Safety-Bounded Space Robot Navigation via Vision-Language Model Integration

Jimmy Tran, Chahat Deep Singh
Department of Robotics, University of Colorado Boulder

PRAISe Lab

Abstract

Planetary exploration robots operate under constraints that challenge modern autonomy: communication latency limits learning from human intervention, mass and power budgets restrict sensing and compute, and training data is scarce. Geometric perception is commonly used, but often fails to capture semantically meaningful hazards not well defined by geometry alone. Vision-language models (VLMs) offer a way to reason about such uncertainty, but their unpredictable failure modes limit use in safety-critical systems. We propose that robust autonomy is better achieved through architectural integration of geometric and semantic perception. We introduce a framework where the VLM acts as a conservative semantic safety advisor, augmenting a geometric planner with safety bounds. We evaluate three integration strategies: single-pass zero-shot detection, multi-stage decomposed reasoning with temporal filtering, and proposal verification with iterative refinement. Preliminary results demonstrate improved safety over geometric-only baseline in simulated navigation tasks.

Problem Motivation

Opportunity Rover stuck in Purgatory Dune. [1]

The 'status quo':
⇒ Geometric sensors (stereo, LiDAR, etc.) + Auxiliary sensors (thermal, infrared).

Problem #1:
⇒ Lack of semantic ability to understand hazards not well defined by available sensing modalities.

Initial 'solution':
⇒ Use VLMs! Stand in replacement for human intervention, and it replaces multi-step pipelines with single inference pass.

Problem #2:
⇒ VLM failure modes are not yet characterizable (no statistical guarantee on safety bounds or means of failure)...
⇒ Launch costs limit payload capabilities, including compute, so model size is also limited…

So what do we do?

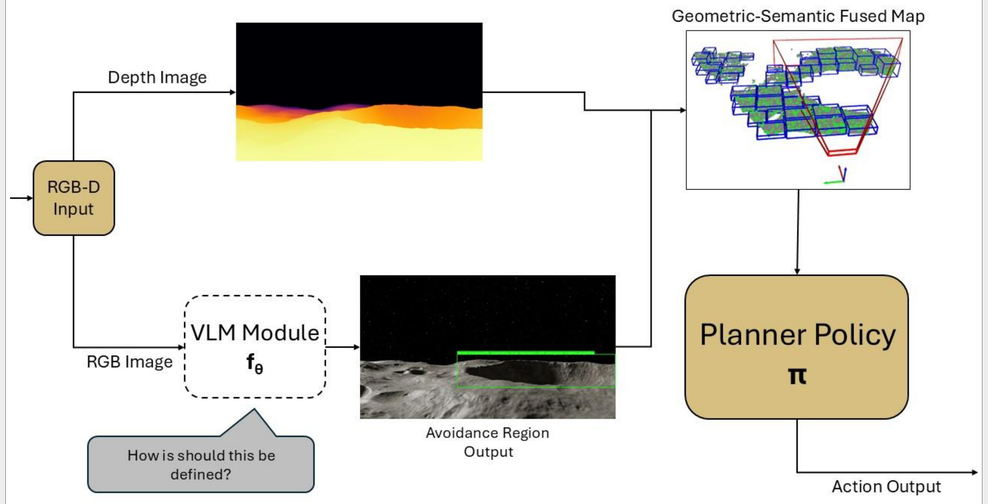
Operational setup: Let's say the following are true…

1) We are currently in a communications blackout with satellites communicating with ground stations
2) We do not have orbital maps of the area we are in (either due to lack of satellite imagery, operation in occluded regions)
3) We are in a state that is either currently unsafe, or has the potential to become unsafe in the future

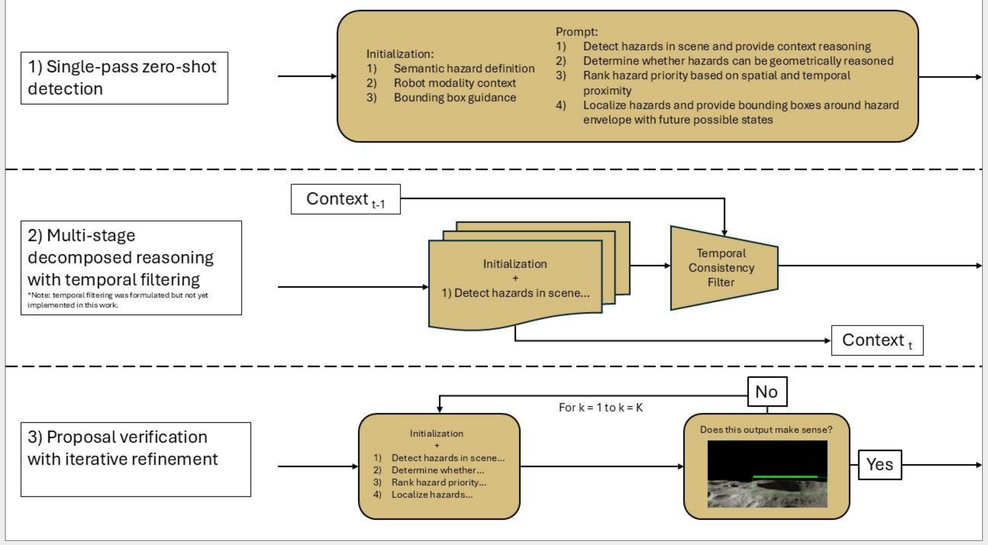
⇒ So inaction is not necessarily the 'safest' action…

Image showing damage to Curiosity Rover's wheels caused by sharp rock outcrops. [2]

System Diagram

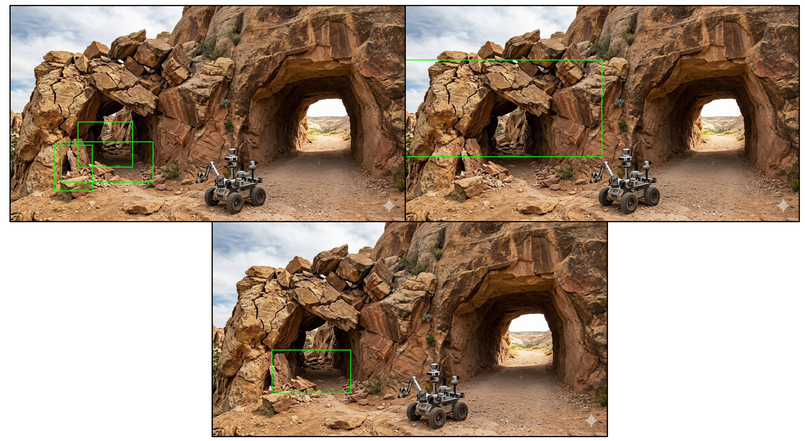


Architecture [f_θ] Design



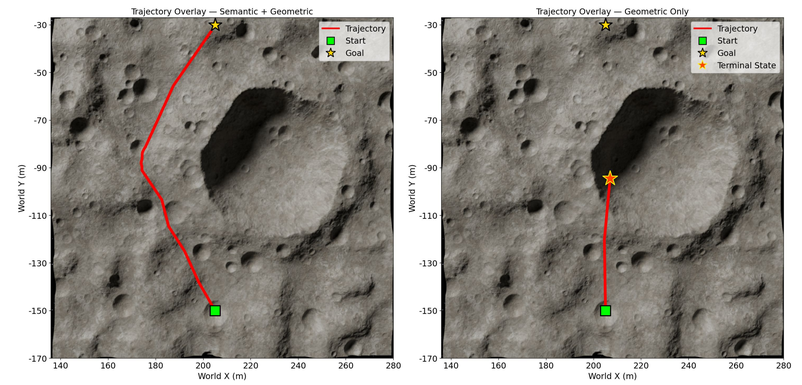
[1] NASA Jet Propulsion Laboratory, "Looking Back at 'Purgatory Dune' (PIA07999)," June 21, 2005.
[2] R. E. Arvidson *et al.*, "Relating geologic units and mobility system kinematics contributing to curiosity wheel damage at gale crater, mars," Journal of Terramechanics, vol. 73, pp. 73–93, 2017.
[3] Google, "Gemini: Multimodal generative AI," https://gemini.google. com/, accessed: March 2026.

Results



1) Single-pass zero-shot detection: capable of recognizing the pathway on the left as potentially unsafe, but also picks up on irrelevant elements in the scene
2) Multi-stage decomposed reasoning: demonstrates stronger reasoning focus, localizing a single bounding box that spans the hazardous region
3) Proposal verification with iterative refinement: Further refines localization of box to block only the potential traversal region of the rover, instead of the entire archway

Note: Numbering ⇒ Top → Down | Left → Right, test images generated with Gemini. [3]



Note #1: Crater was chosen as 'semantic hazard' in this Blender-simulated lunar rover scene for its relevance to Opportunity rover example (potential risk for wheel slip or sinkage in loose regolith near crater slopes).

Note #2: Example of single frame outputs can be seen in the system diagram.

Limitations and Future Work

Limitations:
1) Using general-purpose VLM (Qwen3-VL) for testing, instead of model adapted specifically for space robotics.
2) Minimal test dataset utilized (lack of real and relevant scenes to anticipated semantic safety hazards). This work largely focuses on proposal of safety systems architecture for VLM integration into space robotics applications.
3) Assumes ideal visibility conditions for RGB-D sensing.
4) Power consumption is loosened as a constraint. Overly conservative outputs could lead to inefficient power draw from idling commands (specifically for robot modalities that require constant power draw to operate, such as aerial robots).

Future Work:
1) Usage of space-adapted VLMs.
2) Selective semantic segmentation: using segmentation models to more accurately outlined identified regions in the scene. 'Selective' refers to the VLM's decision to segment elements of the scene (e.g. segmentation would make more sense in static hazard scenarios – such as rocky terrain – than for dynamic hazard scenarios – such as dust plumes).
3) Temporal consistency filtering implementations: filtering VLM results across time steps to ensure detections are consistent and not spurious.
4) VLM tool-calling procedures: offload tasks to external tools with bounded success rates.