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

Jimmy Tran and Chahat Deep Singh
University of Colorado, Boulder

***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. While geometric perception is commonly used for navigation, it often fails to capture semantically meaningful hazards that are not well defined by geometry alone. Satellite imagery can help manage growing uncertainty bounds by marking out regions known to be unsafe, but occluded environments do not have this support. Vision-language models (VLMs) offer a way to reason about such semantic uncertainty, but their unpredictable failure modes limit use in safety-critical systems. We propose that robust autonomy in space is better achieved through architectural integration of geometric and semantic perception, rather than relying solely on training better models. We introduce a framework in which a VLM acts as a conservative semantic safety advisor, augmenting a geometric planner with safety bounds such that uncertainty results in over-restriction rather than unsafe actions. 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 that these architectures improve the safety of a geometric-only baseline in simulated navigation tasks.**

I. Introduction

Robot navigation in unstructured environments typically relies on geometric sensors (stereo depth, LiDAR) paired with auxiliary sensors (thermal, infrared) to fill in for the perception gaps that geometry alone can not capture. This design is only effective when hazards can be accurately represented by the available sensing modalities. While LiDAR enables terrain traversability analysis, and thermal can mark out temperature-inoperable regions, these sensors lack the semantic ability to discern partially ambiguous decision boundaries. [1] Sensor limitations are further compounded in space, where launch costs scale exponentially with payload mass, limiting sensor availability. [2] In Earth-based settings, robots can request human help, [3] or deploy more powerful multi-step algorithms (image segmentation, feature detection, region classification, depth quantization, material analysis) to continually re-sense and re-plan [4]; these approaches are infeasible in space environments. VLMs offer a compelling alternative, replacing multi-step pipelines with a single inference pass that provides semantic hazard reasoning at reduced computational cost. While inaction seems safest under uncertainty, hazards can evolve temporally, rendering previously safe states unsafe and forcing autonomous action without human supervision.

'Semantic uncertainty' in space manifests in hazards such as loose regolith slopes with high slip or sinkage risk, fine dust particulates obscuring visibility for feature-based tracking, or operation near structurally compromised bodies– obstacles requiring logical inference beyond geometric sensing to classify. [5] Past examples include the Opportunity rover's five-week entrapment in Purgatory Dune due to wheel sinkage, [6] and the Curiosity rover's wheel damage caused by sharp rock outcrops. [7] While mission planners will typically have these regions marked out on satellite maps, [8] autonomy in space will eventually have to tackle unmapped areas where 'keep-out zones' are not well defined. [9] Given these constraints, how can we design intelligent, robust, and scalable autonomy? The focus for safety-critical systems is shifting from training better models towards identifying architectural gaps and integration structures that enable formal safety guarantees. [10] This work evaluates three VLM integration architectures that proactively prevent reaching irrecoverable states while maintaining two principles: (a) advisory-only guidance decoupled from direct control inputs, and (b) minimal resource requirements.

The contributions of this work are:

- An evaluation of three integration architectures for incorporating vision-language models into space robot navigation for semantic hazard detection.
- A formulation of semantic safety bounding as an advisory layer over geometric planning, ensuring that VLM uncertainty results in conservative over-restriction rather than unsafe actions while preserving baseline safety guarantees.

II. Problem Formulation

Consider a robot with state space $\mathcal{S}$ and action space $\mathcal{A}$. A trajectory τ is defined as a set of states and actions over time horizon T, $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$ where $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$. A simple geometry-only navigation policy is defined as:

$$a_t = \pi(M_t^{geom}, s_t, g), \quad M_t^{geom} = \Phi_g(z_t, s_t) \qquad (1)$$

Where M_t^{geom} encodes the geometric map at time t which the policy plans over, and g is some goal condition for which a heuristic could be defined. We define an operator Φ, that takes in 'sensor inputs' (depth values z_t for geometric case) paired with the state to construct a map representation. As previously mentioned, this policy is insufficient when geometric sensing does not fully represent traversal risk. We consequently define a new term M_t^{sem}, which encodes the semantic map representation of the environment, and gets

fused with the geometric map through a union operation:

$$a_t^* = \pi(M_t^{geom} \cup M_t^{sem}, s_t, g) \quad (2)$$

The formulation for the new term M_t^{sem} is defined as:

$$M_t^{sem} = \phi_s([\mathcal{B}_t, (z_{min}, z_{max})], s_t), \quad \mathcal{B}_t, C_t = f_\theta(I_t) \quad (3)$$

The 'sensor inputs' in the semantic case include $\mathcal{B}_t$ (the 2D image space bounding boxes outputted by the VLM), and (z_{min}, z_{max}), the actionable depth range provided to project $\mathcal{B}_t$ into M_t^{sem} (since the VLM has no discrete depth rectification). We define f_θ as an operator that involves querying the VLM, with the input being the RGB image I_t, and the outputs being $\mathcal{B}_t$ and corresponding context labels C_t. The key problem that this work aims to define is how f_θ should be integrated into existing frameworks in a way that is spatially grounded, robust, and computationally practical.

III. Methodology

A. Geometric Layer Design

The geometric layer constructs an obstacle map from depth observations using standard point cloud processing techniques (downsampling, voxelization, and clustering to form boxed regions). This map defines the baseline planning space.

B. Semantic Layer Design

A vision-language model takes in RGB inputs and outputs spatial bounds over the robot action space. [11] We test three different architectures for integrating f_θ. These are formulated in Algorithms 1, 2, and 3.

Algorithm 1 Single Zero-Shot Prompt: standalone zero-shot prompts at every VLM query call. The prompt contains information on hazard, spatial, and output formatting guidance.

Require: Depth readings z_t, RGB image I_t, state s_t
Ensure: Action a_t

Construct geometric map from depth readings:
  $M_t^{\text{geom}} = \Phi_g(z_t, s_t)$
Single-pass query to VLM for semantic hazard 2D bounding boxes and context labels:
  $\mathcal{B}_t, C_t = f_\theta(I_t)$
Project 2D bounding boxes to 3D semantic map:
  $M_t^{\text{sem}} = \Phi_s([\mathcal{B}_t, (z_{min}, z_{max})], s_t)$
Fuse geometric and semantic maps:
  $M_t^{\text{fused}} = M_t^{\text{geom}} \cup M_t^{\text{sem}}$
Plan over fused map:
  $a_t = \pi(M_t^{\text{fused}}, s_t, g)$

In all architectures, the model received an initial safety context prompt with robot specifications (size and modality) and hazard determination rules across robot modalities (e.g. aerial robots are generally indifferent to ground hazards). The final 2D bounding box outputs are then projected into 3D space over actionable depth (z_{min}, z_{max}). In this case, 'actionable depth' refers to setting z_{min} and z_{max} to the geometric depth sensor's sensing range for map alignment, but also to enforce a 'proactive' planning intent–minimizing the risk of reaching irrecoverable states early on.

Algorithm 2 Multi-Stage Decomposed Reasoning with Temporal Filter: prompting is broken down across decision boundaries, each contextualized by the preceding stage's outputs, and filtered against each other over temporal instances.

Require: Depth readings z_t, RGB image I_t, state s_t, previous semantic map M_{t-1}^{sem}
Ensure: Action a_t

Construct geometric map from depth readings:
  $M_t^{\text{geom}} = \Phi_g(z_t, s_t)$
Triage prompt to VLM for hazard category presence:
  $\mathcal{T}_t = f_\theta^{\text{triage}}(I_t)$
Localize hazard regions for each flagged category:
  $\mathcal{H}_t = f_\theta^{\text{prox}}(I_t, \mathcal{T}_t)$
Extract spatial bounds from hypothesis regions:
  $\mathcal{B}_t, C_t = f_\theta^{\text{bbox}}(I_t, \mathcal{H}_t)$
Project raw detections to 3D semantic map:
  $\hat{M}_t^{\text{sem}} = \Phi_s([\mathcal{B}_t, (z_{min}, z_{max})], s_t)$
Filter detections using temporal consistency and previously determined context:
  $M_t^{\text{sem}} = g_{\text{filter}}(\hat{M}_t^{\text{sem}}, M_{t-1}^{\text{sem}}, C_{t-1})$
Fuse maps:
  $M_t^{\text{fused}} = M_t^{\text{geom}} \cup M_t^{\text{sem}}$
Plan:
  $a_t = \pi(M_t^{\text{fused}}, s_t, g)$

Algorithm 3 Proposal Verification with Iterative Refinement: this approach builds off of Alg. 1, adding a secondary stage that passes the image with overlayed 2D bounding boxes back to the model (paired with context) for self-validation. Note: $\hat{\chi}$ represents proposal for value of χ.

Require: Depth readings z_t, RGB image I_t, state s_t, max validation iterations K
Ensure: Action a_t

Construct geometric map from depth readings:
  $M_t^{\text{geom}} = \Phi_g(z_t, s_t)$
Forward-pass query:
  $\hat{\mathcal{B}}_t, \hat{C}_t = f_\theta(I_t)$
  $\mathcal{B}_t \leftarrow \hat{\mathcal{B}}_t, C_t \leftarrow \hat{C}_t$ ▷ Initialize Storage
for $k = 1$ to K **do**
  Overlay bounding boxes on image:
    $I_t^{\text{overlay}} = \text{draw}(I_t, \hat{\mathcal{B}}_t)$
  Verification query:
    $v_k, \mathcal{B}_t^{(k)}, C_t^{(k)} = f_\theta^{\text{verify}}(I_t^{\text{overlay}}, \hat{C}_t)$
  if $v_k =$ `approved` **then**
    $\mathcal{B}_t \leftarrow \mathcal{B}_t^{(k)}$, $C_t \leftarrow C_t^{(k)}$ ▷ Update Best
    break
  else
    $\hat{\mathcal{B}}_t \leftarrow \mathcal{B}_t^{(k)}$, $\hat{C}_t \leftarrow C_t^{(k)}$ ▷ Update Latest
  end if
end for
Project to 3D semantic map:
  $M_t^{\text{sem}} = \Phi_s([\mathcal{B}_t, (z_{min}, z_{max})], s_t)$
Fuse maps:
  $M_t^{\text{fused}} = M_t^{\text{geom}} \cup M_t^{\text{sem}}$
Plan:
  $a_t = \pi(M_t^{\text{fused}}, s_t, g)$

C. Experimental Setup

We use the Qwen3-VL 8B instruct model [12] and evaluate on: (1) single-frame synthetic scenes generated using Gemini, a commercial multi-modal generative model with text-to-image capabilities, [13] and (2) a full trajectory in a Blender-simulated lunar rover scene, comparing Algorithm 1 against a geometric-only baseline. The lunar scene was chosen for its relevance to the Opportunity rover's Purgatory Dune entrapment. The substrate near and around a crater is both sloped and structurally ambiguous, posing potential slip or sinkage risks. The planner policy used was a custom collision-free sampling-based approach with a weighted distance-to-goal heuristic.

IV. Results

(a) *Single zero-shot prompt* (b) *Multi-stage reasoning*

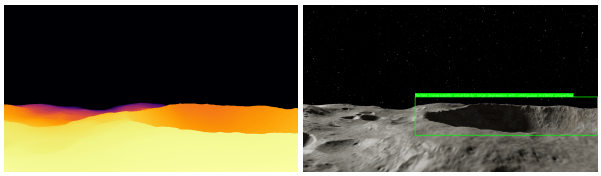
(c) *Proposal verification*

Fig. 1: *Single instance testing for a synthetic scene with a structurally compromised overhang.*

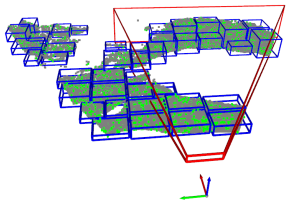
1) Single Zero-Shot Prompting: This method avoided high-uncertainty regions but exhibited selective attention failures, often over-analyzing certain features in the scene while ignoring others. Additionally, the bounding box outputs were not always well-defined (insufficient hazard-extent coverage, boxes 'bleeding' into safe zones, overlapping boxes with similar context labels–refer to Fig. 1(a)), demonstrating weak spatial understanding within the image. The current working hypothesis is that this effect is due to the model having to simultaneously handle multiple decision boundaries in a single prompt (hazard reasoning, spatial localization, pixel-wise inference), increasing cognitive load and reducing performance on the individual tasks themselves.

2) Multi-stage Decomposed Reasoning: Sub-task decomposition was able to alleviate some of the cognitive load experienced in single zero-shot prompting by splitting the context window across multiple prompts. [14] Each stage was able to focus on handling its own respective task, with the outputs of each stage feeding into the next. This architecture allowed for stronger grounding across the individual tasks, represented by more concisely defined bounding box outputs as seen in Fig. 1(b). The model successfully identified the single dominant hazard in the scene–the structurally compromised overhang–and did not bound elements that largely will not interfere with the robot's operation (e.g. the small cluster of rocks near the base of the left overhang entrance).

3) Proposal Verification with Iterative Refinement: The third architecture introduced a validation layer that enables the VLM to refine its own predictions. By re-evaluating bounding boxes overlaid on the input image, the model can correct spatial inconsistencies from the initial forward pass. This is demonstrated in Fig. 1(c), where the overlapping bounding boxes of Fig. 1(a) were combined, and the bounded region of Fig. 1(b) was shrunk to only bound the robot's potential path instead of the entire archway.



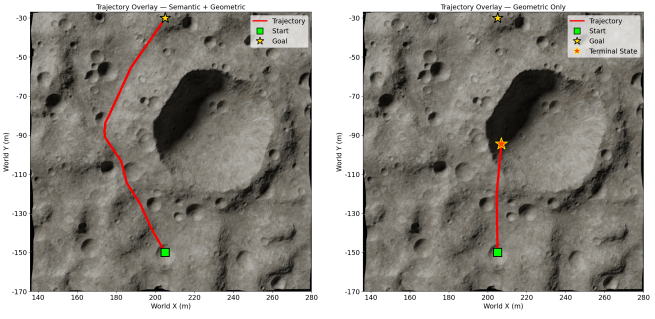
(a) *Depth Map* (b) *Semantic RGB Overlay*



(c) *Semantic-Geometric Map*

Fig. 2: *Single frame outputs for full trajectory testing in lunar rover scene.*

4) Full-trajectory Testing: Even with the limited capabilities of Algorithm 1, this method demonstrated its potential capabilities in a synthetic space-robotics navigation scenario. As shown in Fig. 2, the model correctly identified the crater as the region of high safety uncertainty. To reduce computational load, the VLM was queried every five time steps, assuming limited scene variation between queries. Despite this sparse querying, the semantic layer remained effective in guiding safe navigation. The plots of the full trajectory with and without the semantic-safety bounding layer can be viewed in Fig. 3.



(a) *Semantic + Geometric* (b) *Geometric only*

Fig. 3: *Top-down trajectory overlay comparing semantic-geometric and geometric only planners. The geometric-only baseline fails to detect the crater as an 'obstacle.'*

The full testing files can be viewed at this link.

V. Limitations

This work depends on the choice of VLM and the underlying geometric planning framework. As such, performance is influenced by model capability and perception quality. As a study on safety systems architecture, the goal of the paper was to improve the baseline safety, as well as steer a path toward the future integration of VLM agents into mission-critical systems in a way where formal safety guarantees can be defined, and failure modes can be characterized. With the steady advancement of more capable lightweight vision foundation models, as well as models that are adapted specifically to space environments, [15] [16] the detection accuracies and overall safety enhancements demonstrated in this work will be inherently improved. Additionally, some operational assumptions were made:

- For RGB-D systems, image inputs to the VLM can be affected by unavoidable lighting conditions (darkness, glare; conditions such as dust can be relatively planned around). Multi-modal models that can reason over 'common-sense' cues present in other sensor types could be a way forward.
- For power consumption, ground robots can afford over-conservative outputs from the VLM (i.e. idling commands), whereas aerial robots will likely not be able to due to limited battery capacity. Mission design will have to account for this, or employ strategies that can plan around such constraints (e.g. using an aerial robot primarily to scout out paths in front of a rover).

The current integration architectures also leave room for further augmentation. Future avenues of work could include:

- Introduction of a semantic segmentation layer after extracting hazard bounding boxes to define more spatially accurate regions of interest.
- Context labels are generated in this work, but are not yet fully used; the consistency of labels across frames can help reject spurious detections and improve robustness in temporal filtering and linear temporal logic formulations.
- Offloading tasks spawned by the VLM to external tools via tool-calling procedures.

References

[1] M. Elnoor, K. Weerakoon, G. Seneviratne, R. Xian, T. Guan, M. K. M. Jaffar, V. Rajagopal, and D. Manocha, "Robot navigation using physically grounded vision-language models in outdoor environments," 2024. [Online]. Available: https://arxiv.org/abs/2409.20445

[2] B. A. y Arcas, T. Beals, M. Biggs, J. V. Bloom, T. Fischbacher, K. Gromov, U. Köster, R. Pravahan, and J. Manyika, "Towards a future space-based, highly scalable AI infrastructure system design," 2025. [Online]. Available: https://arxiv.org/abs/2511.19468

[3] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar, "Robots that ask for help: Uncertainty alignment for large language model planners," 2023. [Online]. Available: https://arxiv.org/abs/2307.01928

[4] S. K. Ravipati, E. Latif, R. Parasuraman, and S. M. Bhandarkar, "Object-oriented material classification and 3d clustering for improved semantic perception and mapping in mobile robots," 2024. [Online]. Available: https://arxiv.org/abs/2407.06077

[5] M. Oh, C. Kim, S.-W. Seo, and S.-W. Kim, "Language as cost: Proactive hazard mapping using vlm for robot navigation," 2025. [Online]. Available: https://arxiv.org/abs/2508.03138

[6] R. E. Arvidson, J. W. Ashley, J. F. Bell, M. Chojnacki, J. Cohen, T. E. Economou, W. H. Farrand, R. Fergason, I. Fleischer, P. Geissler, R. Gellert, M. P. Golombek, J. P. Grotzinger, E. A. Guinness, R. M. Haberle, K. E. Herkenhoff, J. A. Herman, K. D. Iagnemma, B. L. Jolliff, J. R. Johnson, G. Klingelhöfer, A. H. Knoll, A. T. Knudson, R. Li, S. M. McLennan, D. W. Mittlefehldt, R. V. Morris, T. J. Parker, M. S. Rice, C. Schröder, L. A. Soderblom, S. W. Squyres, R. J. Sullivan, and M. J. Wolff, "Opportunity mars rover mission: Overview and selected results from purgatory ripple to traverses to endeavour crater," *Journal of Geophysical Research: Planets*, vol. 116, no. E7, p. E00F15, 2011.

[7] R. E. Arvidson, P. DeGrosse Jr, J. P. Grotzinger, M. C. Heverly, J. Shechet, S. J. Moreland, M. A. Newby, N. Stein, A. C. Steffy, F. Zhou, A. M. Zastrow, A. R. Vasavada, A. A. Fraeman, and E. K. Stilly, "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.

[8] V. Verma, J. Nash, L. Saldyt, Q. Dwight, H. Wang, S. Myint, J. Biesiadecki, M. Maimone, A. Tumbar, A. Ansar, G. Kubiak, and R. Hogg, "Enabling long precise drives for the perseverance mars rover via onboard global localization," 03 2024, pp. 1–18.

[9] R. Thakker, A. Patnaik, V. Kurtz, J. Frey, J. Becktor, S. Moon, R. Royce, M. Kaufmann, G. Georgakis, P. Roth, J. Burdick, M. Hutter, and S. Khattak, "Risk-guided diffusion: Toward deploying robot foundation models in space, where failure is not an option," 2025. [Online]. Available: https://arxiv.org/abs/2506.17601

[10] Z. Ravichandran, A. Robey, V. Kumar, G. J. Pappas, and H. Hassani, "Safety guardrails for llm-enabled robots," 2026. [Online]. Available: https://arxiv.org/abs/2503.07885

[11] M. Tölle, T. Gruner, D. Palenicek, T. Schneider, J. Günster, J. Watson, D. Tateo, P. Liu, and J. Peters, "Towards safe robot foundation models using inductive biases," 2025. [Online]. Available: https://arxiv.org/abs/2505.10219

[12] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2511.21631

[13] Google, "Gemini: Multimodal generative AI," https://gemini.google.com/, accessed: March 2026.

[14] S. Zhu, D. Li, L. Mou, Y. Liu, N. Xu, and H. Zhao, "Saro: Space-aware robot system for terrain crossing via vision-language model," 2025. [Online]. Available: https://arxiv.org/abs/2407.16412

[15] M. Foutter, P. Bhoj, R. Sinha, A. Elhafsi, S. Banerjee, C. Agia, J. Kruger, T. Guffanti, D. Gammelli, S. D'Amico, and M. Pavone, "Adapting a foundation model for space-based tasks," 08 2024.

[16] M. Foutter, D. Gammelli, J. Kruger, E. Foss, P. Bhoj, T. Guffanti, S. D'Amico, and M. Pavone, "Space-llava: A vision-language model adapted to extraterrestrial applications," 03 2025, pp. 1–23.