Adapting a Foundation Model for Space-based Tasks

Matthew Foutter[1], Praneet Bhoj[2], Rohan Sinha[3], Amine Elhafsi[3], Somrita Banerjee[3], Christopher Agia[2], Justin Kruger[3], Tommaso Guffanti[3], Daniele Gammelli[3], Simone D'Amico[3] and Marco Pavone[3,4]

***Abstract*—Foundation models, e.g., large language models, possess attributes of intelligence [23] which offer promise to endow a robot with the contextual understanding necessary to navigate complex, unstructured tasks in the wild. In the future of space robotics, we see three core challenges which motivate the use of a foundation model adapted to space-based applications: 1) *Scalability* of ground-in-the-loop operations; 2) *Generalizing* prior knowledge to novel environments; and 3) *Multi-modality* in tasks and sensor data. Therefore, as a first-step towards building a foundation model for space-based applications, we automatically label the AI4Mars dataset [22] to curate a language annotated dataset of visual-question-answer tuples. We fine-tune a pretrained LLaVA checkpoint on this dataset to endow a vision-language model with the ability to perform spatial reasoning and navigation on Mars' surface. In this work, we demonstrate that 1) existing vision-language models are deficient visual reasoners in space-based applications, and 2) fine-tuning a vision-language model on extraterrestrial data significantly improves the quality of responses even with a limited training dataset of only a few thousand samples.**

I. Introduction

Advancements in the development of internet-scale machine learning models trained through self-supervision on a corpus of human knowledge, i.e., Foundation Models (FMs) [3], provide an opportunity to automate complex decision making and reasoning tasks transcribed through language, video, and speech. State-of-the-art (SoTA) large language models (LLMs) already display strong commonsense reasoning and understanding capabilities that, for example, enable them to score in the upper quartile on a variety of standardized exams [17]. These commonsense reasoning capabilities make the use of FMs attractive in space robotics, satellite operations, and other space-related domains, where they show potential to mitigate core challenges such as: 1) *Scalability* of ground-in-the-loop operations; 2) *Generalizing* prior knowledge to novel environments; and 3) *Multi-modality* in tasks and sensor data.

Accordingly, in this paper, we conduct a preliminary investigation of the application of pretrained multi-modal FMs to the space domain. As a first step towards developing *a space foundation model*, we focus on a space robotics application in which a rover navigates a planetary environment (Fig. 1). We programmatically generate language annotations on the AI4Mars image dataset [22] to adapt and evaluate vision-language models (VLMs) across several spatial reasoning and navigation tasks. These tasks are inspired by the detailed sensory reasoning necessary to, e.g., identify sites of scientific interest or validate candidate motion plans. Our evaluations demonstrate that 1) existing VLMs are deficient visual reasoners in space-based applications, and 2) fine-tuning a VLM on our programmatically generated tasks significantly improves the quality of the VLM outputs across all the visual reasoning tasks, even when the training dataset only consists of few thousand images that are reused for different QA pairs. Subsequently, we propose pathways for extending the approaches and concepts discussed in this paper to orbital in-space applications. As such, this paper represents a promising exploration towards the development of generalist models for space.

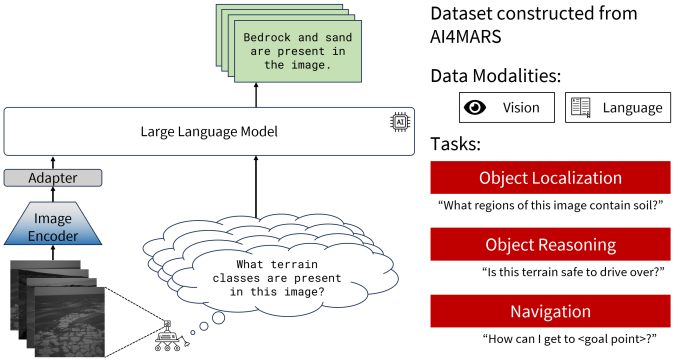


Fig. 1: We present Space-LLaVA, initialized from a pretrained LLaVA model [13] and fine-tuned with domain-specific data, namely Martian imagery [22]. Space-LLaVA can reason about localization and navigation tasks in planetary environments.

II. Related Work

Vision-Language Models: Recent advances in natural language and image processing have enabled the development of large-scale VLMs trained on internet-scale data. Early work develops an understanding of vision and language by using vision and text encoders [19], while VLMs build atop a language model to allow for open-ended visual reasoning such as Visual-Question-Answering (VQA) [11, 5, 14, 7]. In this work, we investigate LLaVA v1.5 [13] as the base model for fine-tuning given it is SoTA among open-source models on standard VQA benchmarks [1, 10].

Foundation Models in Robotics: Prior work incorporates FMs at various levels of the autonomy stack, ranging from planning/decision making [24, 12] to semantic [8] and visual [21] reasoning. There is also emerging work on adapting FMs to space-based applications. SpaceTransformers [2] fine-tunes variations of BERT [6] on a corpora of systems engineering texts and an augmented mission standards dataset to recognize space mission requirements, while Rodriguez-Fernandez et al. [20] leverages GPT-3.5 [4] as the policy backbone for language-based autonomous satellite operations. We extend these works by incorporating both vision and language into a shared representation for enhanced reasoning.

Large-scale Dataset Curation: Related work in large-scale data collection includes Gao et al. [9] that develops a dataset of objects annotated with physical properties for image classification. We aim to extend this work by developing

[1]Dept. of Mechanical Engineering, Stanford University. [2]Dept. of Computer Science, Stanford University. [3]Dept. of Aeronautics and Astronautics, Stanford University. [4]NVIDIA. Contact: {mfoutter, praneet, rhnsinha, amine, somrita, cagia, jjkruger, tommaso, gammelli, damicos, pavone}@stanford.edu.

a dataset for *visual reasoning* for terrain-aware navigation on Mars. Marcu et al. [16] and Ma et al. [15] curate a large-scale autonomous driving VQA benchmark to enable perception, prediction, and planning; however, they require language annotations from human operators, which is likely incompatible with long-horizon data collection at scale. Consequently, our work is distinguished from existing work by programmatically generating a dataset of language annotations for visual reasoning to fine-tune a VLM in the context of space robotics.

III. Architecting Space-LLaVA

In this work, we adapt a FM to two space-based applications using the AI4Mars dataset which encompasses 35k images with crowd-sourced semantic segmentation masks of Mars' terrain gathered from the Curiosity, Opportunity, and Spirit rovers. A representative example of raw terrain from the AI4Mars dataset and its associated semantic masks for each terrain class is provided in Fig. 2.

Given the AI4Mars dataset, we require a high-quality and scalable technique to generate QA pairs in natural language to endow an open-source VLM with the ability to perform spatial reasoning and high-level motion planning on withheld terrain.

We ground a VLM in the visual and semantic features of Mars' terrain by fine-tuning LLaVA v1.5 13B [13] on our augmented dataset with the standard auto-regressive language modeling loss. Suppose we curate a dataset $\mathcal{D}=\{(\mathbf{I}^{(i)},\mathbf{Q}^{(i)},\mathbf{A}^{(i)})\}_{i=1}^{n}$ consisting of n image $\mathbf{I}^{(i)}\in\mathbb{R}^{h\times w\times 3}$, question $\mathbf{Q}^{(i)}\in\mathbb{R}^{T_Q}$, and answer $\mathbf{A}^{(i)}\in\mathbb{R}^{T_A}$, tuples where T_Q and T_A denote the maximum tokenized question and answer sequence length, respectively, with padding. We fine-tune LLaVA by freezing certain parameters in the model, e.g., only fine-tuning the language backbone, to optimize the objective

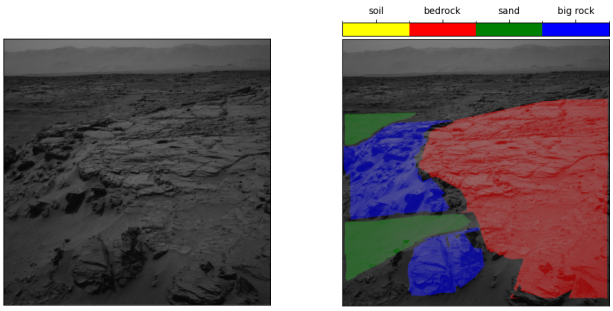
$$\min_{\hat{\theta}\subseteq\Theta} L(\hat{\theta}\,|\,\mathcal{D}), \tag{1}$$

where we construct $L(\hat{\theta}\,|\,\mathcal{D})$ as the negative log-likelihood loss on token generation assuming samples are independent and identically distributed and using the chain rule factorization for auto-regressive generation. More formally, we define:

$$L(\hat{\theta}\,|\,\mathcal{D})=-\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\log p_{\hat{\theta}}(x_{t+1}^{(i)}\,|\,\mathbf{I}^{(i)},\mathbf{Q}^{(i)},\mathbf{A}_{1:t}^{(i)}), \tag{2}$$

where each term in the summation represents the log-likelihood, under the model's current weights $\hat{\theta}$, to predict the ground-truth next text token in the answer sequence $\mathbf{A}^{(i)}=\{x_1^{(i)},x_2^{(i)},...,x_T^{(i)}\}$ conditioned on the sample's visual input, associated question and the full answer sequence preceding $x_{t+1}^{(i)}$. Here, $\hat{\theta}\subseteq\Theta$ indicates that the unfrozen weights are a subset of the model's weights Θ. We evaluate the quality of the fine-tuned model's responses in comparison to a base model by prompting GPT-4 to choose the preferable response conditioned on the ground-truth answer for a particular question. A template of the prompt we provide to GPT-4 is provided in Appendix VI-A with further discussion on the prompt's construction.

That is, through fine-tuning a VLM on semantically annotated terrain from the AI4Mars dataset, we measure whether the fine-tuned model outperforms SoTA VLMs on the same task without adaptation, i.e., zero-shot.



(a) Raw MSL NAVCAM capture.

(b) MSL NAVCAM crowd-sourced terrain labels.

Fig. 2: The AI4Mars dataset [22] provides crowd-sourced annotations for "soil", "bedrock", "sand", "big rock" on Mars.

IV. Data Generation Pipeline

In order to adapt a FM to the unique visual and semantic features on Mars' surface, we develop a language QA generation pipeline on AI4Mars' semantic segmentation masks for spatial reasoning and high-level motion planning. Explicitly, we choose to curate a dataset of spatial reasoning and navigation samples since these two tasks together require semantic analysis on an extraterrestrial environment and leverage an existing, high-quality dataset. We first present our programmatic solution to curate spatial reasoning QA pairs after which we present a similar methodology to curate high-level motion planning QAs.

A. Spatial Reasoning Dataset

We translate semantic segmentation masks into QA pairs requiring spatial reasoning through two programmatic measures of position in an image. First, we process each terrain's segmentation mask using KMeans clustering with K = 1 to identify a surrogate for the centroid. While choosing K = 1 does represent a strong inductive bias in clustering, we have noticed that dominant collections of terrain tend to naturally cluster together on Mars' surface, i.e., rarely is it the case two large instances of soil or rocks are isolated on opposite ends of an image, in which case K = 1 represents a reasonable approximation of the terrain's general position. Further, we mitigate the impact of outliers in terrain classification, i.e., small patches of classified terrain distant from the dominant patch(s), through multiple random initializations for clustering and accept the cluster center with the lowest total variance across all runs.

Also, we develop a second measure of each class' position by dividing the image into a 3x3 grid and measuring the population, or number of pixels, for each class in each grid. This simple measure of population helps formalize a notion of terrain density. A weakness of KMeans clustering is that we lose global information on the spread of terrain in an image, e.g., does 75% of soil or only 30% of soil in view exist in the bottom right corner of the image? Separately, we also enclose each semantic mask with bounding box annotations. Therefore, we supplement each cluster center with grid population counts and bounding boxes to provide a more holistic measure of terrain position.

We use these three measures of a terrain's position to programmatically generate QA pairs to elicit spatial reasoning based on the AI4Mars dataset. That is, we develop a set of template prompts for seven styles of QA pairs: terrain description; terrain localization; multi-instance terrain localization; relative terrain localization; terrain coverage;

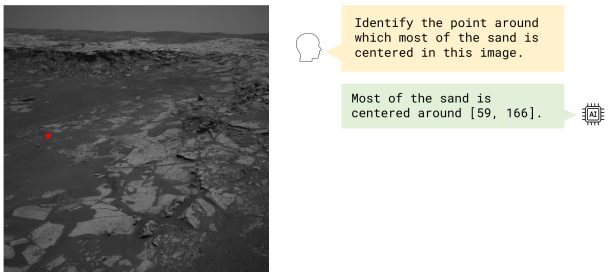


Fig. 3: Automatically generated terrain localization VQA using sand's true cluster center scaled to LLaVA's image resolution of (336×336); (0, 0) corresponds to the top left corner. A red star marks sand's cluster center, which is not available at training.

terrain size; and terrain traversability. Each question type is templated by a terrain's cluster position, bounding box and/or grid populations. For example, as shown in Fig. 3, to instantiate a question on sand localization we use the terrain's centroid.

Within each question type, we use GPT-4 to curate between 3-6 variations of the question's wording to discourage overfitting to the particular prose used when the question is asked.

B. High-level Motion Planning Dataset

In order to curate high-level motion plans that are cognizant of Mars' terrain, we curate a sequence of cardinal direction maneuvers, e.g., up, down, left, or right, connecting a start and goal point through a small grid over each AI4Mars sample. First, we discretize the image into a 5x5 grid and classify each grid into a terrain class according to majority representation. Then, we can mask out a particular class, e.g., large rocks, and if a feasible path to the goal exists, we use the $A^\star$ search algorithm to plan a path from the start to the goal. We choose the starting point as the lowest point along the center column of the image which is not occupied by 1) the rover itself and 2) the class we choose to mask from the image. For each sample, assuming a feasible path exists, we choose multiple path endpoints among the unmasked grids.

For each image with an annotated mask to identify the rover, i.e., the MSL dataset within AI4Mars, we ask 2-3 question types for every start and end point pair: 1) whether a feasible path to the goal exists, 2) whether a candidate path overlaid on the image is both feasible and reaches the desired goal and 3) only if a feasible path exists, we request a path in natural language to the goal. An example for each question type is presented in Appendix VI-B.

V. Experiments and Discussion

Having outlined our approach to data collection, we fine-tune LLaVA on our augmented planetary dataset, ablate different training configurations, and compare the model's performance to SoTA FMs applied zero-shot to space-based applications. Explicitly, we investigate whether adaptation is necessary for SoTA VLMs in an extraterrestrial environment, and if so, to what degree components of the model must be adapted. Then, we discuss future work to extend *a space foundation model* to applications in orbital space.

A. Experimental Setup

Recall from Fig. 1 that the LLaVA model is comprised of three components: a vision encoder (VE), a multi-modal adapter (MMA), and a language model (LM). In this work, we experiment with training four combinations of LLaVA's components: (1) training the vision encoder and multi-modal adapter together; (2) training only the multi-modal adapter; (3) training only the language model backbone; and (4) training all components of the model. A complete description of the experimental details is available in Appendix VI-C.

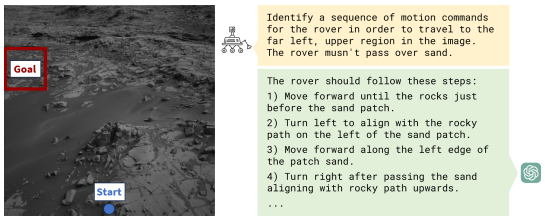
Frozen			Response Preference	
VE	MMA	LM	Ours vs. LLaVA	Ours vs. GPT-4o
×	×	✓	73.7%	69.2%
✓	×	✓	83.3%	78.8%
✓	✓	×	81.6%	77.9%
×	×	×	87.2%	85.4%

TABLE I: Response preference for each fine-tuned configuration against the base LLaVA model and GPT-4o.

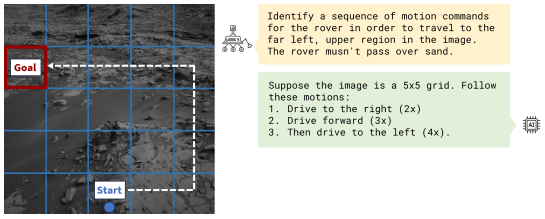
B. Experimental Results

The results of fine-tuning each configuration using our augmented AI4Mars dataset in comparison to zero-shot LLaVA and GPT-4o are presented in Table I. Further, we provide an example generation from Space-LLaVA and a qualitative comparison to GPT-4o in Fig. 4. Based on these results, it is immediately apparent that SoTA VLMs out-of-the-box are ill-equipped to process the novel semantic features on Mars likely due to a visual domain gap. Indeed, SoTA VLMs produce inferior content relative to an adapted model in Table I. For example, in Fig. 4a, GPT-4o hallucinates a path of bedrock on the left and leads the rover into a smooth sand patch in pursuit of the goal; However, in Fig. 4b, Space-LLaVA suggests a more favorable plan which leads the rover along the bedrock in view and thereby reaches the goal without exposure to sand.

We find that jointly training the language model and visual components provides the largest benefit for spatial reasoning tasks. As shown in Table I, fine-tuning the language and vision components in concert produces the best-performing model at just over 87% and 85% response preference to LLaVA zero-shot and GPT-4o, respectively, as may be expected given



(a) GPT-4o: Terrain-aware navigation on Mars' surface.



(b) SpaceLLaVA: Terrain-aware navigation on Mars' surface.

Fig. 4: GPT-4o and Space-LLaVA perform navigation on Mars' surface. The 5x5 grid, selected path, start and end grid point are highlighted for visual clarity.

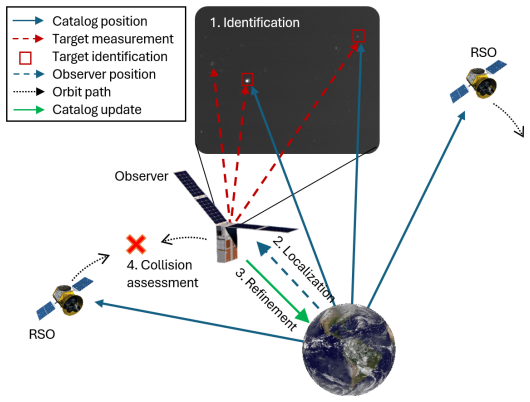


Fig. 5: A notional illustration of the FALCON problem and its four core tasks: target identification, observer positioning, SSA catalog refinement, and RSO collision assessments.

the full flexibility of the model to adapt to Mars' semantic features. In contrast, the configuration with a frozen language model demonstrates the worst relative performance against both zero-shot models. In particular, we posit that fine-tuning the language model significantly improves adaptation given that our terrain classification and navigation tasks require the VLM to perform *fine-grained semantic reasoning*, whereas the VLM from Gao et al. [9], which only trains InstructBLIP's equivalent to LLaVA's multi-modal adapter, is fine-tuned to perform the simpler task of image classification.

One of the most consistent cases in which the fine-tuned model is preferred to the base model is in the task of navigation. Inaccuracies in the baseline model's spatial reasoning and terrain classification manifest as an inability to plan paths around the terrain types to be avoided as shown in Fig. 4. On the other hand, the most common failure mode for our fine-tuned model appears to be image QA pairs requiring the model to make fine-grained distinctions between soil and sand terrain classes. Soil and sand do often have a very similar appearance in the AI4Mars dataset, especially given that the images are grayscale. With the lack of color features in these images, the model will likely need to undergo additional training to develop an understanding of grayscaled imagery, and capture nuanced differences between soil and sand features.

C. Future Applications: Orbital Space

Extending the concept of *a space foundation model* to orbital operations presents additional challenges, because the characteristics of orbital scenarios and data sources are even further removed from typical terrestrial applications. Nevertheless, it remains desirable to leverage the advantages of FMs when pursuing complex in-orbit objectives such as resilient positioning, navigation and timing (PNT), space situational awareness (SSA), and collision avoidance.

To provide a holistic example, we propose a novel optical PNT/SSA framework named Fast Autonomous Lost-in-space Catalog-based Optical Navigation (FALCON). FALCON runs on board one or multiple observer spacecraft and uses bearing angles to visible Resident Space Objects (RSO) to: 1) determine the observer's orbit; 2) refine the orbits of tracked RSO targets; and 3) provide RSO collision avoidance alerts. FALCON operates in a "lost in space" manner, i.e. without a-priori observer orbit knowledge. The key idea is to detect RSO in images from an on-board camera, match RSO to existing identities in an RSO catalog, and use the known positions of identified RSO as optical beacons for positioning. Subsequently, RSO catalog data can be used to provide collision assessments and trigger collision avoidance maneuvers. Figure 5 provides a notional illustration.

A key challenge presented by FALCON is its interaction with RSO catalog data. Existing catalogs feature tens of thousands of objects with significant orbit uncertainties, and the catalog identification, positioning, refinement, and collision detection tasks require, e.g., solving intensive geometric optimization problems or extensive orbit propagation. Therefore, we may be able to use an FM's generalist, semantic prior to improve flexibility or reduce computation costs on board. However, satellite data is even further out-of-distribution than planetary rover imagery and existing FMs will likely require extensive training and fine-tuning to bridge the domain gap.

To address this we propose a three-pronged approach. First, a pre-trained open-source FM is selected as the basis, motivated by desired modalities (e.g. image sequences, catalog data) and capabilities. Besides LLaVA, we consider models such as ViNT [21] (tuned towards solving large-scale navigation problems) or Tool-LLM [18] (to facilitate interaction with algorithmic tools). Second, additional training is performed using space-specific datasets: remote sensing data, satellite telemetry, space object catalogs, orbit trajectories, spacecraft hardware databases, among others, to improve generalization to spaceborne data. Third, the model is fine-tuned for tasks of interest using labeled datasets. For FALCON, this includes generating high-fidelity spaceborne images with accompanying ground-truth RSO labels (and/or labeling real spaceflight images), and generating propagated orbits and ground-truth collision estimates from space catalog data.

VI. Conclusion

In this paper, we argue that future challenges in space robotics motivate the development of *a space foundation model*. We introduce a first step towards *a space foundation model* by automatically labeling the AI4Mars dataset with QA pairs to adapt a VLM to terrain classification and navigation tasks for a planetary rover on Mars. We demonstrate that 1) existing VLMs are deficient visual reasoners in space-based applications, and 2) fine-tuning a VLM on automatically labeled in-situ extraterrestrial data significantly improves the quality of responses even with a limited training dataset of only a few thousand samples. We also propose new applications of foundation models to satellite scenarios focusing on highly complex PNT, SSA, and collision avoidance tasks, infeasible using traditional onboard algorithms. Future work in the development of a foundation model for space will incorporate: 1) collecting a sufficiently large and diverse space dataset, e.g., remote sensing data, spaceflight simulations, and space object catalogues, for space-related tasks and 2) developing data encoders to process the diverse modalities (LiDAR, GPS, etc) inherent to these data in order to create a meaningful representation for decision making.

Acknowledgments

The authors acknowledge Blue Origin and Redwire for their support and thank the reviewers for their helpful comments.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Audrey Berquand, Paul Darm, and Annalisa Riccardi. Spacetransformers: Language modeling for space systems. *IEEE Access*, 9:133111–133122, 2021.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257364842.

[8] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa A. D. Nesnas, and Marco Pavone. Semantic anomaly detection with large language models. *Autonomous Robots*, 47:1035 – 1055, 2023. URL https://api.semanticscholar.org/CorpusID:258823112.

[9] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[10] Drew A. Hudson and Christopher D. Manning. Gqa: a new dataset for compositional question answering over real-world images. *ArXiv*, abs/1902.09506, 2019. URL https://api.semanticscholar.org/CorpusID:67855531.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256390509.

[12] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: from natural language instructions to feasible plans. *Autonomous Robots*, Nov 2023. ISSN 1573-7527. doi:10.1007/s10514-023-10131-7. URL https://doi.org/10.1007/s10514-023-10131-7.

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023. URL https://api.semanticscholar.org/CorpusID:263672058.

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

[15] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving, 2023.

[16] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. Lingoqa: Video question answering for autonomous driving. *ArXiv*, abs/2312.14115, 2023. URL https://api.semanticscholar.org/CorpusID:266435950.

[17] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

[18] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

[20] Victor Rodriguez-Fernandez, Alejandro Carrasco, Jason Cheng, Eli Scharf, Peng Mun Siew, and Richard Linares. Language models are spacecraft operators. *arXiv preprint arXiv:2404.00413*, 2024.

[21] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=-K7-1WvKO3F.

[22] R. Michael Swan, Deegan Atha, Henry A. Leopold, Matthew Gildner, Stephanie Oij, Cindy Chiu, and Masahiro Ono. Ai4mars: A dataset for terrain-aware autonomous driving on mars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1982–1991,

2021. doi:10.1109/CVPRW53098.2021.00226.

[23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

[24] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

APPENDIX

A. Template for GPT-4 Preference Evaluation

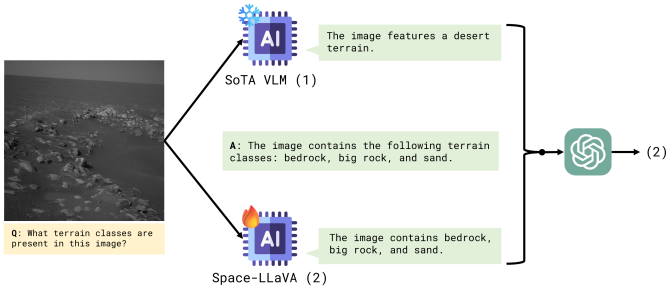


Fig. 6: Preference analysis used to compare Space-LLaVA's performance relative to SoTA VLMs, i.e., GPT-4o and the base LLaVA model.

In order to demonstrate improved proficiency on Mars spatial reasoning and path planning tasks for our fine-tuned models compared to existing pre-trained models, we leverage the GPT-4 language model as an automated text evaluator. In this side-by-side comparison, each model's response is collected along with the ground-truth response for the particular question. This tuple of three natural language responses is provided to GPT-4, prompted according to the following template, to determine which model's answer is most similar to the ground-truth response in terms of content. Specifically, we ask GPT-4 to focus on the content of each model's response in comparison to the ground-truth label rather than the style or prose used, which the fine-tuned model is expected to mirror. We ask GPT-4 to provide its output as a single number indicating its preference between the two model answers. If GPT-4 prefers the base model answer, it outputs 1; if it prefers the fine-tuned model answer, it outputs 2; and if it determines that neither of the model's answers are correct, it outputs 0. We visual describe this evaluation setup in Fig. 6.

Template Prompt for Preference Analysis with GPT-4

System message: "You are evaluating a new foundation model for Mars rover missions. You will be presented with a question, the desired response, and generated responses from two foundation models. Your job is to decide which of the two generated responses you think is most similar to the ground truth response based on the response content (disregarding response structure)" **User message**: "The question is ['question']. The desired response is ['ground-truth-answer']. The generated response from model 1 is ['base-model-answer'] and the generated response from model 2 is ['fine-tune-answer]. Which model's response is most similar to the desired response? Provide your answer as a single number (1 or 2) that indicates which model's response you think is most similar to the desired response. If neither model's response is correct relative to the desired response, provide your answer as 0. Do not provide any justification or explanation in your answer."

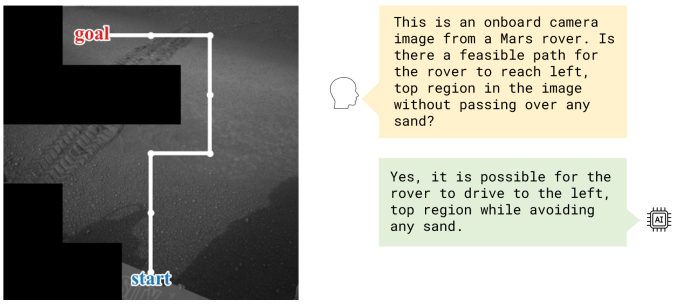
We run this procedure over 12951 image-question pairs in our withheld evaluation dataset, and track how many times the GPT-4 evaluator prefers the fine-tuned model answer. The percentage of evaluation samples on which GPT-4 prefers the fine-tuned model answer is thus used as the response preference metric described in the problem formulation for evaluation. Improved performance on Mars spatial reasoning and high-level path planning tasks for our fine-tuned model should be demonstrated in response preferences that are much higher than 50%.

B. High-level Motion Planning Examples

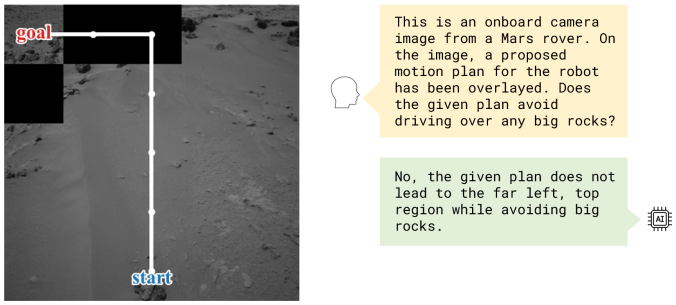
As discussed in IV-B, for each sample in the MSL dataset, i.e., each sample for which a mask exists to identify the rover, we programmatically curate 2-3 QA pairs requiring a understanding of naviation on Mars' terrain. For all MSL samples, we divide the image into a 5x5 grid and curate two questions: 1) we ask the model to identify whether a path exists to connect a start and end point without crossing undesirable terrain and 2) whether a candidate path, which we overlay on the image, is both feasible and reaches the desired end point. As you can see, both of these questions do not require that a feasible path exists between any two particular grid points in the image. Hence, we curate these questions for every MSL image. If a feasible path does in fact exist between the selected start point and at least one distinct end point, then we use the $A^\star$ search algorithm without diagonal movements to identify a path in natural language to navigate the grid. We provide an example of each VQA type in Fig. 7. When training our model on feasibility questions as in Fig. 7a, we provide the *raw camera image* without masking out undesirable terrain or providing a potential path between the start and end grid point: we provide these annotations in Fig. 7a for visual clarity as to why a feasible path exists and how to navigate through the image. Similarly, for Fig. 7c, during training we do not provide the model with masked terrain or path annotations. That is, as in our evaluation of feasibility, we provide the raw camera image. In Fig. 7c we mask infeasible grid sections to justify the ground-truth annotation and overlay the correct path to provide a visual representation of our chosen path in natural language.

C. Training Infrastructure

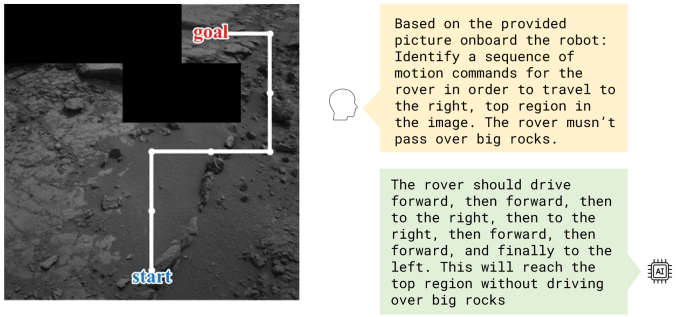
For each configuration, we perform a hyperparameter sweep over the learning rate and weight decay with values inspired by Marcu et al. [16] and Gao et al. [9]. The learning rate is selected from $\{1\times10^{-5},3\times10^{-5},5\times10^{-5}\}$, and the weight decay is selected from $\{0.05,0.1\}$ for a total of 6 settings for each model configuration. Every configuration is trained with an 80-20 train-validation split for 1-2 epochs on a compute infrastructure with 4 80GB A100 GPUs. For each configuration, the model with the best validation loss is evaluated in the preference analysis against the base LLaVA model and GPT-4o. We perform preference analysis as described previously in III, with further detail provided in VI-A.



(a) Feasibility: We are able to automatically identify whether the left, top region of the image is accessible starting from the center, bottom region without traversing sand. In this particular case, a path does exist while traversing soil.



(b) Feasibility overlay: We are able to automatically overlay a candidate path from the center, bottom to the far left, top region of the image while avoiding big rocks. In this particular case, the candidate path is infeasible as it requires the rover to traverse big rocks near the center, top of the image.



(c) Planning: Given that a feasible path exists between two distinct points, we are able to automatically generate a language description of a path from the center, bottom to the right, top region of the image while avoiding big rocks.

Fig. 7: Automatically generated navigation QA pairs on AI4Mars instantiated as a feasibility test, evaluating a candidate path and planning a route in a 5x5 grid on the image. This data collection pipeline only leverages the semantic masks provided by the AI4Mars dataset.