LunarOS-Mini: A Two-Stage Decoupled Visual Language Model for Lunar SAR Imagery

Yi Zheng
Key Laboratory of Information Science of Electromagnetic Waves (MoE)
Fudan University
Shanghai, China
zhengyi24@m.fudan.edu.cn

Dezhang Li
Key Laboratory of Information Science of Electromagnetic Waves (MoE)
Fudan University
Shanghai, China
lidz25@m.fudan.edu.cn

Niutao Liu*
Key Laboratory of Information Science of Electromagnetic Waves (MoE)
Fudan University
Shanghai, China
ntliu@fudan.edu.cn

***Abstract*—Single-modal optical imagery exhibits limitations when adopted for lunar geomorphic interpretation via vision-language models (VLMs), motivating the development of optical-SAR multimodal frameworks. However, visual-language interpretation of lunar SAR imagery still faces two major bottlenecks. First, high-quality lunar SAR-optical multimodal paired datasets remain scarce. Second, optical and SAR modalities suffer alignment obstacles during cross-modal fusion. To address these issues, LunarOS-Mini, a two-stage decoupled multimodal VLM for lunar SAR imagery is proposed. 2,400 real orbital optical-SAR pairs are selected from LOSAD-14k and textual descriptions are added to build the dataset. A structured prompting pipeline based on ChatGPT-5.5 High is devised to produce geological annotations spanning geomorphic classes, optical morphological characteristics, SAR backscatter textures, and cross-modal correlations. A two-stage decoupled training framework is developed for cross-modal alignment. Stage-wise LoRA tuning unifies lunar SAR, optical and textual features. Experiments use three LLM-based judges, Deepseek-V4, GPT-5.5 High, and Gemini-3.5, to evaluate the model from five dimensions. The proposed two-stage model achieves superior scores on all metrics. With full scores for question relevance and GTA, indicating the effectiveness of the proposed dataset and two-stage decoupled framework under the adopted evaluation protocol.**

Keywords—SAR, Visual Language Model, Lunar, Multimodals

I. INTRODUCTION

Large-scale multimodal image-text paired data have provided an important foundation for the development of visual language models (VLMs). Such data enable models to learn associations between visual information and textual descriptions. Models such as CLIP [1], BLIP [2], and Flamingo [3] have demonstrated strong visual perception and understanding capabilities. With supervised fine-tuning (SFT) [4], VLMs can acquire cross-task generalization and adapt to vertical-domain visual-language tasks. This enables models to interpret visual content and perform logical reasoning in a more human-like manner. Such techniques have also been applied to lunar remote sensing interpretation [5].

However, single-modal planetary remote sensing data cannot fully characterize lunar geomorphic features. Multimodal interpretation suffers from a shortage of standardized paired datasets and intrinsic cross-modal alignment difficulties. For example, the currently available open-source lunar VLM, LLaVA-LE [5], uses only optical image and text, and therefore provides incomplete interpretation capability for lunar scenes. These limitations make existing resources insufficient for training and evaluating lunar visual-language interpretation models. A high-quality image-text dataset based on real lunar exploration data and multiple modalities is therefore needed to support subsequent research.

To tackle the scarcity of multimodal datasets, 2,400 real orbital optical-SAR pairs are selected from LOSAD-14k, and GPT-5,5 high is leveraged to generate corresponding descriptive captions and dialogue samples. Based on the Qwen-VL [6] multimodal large-model architecture and real Mini-RF and NAC observations. A dedicated SAR-NAC-text dataset for lunar exploration tasks is constructed. The dataset supports domain adaptation of the model to lunar scenarios. To provide semantic guidance, we design a structured prompting pipeline and use ChatGPT-5.5 High to generate scientific annotations. These annotations include geomorphic scene categories, optical geomorphic features, SAR backscattering textures, and bimodal cross-association information, and provide supervision for learning geospatial understanding. The multimodal image-text pairs are then converted into diverse instruction-answer formats for SFT. It enables the model to perform image captioning, visual question answering, and cross-modal reasoning.

The remainder of this paper is organized as follows. Section 2 introduces the adopted SAR-optical registered dataset, LunarOS-Mini, and the proposed two-stage model training architecture. Section 3 presents the experiment results. Section 4 concludes the paper.

II. DATA AND METHOD

To address the shortage of lunar multimodal image-text datasets, we construct a dedicated SAR-NAC image-text dataset based on the open-accessed LOSAD-14K [7] dataset. LOSAD-14K contains a large number of physically simulated optical samples, whose feature distribution differs from that of real optical observations. 2,400 real orbital optical-SAR pairs are selected from LOSAD-14k and are divided into training, validation, and test sets at a ratio of 8:1:1. To match this data scale, Qwen3-VL-4B-Instruct is selected as the base model for downstream training and evaluation.

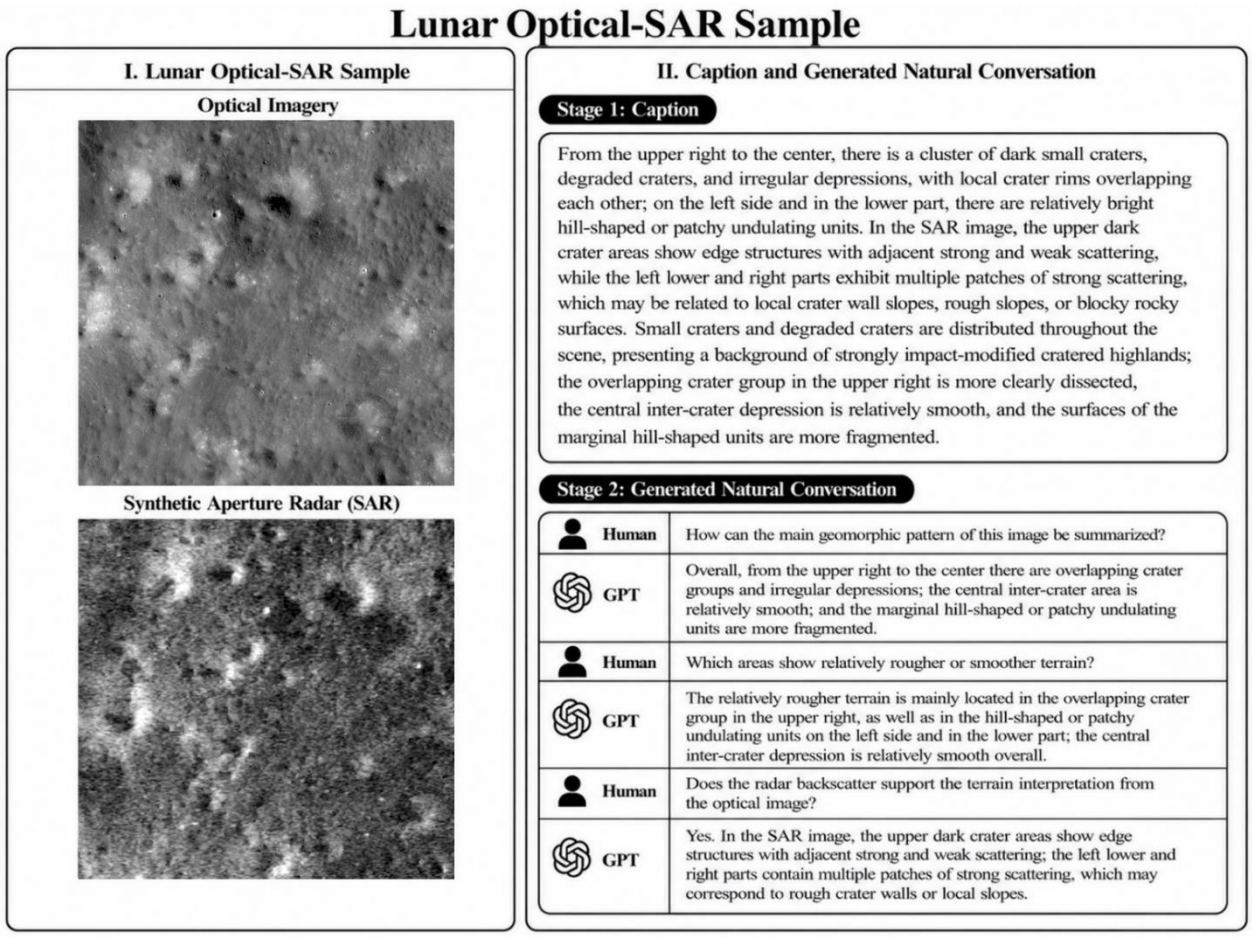

Fig. 1. An example of LunarOS-Mini dataset.

A. LunarOS-Mini Dataset

To support interpretation of the real lunar surface, LunarOS-Mini is constructed, a small-scale dataset for training VLMs. The dataset contains two parts. The first part is an image-text description, dataset with 2,400 registered SAR-NAC-text triplets. It is designed to learn mappings between visual features and geological descriptions. In the second part, the annotation texts are converted into question-answer interaction formats. For each text description, three different questions are randomly generated to simulate human-machine dialogue, resulting in 2,400 three-turn dialogue samples.

Although multimodal datasets such as LUCID [5] have been released for planetary remote sensing, their modalities remain limited. LunarOS-Mini addresses this gap by introducing the SAR modality. To enrich the semantic supervision, we further design a structured prompting pipeline and use ChatGPT-5.5 High to generate scientific annotations. These annotations provide the semantic basis for both the first-stage caption data and the second-stage SFT data.

B. Two-Stage Model Training Architecture

To adapt to the small-scale LunarOS-Mini dataset, we adopt Qwen3-VL-4B-Instruct as the base weights and train the model in two progressive stages: concept alignment and instruction tuning.

1) Concept Alignment: This stage establishes cross-modal alignment among lunar SAR images, optical images, and text descriptions. During training, the visual encoder and the visual-language adapter remain frozen. LoRA [8] low-rank matrices are injected only into the Q, K, V, and O projection layers of the Transformer [9] modules for parameter-efficient fine-tuning. This keeps the number of trainable parameters at the million level while preserving the model's general multimodal knowledge. The CausalLM loss is used as the optimization objective. After this stage, the model can map SAR and optical geomorphic features into the language embedding space and learn basic associations between visual cues and lunar geological semantics.

2) Instruction Tuning: The second stage improves the model's ability to follow instructions and generalize across visual-language tasks. Based on the aligned representations from the first stage, the annotated image-text pairs are converted into diverse instruction-answer formats. Specifically, each text description is randomly expanded into three question-answer pairs to simulate human-machine interaction, yielding 2,400 three-turn dialogue samples. In this stage, LoRA is used to jointly fine-tune the visual-language adapter and the large language model. This helps adapt the model to the target domain while reducing catastrophic forgetting. The model generates responses autoregressively from visual tokens and text instructions,

using the standard cross-entropy loss for next-token prediction. The loss is applied only to response tokens. The instruction text serves as conditional context and is not included in the loss calculation. Instruction tuning enables zero-shot cross-task generalization for image captioning, visual question answering, and cross-modal reasoning in lunar exploration scenarios.

III. Experiments and Results

A random 10% subset of LunarOS-Mini is used as the test set and is excluded from the two-stage training process. GPT-5.5 High is used to generate 720 evaluation questions involving geomorphic interpretation and joint SAR-NAC interpretation. The model answers each question using the imagery and the question, without additional prompts. During evaluation, the real annotations of LunarOS-Mini are used as references. GPT-5.5, Deepseek-V4, and Gemini-3.5 automatically score the responses from five dimensions: fluency (FL), question relevance (QR), ground-truth annotation alignment (GTA), multi-turn consistency (MC), and professional information density (PID). Each dimension has a maximum score of 20, for a total score of 100. All experiments are conducted on four Nvidia RTX 3090 GPUs with 24 GB of video memory.

Fig. 2 compares the response quality of the Qwen3-VL-4B-Instruct base model and its Stage 1 and Stage 2 fine-tuned versions on a single example. The validation column denotes the reference annotation in the test set. The base model identifies some geomorphic features, but its responses are broad and weakly organized. After Stage 1 fine-tuning, the responses become more structured and begin to capture the relationship between SAR textures and slope-orientation changes. After Stage 2 fine-tuning, the model produces more concise and domain-specific responses. Demonstrating consistency with the reference annotation. This progressive improvement indicates the effectiveness of the two-stage fine-tuning strategy.

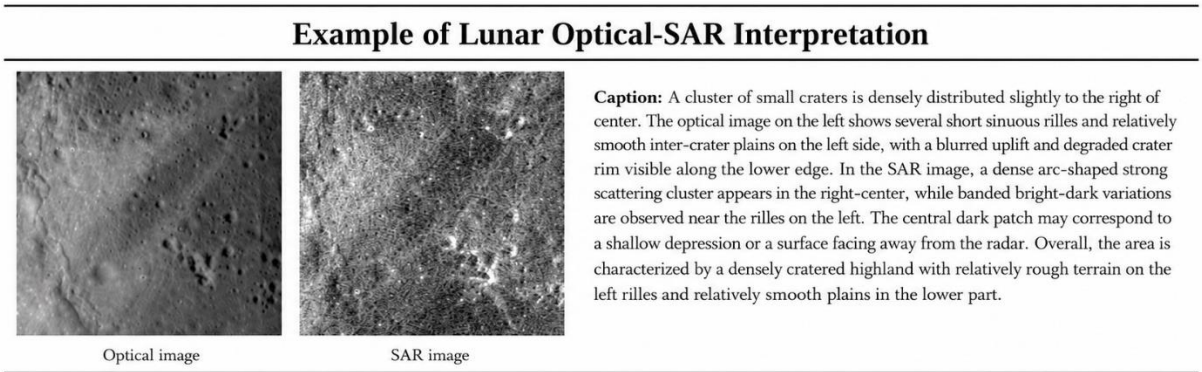

Fig. 2. An example of LunarOS-Mini interpretation demo.

Model-generated answers are evaluated from five dimensions: FL, QR, GTA, MC, and PID. Fluency measures the naturalness and logical coherence of the response. QR examines whether the response matches the user's question intent. GTA evaluates the consistency between the response and the LunarOS-Mini reference annotation in format, terminology, and style. MC tests whether the model maintains logical coherence and complementary information in multi-turn dialogue. PID measures the density of useful geomorphic information. These five dimensions jointly assess the model's geomorphic interpretation capability. Each dimension has a maximum score of 20, for a total score of 100. GPT-5.5, Deepseek-V4, and Gemini-3.5 independently assign scores to reduce the bias of any single LLM-based judge.

The experimental results in Table I show that the two-stage fine-tuning strategy improves the base model across all five evaluation dimensions. The Qwen3-VL-4B-Instruct base model obtains an average score of 71.7 under the three LLM-based judges. Stage 1 increases the score to 80.0, and Stage 2 further raises it to 93.0. This monotonic trend indicates the effectiveness of the proposed training strategy under the adopted evaluation protocol. Among the five dimensions, fluency is the strongest dimension of the base model, with an average score of 16.3. After two-stage fine-tuning, it reaches 18.7, suggesting that the base model already has a relatively solid language-generation capability. GTA and PID show larger gains. GTA increases from 12.7 for the base model to 19.0 at Stage 2, while PID improves from 14.0 to 18.0. These results suggest that Stage 2 helps the model produce more concise and domain-specific geomorphic interpretations. The three LLM-based judges show consistent trends, with Stage 2 outperforming Stage 1 and Stage 1 outperforming the Base model. This suggests that the improvement is consistent across different LLM-based judges. Gemini-3.5 assigns the Stage 2 model an overall score of 97, with full scores for QR and GTA, further indicating the benefit of the proposed method for lunar remote sensing geomorphic interpretation.

TABLE I
Performance Evaluation by Different LLM Judges

Judge	Model	FL	QR	GTA	MC	PID	Overall
Deepseek-V4	Qwen3-VL-4B-Instruct	18.0	16.0	14.0	15.0	13.0	76.0
	Qwen3-VL-4B-Instruct-Stage1	18.0	18.0	17.0	17.0	16.0	86.0
	Qwen3-VL-4B-Instruct-Stage2	**19.0**	**19.0**	**19.0**	**19.0**	**18.0**	**94.0**
GPT-5.5 High	Qwen3-VL-4B-Instruct	16.0	10.0	7.0	9.0	14.0	56.0
	Qwen3-VL-4B-Instruct-Stage1	17.0	11.0	9.0	10.0	15.0	62.0
	Qwen3-VL-4B-Instruct-Stage2	**18.0**	**18.0**	**18.0**	**17.0**	**17.0**	**88.0**
Gemini-3.5	Qwen3-VL-4B-Instruct	18.0	17.0	16.0	17.0	15.0	83.0
	Qwen3-VL-4B-Instruct-Stage1	18.0	19.0	18.0	18.0	19.0	92.0
	Qwen3-VL-4B-Instruct-Stage2	**19.0**	**20.0**	**20.0**	**19.0**	**19.0**	**97.0**

The best results are in blod.

IV. Conclusion

LunarOS-Mini is presented as a SAR-optical visual-language framework for lunar remote sensing interpretation. By constructing a real-observation-based SAR-NAC-text dataset and adopting a two-stage LoRA fine-tuning strategy, the proposed method improves the model's ability to generate domain-specific geomorphic descriptions and answer SAR-NAC reasoning questions. The framework uses concept alignment to connect SAR-optical geomorphic features with geological semantics and then applies instruction tuning to improve task generalization.

Evaluation shows that the two-stage model achieves stable improvements over both the original Qwen3-VL-4B-Instruct base model and the single-stage aligned model. The average overall score reaches 93 across the three LLM-based judges, and the improvement trend is consistent across fluency, QR, GTA, MC, and PID. Under the Gemini-3.5 evaluation system, the model achieves an overall score of 97 and obtains full scores for QR and GTA. These results indicate that integrating SAR backscattering information with optical imagery is beneficial for fine-grained lunar geomorphic interpretation under the adopted evaluation protocol.

Overall, LunarOS-Mini delivers a solution for small-sample multimodal interpretation of lunar multi-source remote sensing data and lays a foundation for developing SAR-optical VLMs dedicated to planetary exploration. However, the model is limited to basic semantic interpretation at the current stage. Exploring its extension to tackle professional lunar scientific tasks remains an open research direction [10]. In the future, the upcoming Chang'E-7 mission will acquire abundant polar optical-SAR multi-source observations, which can further promote the practical deployment of such multimodal vision-language frameworks for lunar scientific analysis [11].

Acknowledgment

The Mini-RF data, NAC data are available on NASA's Planetary Data System (http://pds-geosciences.wustl.edu). The authors would thank the LOSAD-14k dataset produce team for their great work in producing and maintaining the data used in this study.

References

[1] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *ArXiv,* vol. abs/2103.00020, 2021.

[2] J. Li, D. Li, C. Xiong *et al.*, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation."

[3] J.-B. Alayrac, J. Donahue, P. Luc *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," *ArXiv,* vol. abs/2204.14198, 2022.

[4] L. Ouyang, J. Wu, X. Jiang *et al.*, "Training language models to follow instructions with human feedback," *ArXiv,* vol. abs/2203.02155, 2022.

[5] G. İnal, P. Navard, and A. Yilmaz, "LLaVA-LE: Large Language-and-Vision Assistant for Lunar Exploration," *ArXiv,* vol. abs/2603.24696, 2026.

[6] J. Bai, S. Bai, S. Yang *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond."

[7] T. Xia, N. Liu, Y. Zheng *et al.*, "LunarS2O: SAR-to-optical image translation for permanently shadowed region exploration and its application," *Icarus,* vol. 457, pp. 117178, 2026/10/01/, 2026.

[8] J. E. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *ArXiv,* vol. abs/2106.09685, 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is All you Need."

[10] R. Wang, W. Wang, Y. Shao *et al.*, "First Bistatic Demonstration of Digital Beamforming in Elevation With TerraSAR-X as an Illuminator," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 54, no. 2, pp. 842-849, 2016.

[11] H. Yue, Y. Wang, P. Wang *et al.*, "Ground Calibration Method for the Chang'E-7 Lunar Microwave Imaging Radar," *Chinese Journal of Space Science*, 2026.