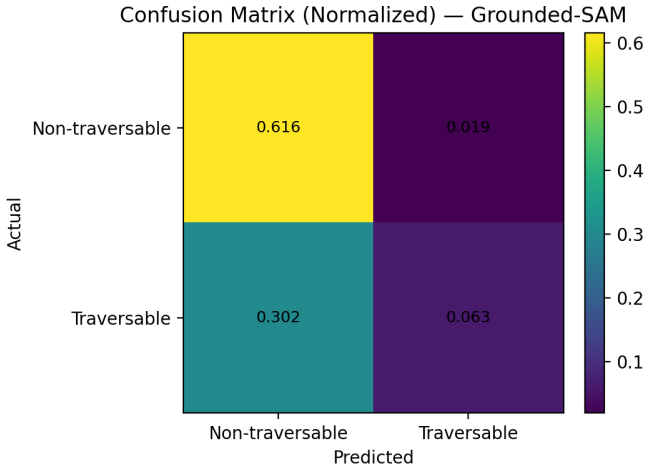

Fig. 3. A confusion matrix for Grounded SAM comparing Predicted and Actual traversability.

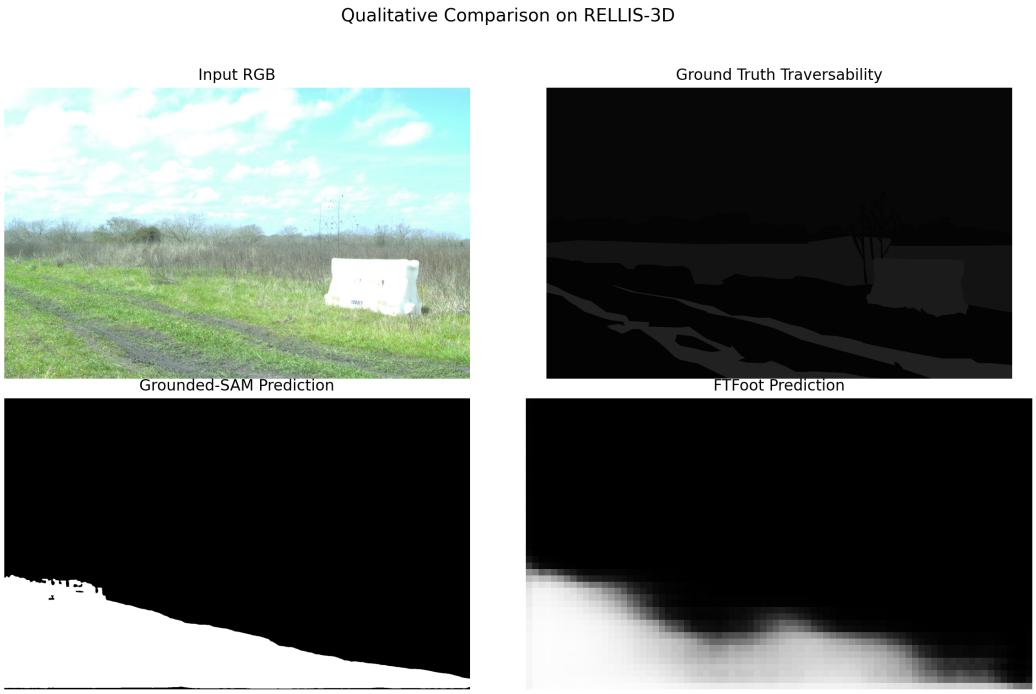

Fig. 4. A qualitative comparison of the raw image, the annotated image, and the binary masks generated by both models.

B. Qualitative Evaluation

Figure 4 presents representative qualitative examples for both methods. Three consistent behaviours are observed:

1) **FTF produces smooth, spatially coherent traversability maps**. It often captures the overall shape of walkable or driveable terrain, but tends to overestimate safe regions in ambiguous areas such as tall grass or uneven ground.
2) **GSAM generates conservative traversability masks**. It marks compact regions that strongly match terrain-related prompts, such as "dirt" or "trail", but under-segments broader ground areas, leaving substantial traversable regions unlabelled.
3) **Prompt sensitivity shapes GSAM's predictions**. Where prompts do not align well with the visual appearance of the scene, GSAM may return very small or even empty masks. Although such extreme cases are rare, they highlight the dependence on prompt design and semantic coverage.

Overall, the qualitative results reinforce the quantitative findings: FTF behaves as a high-recall, low-precision model, while GSAM behaves as a high-precision, low-recall model. These complementary behaviours are examined in more detail in the Discussion.

VI. Discussion

The results highlight fundamentally different behaviours between the specialized FTF model and the prompt-driven GSAM pipeline. Although both achieved similar overall F1 and IoU scores, their error profiles and operational characteristics diverge sharply. This section interprets these findings in relation to the research questions and the broader context of off-road traversability estimation.

A. Model Behaviour and Research Questions

Research Question 1 asked how effectively GSAM can generate traversability masks relative to FTF. Quantitatively, GSAM matches FTF in F1 and IoU, showing that a foundation-model pipeline can approximate a supervised model in overall performance. However, GSAM's high precision and low recall reveal a conservative strategy. It is reliable when predicting traversable pixels, but it does not capture much of the available terrain. FTF displays the opposite pattern, with extremely high recall and poor precision. These contrasting profiles point to different safety trade-offs. GSAM prioritizes avoiding risky predictions, whereas FTF prioritizes covering as much safe terrain as possible.

Research Question 2 focused on qualitative differences in prediction behaviour. FTF produces smooth, continuous traversability maps but often overestimates safe areas in ambiguous vegetation or uneven ground. This is consistent with a model trained to recover all potentially drivable surfaces but constrained to RGB-only inputs. GSAM's predictions are shaped by prompt semantics and pre-trained representations. It segments surfaces that clearly match prompts such as "dirt" or "trail", but under-segments broader ground regions and may return minimal masks when prompts do not align well with local appearance. This prompt-aligned behaviour is predictable

but restrictive, as semantic categories do not always coincide with drivability.

Research Question 3 addressed the suitability of prompt-driven models for off-road robotics. The findings suggest that GSAM is best viewed as a high-precision component within a larger perception stack, especially where false positives carry high risk. Its low recall limits its standalone viability for navigation, as overly conservative masks may lead to missed path opportunities or overly cautious planning. FTF's high recall makes it attractive for maximizing workspace, but its false positives must be mitigated through downstream filtering, additional sensors, or more expressive cost formulations.

B. Contextualizing Results with Prior Work

These observations are consistent with trends from previous works. Specialized methods such as FTF, RoadRunner, and Wild Visual Navigation show strong performance when trained directly on traversability data and tailored to the task [8], [12], [19]. The present results support this, even under the simplified RGB-only setting. Foundation-model pipelines, represented here by GSAM, behave similarly to systems such as Grounded SAM and AnyTraverse [33], [34]. They offer rich semantic understanding, but lack explicit modelling of affordances and terrain continuity. The strong dependence on prompt design also reflects findings from open-vocabulary segmentation, where prompt choice significantly affects mask quality.

C. Limitations and Implications

The main limitation of this study is that FTF was evaluated without its geometric input channels, which likely suppressed its precision and boundary accuracy. A multimodal variant would better represent real-world deployment and may widen the performance gap. For GSAM, prompt engineering is a practical limitation. Prompt sets may require environment-specific tuning, and foundation models are not trained for traversability, so their semantics only indirectly reflect terrain affordances. Finally, the use of a single dataset means that results may not generalize to all off-road domains, although RELLIS-3D remains a challenging and relevant benchmark.

D. Recommendations and Future Directions

A natural next step is to explore hybrid systems that take advantage of the complementary strengths of both approaches. For instance, GSAM could be used to suppress high-confidence non-traversable regions, while FTF expands traversable areas using recall-driven predictions. Alternatively, GSAM masks could serve as additional supervision or regularization signals during training. Future work should also investigate multimodal FTF variants, automated or learned prompt tuning for GSAM, and evaluations across multiple off-road datasets to assess generalization. Overall, the results indicate that while foundation-model pipelines provide useful semantic priors and flexibility, specialized models currently remain more reliable for dense off-road traversability estimation.

VII. Conclusion

This project evaluated two contrasting approaches to off-road traversability estimation on the RELLIS-3D dataset. The first being a specialized supervised model (FTF) and the second being a prompt-driven segmentation pipeline built entirely from vision foundation models (GSAM). By assessing both quantitative performance and qualitative behaviour, the study examined how well each method could support binary traversability prediction using only monocular RGB imagery.

The results showed that although FTF and GSAM achieved similar overall F1 and IoU scores, their error profiles and operational characteristics differ substantially. FTF produced smooth and expansive traversability maps with very high recall, but at the cost of low precision. GSAM, in contrast, delivered high-precision but conservative predictions shaped by prompt semantics. It also frequently missed large regions of terrain that were genuinely traversable. These behaviours illustrate a fundamental trade-off between task-specific supervised learning and flexible, prompt-based general-purpose segmentation.

The findings suggest that while foundation-model-based pipelines offer strong semantic prior knowledge and ease of deployment without retraining, they are not yet suitable replacements for specialized traversability models in autonomous off-road systems. However, their complementary properties indicate potential value as components within a hybrid pipeline.

Future work should explore extending this comparison to multimodal versions of FTF, automated prompt tuning for GSAM, and evaluations across diverse off-road datasets. Collectively, these directions may help bridge the gap between high-level semantic perception and terrain-aware affordance estimation for robust off-road autonomy.

References

[1] A. Angelova, L. Matthies, D. Helmick, and P. Perona, "Learning and prediction of slip from visual information," *Journal of Field Robotics*, vol. 24, no. 3, pp. 205–231, 2007.

[2] Anaconda Inc., "Miniconda — Minimal Installer for Conda," 2025. [Online]. Available: https://docs.conda.io/en/latest/miniconda.html

[3] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[4] BrownAssassin, *AIML-Research-Project*. GitHub Repository, 2025. [Online]. Available: https://github.com/BrownAssassin/AIML-Research-Project

[5] B. Cheng et al., "Masked-attention Mask Transformer for universal image segmentation," *arXiv:2112.01527*, 2022.

[6] A. Clark, "Pillow (PIL Fork) Documentation," 2025. [Online]. Available: https://python-pillow.org/

[7] Facebook Research, *Mask-Adapted CLIP*. GitHub Repository, 2023. [Online]. Available: https://github.com/facebookresearch/ov-seg

[8] J. Frey et al., "RoadRunner: Learning Traversability Estimation for Autonomous Off-road Driving," *arXiv:2402.19341*, 2024.

[9] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

[10] IDEA Research, *GroundingDINO*. GitHub Repository, 2023. [Online]. Available: https://github.com/IDEA-Research/GroundingDINO

[11] IDEA Research, *Grounded-Segment-Anything*. GitHub Repository, 2023. [Online]. Available: https://github.com/IDEA-Research/Grounded-Segment-Anything

[12] Y. Jeon, E. I. Son, and S.-W. Seo, "Follow the Footprints: Self-supervised Traversability Estimation for Off-road Vehicle Navigation based on Geometric and Visual Cues," *arXiv:2402.15363*, 2024.

[13] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D Dataset: Data, Benchmarks and Analysis," *arXiv:2011.12954*, 2022.
[14] A. Kirillov et al., "Segment Anything," *arXiv:2304.02643*, 2023.
[15] L. H. Li et al., "Grounded Language-Image Pre-training," *arXiv:2112.03857*, 2022.
[16] F. Liang et al., "Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP," *arXiv:2210.04150*, 2023.
[17] S. Liu et al., "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," *arXiv:2303.05499*, 2024.
[18] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous Robots*, vol. 18, pp. 81–102, 2005.
[19] M. Mattamala et al., "Wild Visual Navigation: Fast Traversability Learning via Pre-Trained Models and Online Self-Supervision," *arXiv:2404.07110*, 2024.
[20] Microsoft Corporation, "Windows Subsystem for Linux Documentation," 2025. [Online]. Available: https://learn.microsoft.com/windows/wsl/
[21] Microsoft, *GLIP*. GitHub Repository, 2022. [Online]. Available: https://github.com/microsoft/GLIP
[22] Z. Pan, P. Jiang, Y. Wang, C. Tu, A. G. Cohn, "Scribble-Supervised Semantic Segmentation by Uncertainty Reduction on Neural Representation and Self-Supervision on Neural Eigenspace," *arXiv:2102.09896*, 2021.
[23] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.
[24] Python Software Foundation, "argparse — Parser for command-line options, arguments and sub-commands," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/argparse.html
[25] Python Software Foundation, "json — JSON encoder and decoder," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/json.html
[26] Python Software Foundation, "os — Miscellaneous operating system interfaces," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/os.html
[27] Python Software Foundation, "pathlib — Object-oriented filesystem paths," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/pathlib.html
[28] Python Software Foundation, "sys — System-specific parameters and functions," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/sys.html
[29] Python Software Foundation, "typing — Support for type hints," *Python 3 Standard Library Documentation*, 2025. [Online]. Available: https://docs.python.org/3/library/typing.html
[30] PyTorch Team, "PyTorch: An open source machine learning framework," 2025. [Online]. Available: https://pytorch.org/
[31] PyYAML Developers, "PyYAML Documentation," 2025. [Online]. Available: https://pyyaml.org/
[32] A. Radford et al., "Learning transferable visual models from natural language supervision," *arXiv:2103.00020*, 2021.
[33] T. Ren et al., "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," *arXiv:2401.14159*, 2024.
[34] S. Sahu, A. Singh, K. Nambiar, S. Saripalli, and P. B. Sujit, "Any-Traverse: An off-road traversability framework with VLM and human operator in the loop," *arXiv:2506.16826*, 2025.
[35] C. Sevastopoulos and S. Konstantopoulos, "A Survey of Traversability Estimation for Mobile Robots," *arXiv:2204.10883*, 2022.
[36] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning Guided Convolutional Network for Depth Completion," *arXiv:1908.01238*, 2019.
[37] A. Valada, R. Mohan, and W. Burgard, "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation," *arXiv:1808.03833*, 2019.
[38] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

APPENDIX

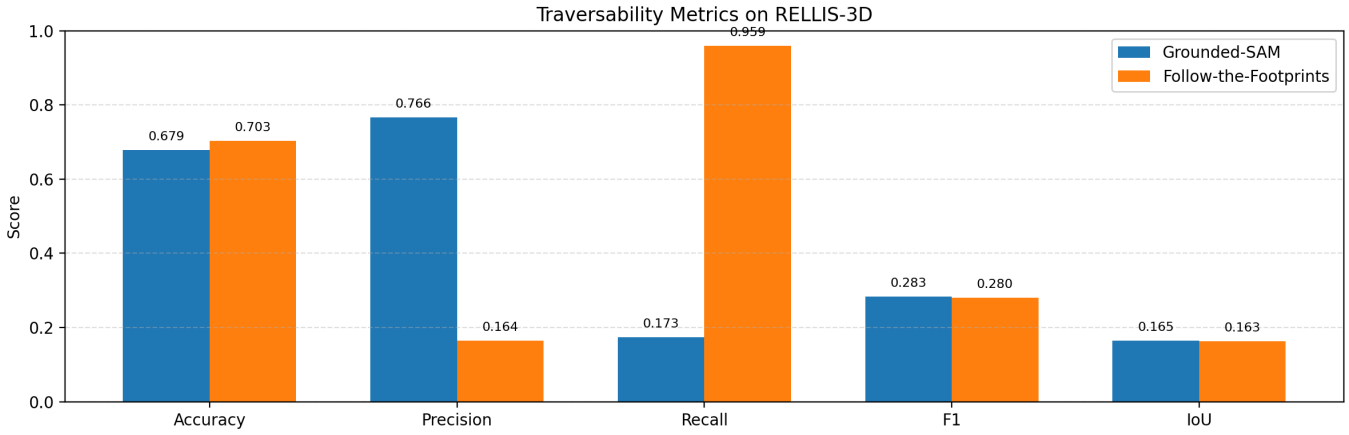

Fig. 5. A chart that makes the metrics more easy to interpret than the table.