LLaVA-LE: Large Language-and-Vision Assistant for Lunar Exploration

Anonymous CVPR submission

Paper ID 10

Abstract

Recent advances in multimodal vision–language models (VLMs) have enabled joint reasoning over visual and textual information, yet their application to planetary science remains largely unexplored. A key hindrance is the absence of large-scale datasets that pair real planetary imagery with detailed scientific descriptions. In this work, we introduce ***LLaVA-LE*** *(Large Language-and-Vision Assistant for Lunar Exploration), a vision–language model specialized for lunar surface and subsurface characterization. To enable this capability, we curate a new large-scale multimodal lunar dataset,* ***LUCID*** *(**LU**nar **C**aption **I**mage **D**ataset) consisting of* ***96k*** *high-resolution panchromatic images paired with detailed captions describing lunar terrain characteristics, and* ***81k*** *question-answer (QA) pairs derived from ∼20k images in the LUCID dataset. Leveraging this dataset, we fine-tune LLaVA using a two-stage training curriculum: (1) concept alignment for domain-specific terrain description, and (2) instruction-tuned visual question answering. We further design evaluation benchmarks spanning multiple levels of reasoning complexity relevant to lunar terrain analysis. Evaluated against GPT and Gemini judges, LLaVA-LE achieves a* ***3.3****× overall performance gain over Base LLaVA and* ***2.1****× over our Stage 1 model, with a reasoning score of* ***1.070*** *— exceeding the judge's own reference score — highlighting the effectiveness of domain-specific multimodal data and instruction tuning to advance VLMs in planetary exploration.*[1]

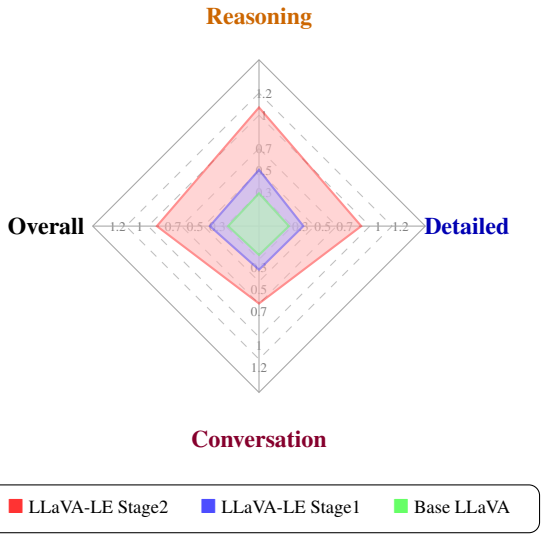


Figure 1. **Category-wise Visual Question Answering Performance Relative to Judge Scores**. We evaluate Base LLaVA, LLaVA-LE Stage 1, and LLaVA-LE Stage 2 on a held-out evaluation set across three question categories: Detailed, Conversation, and Reasoning. Each axis shows the model's score normalized by the judge's score (Score / Judge Score), averaged across GPT-4 and Gemini judges. The Overall axis represents the mean across all three categories. Values greater than 1.0 indicate the model outperforms the judge's reference score.

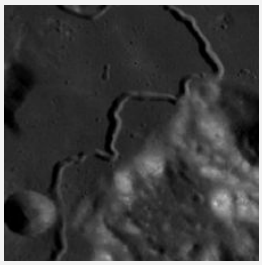
1. Introduction

The wide availability of image-text datasets on the Internet has driven significant progress in vision-language modeling. Large-scale collections of natural images paired with descriptive captions provide an ideal foundation for generative pretraining, where models learn visual-linguistic associations in a self-supervised manner. This approach allows networks to build rich joint representations of vision and language without relying on labor intensive and costly manual labels. Using large-scale image-text pairs, models such as CLIP [35], BLIP [21] and LLaVA [25] have demonstrated strong visual reasoning capabilities in general domain settings. Instruction-tuning has been shown to equip large vision-language models with the ability to generalize across tasks in a zero-shot manner, completing a wide range of vision-language objectives without task-specific supervision. This framework has opened up the way for developing general-purpose multimodal conversational systems that can describe, interpret, and reason about visual scenes in a human-like manner.

Despite the fact that these advances have transformed general-domain adaptation for various applications, they

[1] We release model weights and data upon publication.

Caption:
A sharp sinuous rille crosses the scene from upper right to lower left, separating smooth mare-like plains to the north from rugged, blocky highland material to the south. The rilles narrow, continuous walls and locally steep inner flanks suggest incision into relatively coherent, competent crust. Adjacent mare surfaces show low crater density and embayment of older relief, consistent with relatively young basaltic flooding that may overlie thicker, mechanically stable subsurface material. The rough highland massif, with abundant small craters and irregular relief, likely reflects older, more fractured crust of lower overall strength.



Conversation:
User: What is the main feature visible in the scene?
Assistant: A sharp sinuous rille crosses the scene from upper right to lower left.

User: What types of terrain does the rille separate?
Assistant: It separates smooth mare-like plains to the north from rugged, blocky highland material to the south.

User: What does the appearance of the rille suggest about the type of crust it has incised into?
Assistant: The narrow, continuous walls and locally steep inner flanks suggest incision into relatively coherent, competent crust.

User: What can be inferred about the age of the mare surfaces compared to the highland massif?
Assistant: The mare surfaces show low crater density and embayment of older relief, indicating they are relatively younger due to basaltic flooding, while the highland massif reflects older, more fractured crust.

Figure 2. **An Example of LUCID dataset**: The top shows the the panchromatic image with the associated caption generated using GPT, while the below shows question-answer pairs generated from that caption.

struggle to generalize to planetary exploration where there is a lack of large-scale, high-quality paired image-text corpora. Existing lunar datasets [17, 27, 28] are uni-modal, small in scale, frequently contain mixtures of synthetic simulations and real observations, and often suffer from limited spatial resolution or mismatched distributions between modalities. Notably, the Space-LLaVA [11] fined-tuned LlaVA model on synthetic multi-modal data for extraterrestrial applications. However, neither the data nor the codebase were publicly made available for Space-LLaVA at the time of submission of our paper. Such limitations make them unsuitable for training modern vision-language models motivating construction of a large scale, high quality training data corpus of real lunar imagery.

Planetary remote sensing is fundamentally different from natural image understanding: interpreting a lunar region requires reasoning across multiple physical modalities, not just visual appearance. Although the relationship between images and text is largely self-contained in natural vision tasks, planetary remote sensing is inherently multimodal. In natural images, visual patterns reflect underlying physical and geological processes rather than the semantic object categories typically present in natural imagery. A single optical image often provides an incomplete representation of the geological structure of a region, revealing only surface reflectance patterns. Complementary modalities, such as gravity anomaly maps or terrain slope models, provide additional evidence on subsurface mass distributions, regolith structure, and morphological evolution. These modalities capture different physical aspects of the same region and must be interpreted jointly to produce meaningful geological insights. As a result, planetary analysis fundamentally requires reasoning across multiple co-registered data layers that together characterize both surface morphology and subsurface structure.

Despite the scientific importance of such multimodal reasoning, there currently exists no vision-language assistant trained on large-scale real planetary datasets capable of performing high-level geological interpretation from multimodal remote sensing observations. Prior efforts in multimodal planetary analysis rely primarily on small curated datasets or synthetic simulations, limiting their applicability to real-world exploration scenarios. To address this gap, we introduce **L**arge **L**anguage **A**nd **V**ision **A**ssistant for **L**unar **E**xploration (LLaVA-LE), a vision-language model designed for multimodal geological understanding and reasoning over lunar remote sensing data.

Our approach builds upon multi-modal reasoning capability of the LLaVA framework and adapts it to the planetary science domain through the creation of a large-scale

multimodal lunar dataset derived from real NASA mission observations. Specifically, we curate co-registered data products from the Lunar Reconnaissance Orbiter Camera (LROC) [37], Gravity Recovery and Interior Laboratory (GRAIL) [56], and Lunar Orbiter Laser Altimeter (LOLA) [40] missions, representing high-resolution optical imagery, gravity anomaly measurements, and terrain slope maps. From these mission products, we constructed aligned multimodal image triplets that jointly capture surface morphology and subsurface gravitational structure. To provide rich semantic supervision, we generate detailed scientific descriptions using a structured prompting pipeline with GPT-5, producing captions that describe the geological context, terrain morphology, and inferred subsurface characteristics. These multimodal image–text pairs are subsequently transformed into instruction–response examples used to train a planetary vision-language assistant capable of describing lunar terrain, answering geological questions, and performing multimodal reasoning across planetary data layers.

Our contributions are as follows:

- **LLaVA-LE**. We introduce a vision-language model specialized for planetary science, adapting the LLaVA framework to perform multimodal reasoning over on *real* lunar remote sensing data. Our training pipeline leverages a curriculum-style instruction generation process that enables the model to learn geological interpretation and scientific query answering from multimodal observations.
- **LUCID**. We release the **first** real large-scale *real* multimodal lunar dataset for vision-language learning. LUCID consists of **96k** panchromatic imagery paired with detailed scientific captions, as well as **81K** question-answer pairs as a VQA dataset. LUCID (LUnar Caption Image Dataset) enables scalable training and evaluation of multimodal models for lunar surface and subsurface analysis.
- **Open Source**. To facilitate further research in AI-driven planetary exploration, we publicly release the LUCID dataset, the codebase, and LLaVA-LE model checkpoints.

2. Related Work

Visual Instruction Tuning. Visual instruction tuning aims to align vision-language models (VLMs) with natural language instructions using image-response pairs [15]. Although early work in large language models (LLMs) such as ChatGPT [33], GPT-4 [34], and open-source variants like Alpaca [52], Vicuna [7], and Dolly [9] established the efficacy of instruction following, LLaVA [25] pioneered this in the multimodal domain by connecting a CLIP [35] visual encoder to an LLM. Subsequent architectures such as KOSMOS-1 [16], BLIP-2 [22], and Flamingo [1] explored various fusion strategies. Extensions such as MiniGPT-v2 [4], mPLUG-Owl2 [50], Otter [19], StableLLaVA [24], PandaGPT [42], and LAMM [51] further scaled these methods to multi-tasking and diverse modalities. Our work builds on the LLaVA framework, adapting its conversational capabilities to the specialized constraints of planetary imagery.

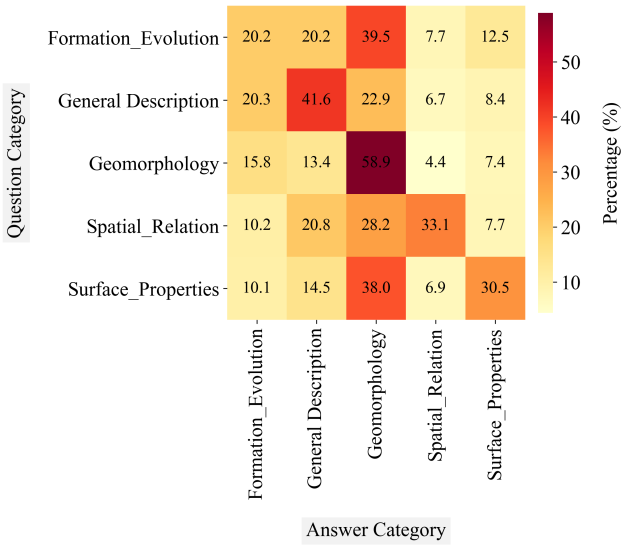


Figure 3. **Distribution of LUCID VQA dataset across question-answer pairs categories.** This heatmap matrix shows how much each answer is contributing, in terms of percentage, to each one of the question categories across the entire LUCID visual question answering dataset. The heatmap is normalized by the row dimension and shows the reasoning capability captured in our dataset to answer each question.

Domain-Specific Visual Assistants. Specialized instruction tuning has seen significant success in technical fields, particularly medicine. Text-based clinical assistants [14, 23, 45, 46, 48] paved the way for multimodal medical models. LLaVA-Med [20] and related frameworks like OphGLM [12], Qilin-Med-VL [26], Med-Flamingo [29], PCLMed [49], MedBLIP [5], CheXagent [6], PathAsst [43], SkinGPT-4 [55], and SigPhi-Med [54] showed that fine-tuning on domain-specific data (e.g., radiology, pathology) significantly outperforms general-purpose models.

Beyond medicine, instruction tuning has been applied to robotics such as PaLM-E [10], EmbodiedGPT [31], visual programming [13, 44], and editing [2]. Although AlphaEarth [3] indicates a growing interest in earth observation multimodal data, the application of these techniques to planetary science remains nascent. We address this gap by introducing a LLaVA-based approach tailored for lunar surface analysis.

3. Method

In this section, we describe the data and training procedure used to develop **LLaVA-LE**. We first introduce LUCID (**L**unar **C**aptioned **I**mage **D**ata) in section 3.1, a

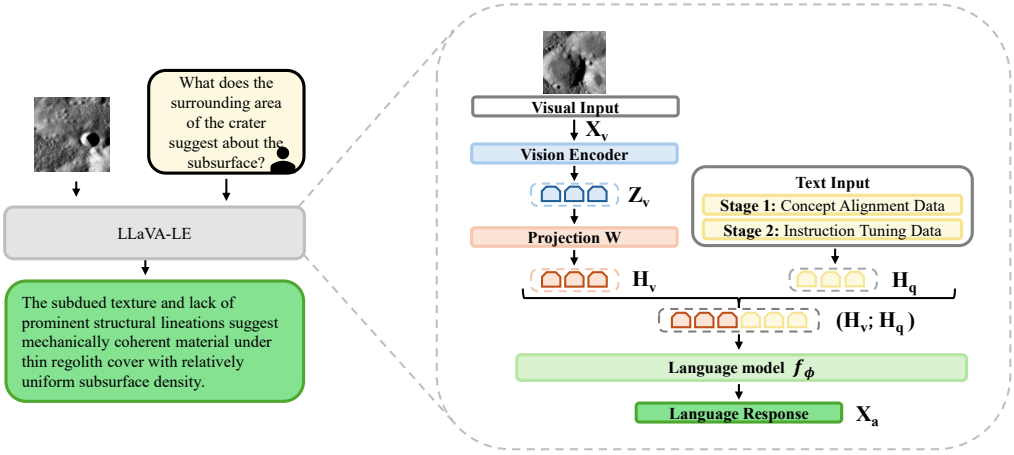


Figure 4. **Overview of LLaVA-LE Training and Inference.** The right panel illustrates the two-stage training pipeline. Visual inputs X_v are encoded by a frozen CLIP vision encoder g to produce visual features $Z_v = g(X_v)$, which are then projected into the language embedding space via a trainable projection layer, yielding H_v to be concatenated with the language token embeddings H_q. The language model backbone f_ϕ autoregressively generates a response X_a conditioned on the joint visual and language context. In **Stage 1 (Concept Alignment)**, caption supervision aligns lunar visual features with domain-specific geological language. In **Stage 2 (Instruction Tuning)**, the aligned model is further fine-tuned on multi-turn question-answer pairs to enable reasoning and conversational interaction. Throughout both stages, the CLIP vision encoder and the pretrained language model backbone f_ϕ remain frozen; only the projection layer and lightweight LoRA adaptation modules are updated. The left panel illustrates a representative inference example after Stage 2, where a user query about lunar surface observations is processed and the model generates a domain-grounded geological interpretation X_a.

large-scale multimodal dataset designed to enable vision-language learning for lunar surface analysis. LUCID is constructed in two stages to support both planetary domain specific learning and instruction-following behavior. We then describe how Llava-LE is adapted from LLaVa using the LUCID dataset in section 3.2.

3.1. Lunar Captioned Image Data (LUCID)

To enable multimodal learning on real planetary observations, we constructed LUCID, a large-scale dataset designed for vision-language training on lunar surface analysis. LUCID contains **96K** samples derived from co-registered lunar remote sensing observations. The dataset is organized into two stages that mirror the training procedure for modern vision-language assistants. The first stage contains **76K** samples used for *concept alignment*, where the model learns to associate visual patterns with geological descriptions. The second stage contains approximately **20K** samples used for *instruction tuning* for LLaVA-LE, where the caption data are transformed into question–answer interactions that simulate how users may query a lunar exploration assistant.

There is currently no publicly available multimodal dataset in planetary science that supports the training of vision-language models on *real* planetary observations. **LUCID** addresses this gap by generating captions for high-resolution panchromatic imagery while leveraging complementary geophysical slope and gravity maps to ground the descriptions in scientifically meaningful observations. In the first stage, we generate image–caption pairs describing the morphology of the lunar surface across diverse terrains. In the second stage, these captions are converted into multi-round question–answer dialogs that emulate interactions between a user and a lunar exploration assistant. Figure 2 shows an instance from our LUCID dataset.

3.1.1. Concept Alignment Data Generation

Decades of lunar remote-sensing research demonstrate that panchromatic images are one of the primary data sources for studying tectonics, volcanic history, and the evolution of stress fields on the Moon [30, 47, 53]. However, accurate geological interpretation typically requires analyzing panchromatic images in conjunction with complementary geophysical measurements. The surface morphology observed in panchromatic images is closely related to the underlying physical properties of the crust, including variations in slope, crustal thickness, and mass distribution measured by gravity missions [32, 36, 38, 39, 41, 53]. Features such as the texture of the plains, the sharpness of topographic boundaries, the morphology of the crater rim, and the organization of structural lineaments are often manifestations of geological processes that also produce characteristic patterns in slope and gravity fields. Consequently, reliable interpretation of lunar surface and subsurface structure benefits from jointly analyzing panchromatic images with slope and gravity data, allowing surface observations to be grounded in the broader geophysical context.

Motivated by these well-established relationships between surface morphology and underlying geological structure, we first assemble a large set of lunar panchromatic image tiles from publicly available resources from the Lunar Reconnaissance Orbiter Camera mission [37]. Each image tile is partitioned into uniform spatial patches to create a consistent training corpus. The original tiles of 1504×832 px at a resolution of 125 m/px were each subdivided into 60 patches of 224×224 px, preserving the native resolution so that each patch spans approximately 28×28 km, corresponding to a surface coverage per-patch of $\sim$784 km^2. For every patch, we generate a corresponding scientific description by querying GPT-5.1 using a structured prompt designed to produce geophysically-grounded detailed captions of the observed terrain.

This caption generation process is informed by the scientific relationships between surface morphology and geophysical structure. When generating captions, GPT-5.1 is instructed to focus on observable terrain attributes such as surface texture, crater morphology, structural lineaments, and shading patterns indicative of local topography. In addition to the panchromatic imagery, co-registered gravity anomaly maps and terrain slope measurements are provided as reference modalities to guide the interpretation of surface patterns during caption generation. Incorporating these complementary data sources encourages captions that reflect both visible surface features and inferred subsurface characteristics using slope and gravity. To the best of our knowledge LUCID is the first publicly available dataset for multi-modal lunar surface analysis.

These captions subsequently serve as the foundation for constructing the instruction-following dataset used in the second training stage.

3.1.2. Instruction-Tuning Data

To align the model with instruction-following behavior in the planetary science setting, we randomly selected $\sim$20K samples from our caption dataset (held out from the first-stage training described in Section 3.2.1) and prompted GPT-5.1 to generate 81K question–answer pairs, forming a VQA dataset derived from LUCID. This process transforms static image–caption pairs in Section 3.1.1 into structured dialogues that resemble how a scientist or mission operator may interact with a lunar exploration assistant.

Given a panchromatic image X_v and its corresponding caption X_c, we construct an instruction query X_q that asks the model to analyze the terrain depicted in the image. In its simplest form, an instruction-following interaction can be expressed as:

$$\begin{aligned} &\text{Human: } X_q,\ X_v\ \langle\text{STOP}\rangle \\ &\text{Assistant: } X_c\ \langle\text{STOP}\rangle \end{aligned}$$

This formulation enables the model to learn to produce descriptive geological explanations conditioned on both the image and the instruction prompt. An example of the generated instruction-following data is shown in Figure 2.

The generated questions-answers are designed to capture several core aspects of lunar terrain analysis, including surface morphology, geological processes that may explain the observed structures, indicators of relative surface age such as crater degradation and resurfacing, and localized spatial relationships between terrain features. For each caption, GPT-5.1 generates between three to five question–answer pairs that follow these guidelines, producing diverse conversational training examples grounded in the same underlying scene description.

Figure 3 demonstrates the strong reasoning ability present in the question and answer pairs of the LUCID dataset across all five categories: Formation & Evolution, General Description, Geomorphology, Spatial Relation and Surface Properties. To perform this analysis, we applied a lightweight NLP pipeline using spaCy to extract verbs and nouns from both question and answer turns, which were then classified into the five categories using a weighted lexicon-based scoring scheme. As shown in Figure 3, answering a question from any single category consistently draws upon reasoning across multiple categories, indicating that the dataset does not encourage narrow responses but instead promotes broad, multi-dimensional reasoning — a key property for training capable vision-language models.

3.2. LLaVA-LE

We initialize LLaVA-LE from a general-domain multimodal conversational model, LLaVA-v1.5-13B [25], and fine-tune it on LUCID for lunar analysis. The method follows a two-stage training strategy designed to first build a basic multimodal understanding and then teach the model how to answer questions about lunar observations more specifically.

3.2.1. Concept Alignment Stage

Stage 1 focuses on aligning lunar visual concepts with the language model representation space before introducing instruction-based reasoning. The goal of Stage 1 is to establish a stable multimodal grounding between lunar imagery and descriptive geological language.

Our architecture consists of three components (1) a pretrained vision encoder from pretrained CLIP-ViT-Large-Patch14 [35] that extracts patch-level image representations (2) a learnable multimodal projection layer that maps visual features into the hidden space of the language model (3) a pretrained autoregressive language model that performs text generation. During Stage 1, the pretrained vision encoder extracts patch-level visual features that encode the local terrain structure. These features are mapped into the language model embedding space through a trainable projection layer, allowing the model to associate planetary surface patterns with a domain-specific language. The vision

Visual Input: A Lunar Patch from LUCID Dataset

Caption: The dominant visible features are a cluster of small to medium impact craters scattered across a relatively smooth plains surface. Several craters in the lower part of the frame display sharp rims and well-defined interior walls, indicating locally steeper relief, whereas surrounding inter-crater areas appear comparatively flat. Subtle linear troughs and grooves trend diagonally across the scene, locally deflecting crater rims and ejecta. The combination of smooth plains and aligned structures may reflect a mechanically coherent, relatively dense substrate beneath a thin regolith cover.

User Question: What does the crater distribution suggest about the underlying substrate?

Instruction Tuned LLaVA-LE ✓	The crater distribution suggests a mechanically coherent, relatively dense substrate beneath a thin regolith cover.
Base LLaVA ✗	The crater distribution on the moon's surface suggests that the underlying substrate is likely composed of a mix of hard and soft materials. The presence of small craters indicates that the surface is not entirely hard and rocky, as these craters have formed due to impacts from smaller objects. The larger crater in the middle of the image, which is more prominent and surrounded by smaller craters, implies that the substrate may have varying levels of hardness and resistance to impacts. This diversity in crater sizes and distribution could be due to the moon's geological history, which has resulted in a complex and varied surface composition.
Chat GPT	The crater distribution, with sharp rims and well-defined interior walls, suggests that the underlying substrate is mechanically coherent and relatively dense. This is indicated by the presence of smooth plains and the alignment of structures, which may reflect a solid substrate beneath a thin regolith cover.
Gemini	The crater distribution, observed across a relatively smooth plains surface, contributes to the overall combination of features that may reflect a mechanically coherent, relatively dense substrate beneath a thin regolith cover.

Figure 5. **Qualitative comparison of LLaVA-LE visual chat and reasoning capabilities in instruction-tuning tasks.** Comparison between the response of LLaVA-13B model and our LlaVA-LE model in during instruction-tuning stage evaluation. The base LLaVA model often produces fluent but generic answers, occasionally lacking domain-specific terminology or geological interpretation. In some cases, the model hallucinates features that are not clearly implied by the scene or shifts toward broad discussions of lunar history rather than addressing the detail in the question. The Gemini and GPT are considered as the performance upper bound because the ground truth caption is fed into both of these models as context to generate the ground truth response.

encoder remains frozen throughout this stage in order to preserve its pretrained visual representation capacity. The base language model parameters are also kept frozen. However, lightweight low-rank adaptation (LoRA) modules are inserted into the transformer layers and optimized jointly with the multimodal projection layer. These low-rank matrices allow controlled adaptation to planetary domain language structure.

To establish domain-specific grounding, we train Stage 1 on 76k lunar image-caption pairs constructed from our curated planetary dataset. Each sample consists of a geo-referenced lunar surface image tile, along with a structured scientific caption describing terrain morphology, crater density, degradation state, slope distribution, and surface texture where applicable. Each image-caption pair is converted into an instruction-style format in which the model is prompted to describe the given image. The target output is the original structured caption. This formulation encourages the model to associate planetary surface features with domain language patterns and scientific terminology.

Training is performed using a causal language modeling objective. Given an input image I and caption tokens $y = \{y_1, \ldots, y_T\}$, we minimize the negative log-likelihood:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I)\,. \tag{1}$$

Here, $y_{<t}$ denotes all previously generated tokens and I represents the encoded visual features projected into the language embedding space. The probability distribution $P(\cdot)$ is modeled autoregressively by the language model, which is conditioned on both the visual representation and the prior tokens.

This stage can be interpreted as aligning visual representations of lunar surface structures with existing language model embeddings in the planetary science domain. At the end of this phase, the model can generate coherent and geologically consistent descriptions of lunar surface struc-

tures. However, it is not yet optimized for complex reasoning or conversational interaction. Those capabilities are introduced in Stage 2 through instruction tuning.

3.2.2. Instruction Tuning Stage

After alignment with the domain data, we fine-tune the model to operate as a planetary conversational assistant capable of answering scientific questions and performing reasoning over multimodal lunar inputs.

To activate instruction-following behavior, we train on our planetary instruction dataset containing more than 81k question-answer turns. Multiple instructions are associated with each image, allowing the model to learn different reasoning patterns grounded in the same visual context. The instructions span several categories, including surface properties, spatial relationships between geological features, geomorphological characterization, and interpretations of formation processes and surface evolution. This multi-turn supervision encourages the model to move beyond caption generation toward scientific reasoning.

Stage 2 is trained using the same causal language modeling objective as defined in Eq. (1). However, the supervision shifts from global caption generation to instruction-conditioned question answering. The model is trained to produce targeted, reasoning-based responses grounded in the same visual representations. The loss is applied only to the assistant response tokens, whereas the instruction tokens are treated as a conditioning context.

Stage 2 is initialized from the low-rank adaptation weights obtained after Stage 1. The LoRA modules injected into the transformer layers remain trainable, and the multimodal projection layer continues to update. The base language model parameters remain frozen. This allows the model to refine the domain adaptation while maintaining the stability of the pretrained backbone.

The multimodal projection layer continues to learn a mapping from the vision encoder feature dimension to the hidden dimension of the language model, while the LoRA modules refine internal attention representations, allowing the model to better integrate planetary terminology, geological structure descriptions, and cross-modal reasoning grounded in lunar imagery.

The two-stage design separates the foundational multimodal grounding from the deeper adaptation to reasoning. This separation improves training stability and allows visual alignment to converge before introducing more complex reasoning objectives.

4. Experiments

We evaluate LLaVA-LE in a controlled setting designed to assess multimodal instruction-following and complex reasoning. Our evaluation focuses on performance under open-ended planetary visual interpretation.

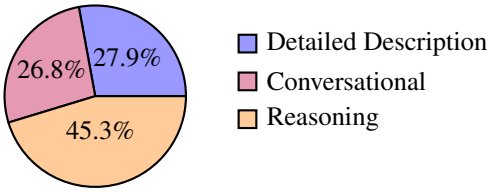


Figure 6. **Category distribution of the evaluation set**: Reasoning category is intentionally emphasized in the evaluation set because they pose a harder challenge beyond general visual description.

4.1. Evaluation Benchmark

To evaluate domain adaptation and reasoning ability, we constructed a held-out evaluation benchmark of 50 lunar patches from the LUCID stage 1 dataset. Using a language-only GPT-5.1, we generated 190 questions covering the categories shown in Figure 6. The language model is instructed to generate responses solely based on the scientific caption. This evaluation set was excluded from both stages of training LLaVA-LE.

As depicted in Figure 6, the reasoning category is intentionally emphasized because it requires interpretation of crater morphology, degradation state, slope distribution, feature relationships, and inferred geological processes. This category represents the most challenging subset and is designed to stress-test multimodal reasoning capability. These questions assess the model's ability to move beyond visual description and perform scientific interpretation.

4.2. Evaluation Protocol

To assess the impact of domain alignment and instruction tuning, we adopt an automatic evaluation protocol based on LLM judges, following the evaluation strategy introduced in LLaVA [25] and LLaVA-Med [20].

Candidate Model Responses. We compare three variants of the model. **Base LLaVA** is the original LLaVA-v1.5-13B model without any lunar-specific adaptation, trained on general-purpose multimodal instruction data, and not exposed to planetary science content. It serves as a strong general-domain multimodal baseline. **Concept-Aligned LLaVA-LE** is our Stage 1 model, trained on caption-based concept alignment data without instruction tuning, which learns to align lunar imagery with scientific descriptions. **Fully Trained LLaVA-LE** incorporates both Stage 1 alignment and Stage 2 instruction tuning on the LUCID VQA dataset, optimizing the model for multimodal scientific dialogue and reasoning. All models receive the raw lunar image and the corresponding question, without access to the scientific caption. Each model generates a free-form textual response based solely on visual content and the question.

Table 1. Performance Ratios Relative to GPT and Gemini Judges (Score / Judge Score)

	GPT Judge				Gemini Judge			
Model	Detailed	Conversation	Reasoning	**Overall**	Detailed	Conversation	Reasoning	**Overall**
LLaVA-LE Stage2	**0.952**	**0.767**	**1.006**	**0.925**	**0.892**	**0.628**	**1.134**	**0.917**
LLaVA-LE Stage1	0.433	0.419	0.480	0.450	0.367	0.362	0.536	0.436
Base LLaVA	0.319	0.314	0.344	0.329	0.220	0.205	0.246	0.227

Reference Answers. To establish a strong textual reference response for open-ended evaluation, we generated answers using GPT-5.1 and Gemini-2.5 from the ground truth captions in the LUCID dataset.

Automatic Judging. To improve robustness and reduce single-model bias, we employ two independent judges: ChatGPT [33] and Gemini [8]. For each question, an LLM judge receives the scientific caption, the reference answer, and the candidate model responses. Importantly, judges do not access the raw image during scoring; instead, evaluation is grounded in the scientific caption, which serves as a structured textual description of visual content. The agreement between a model's response and the reference answer, therefore, indicates successful visual grounding and domain-specific reasoning. Each judge assigns a score from 1 to 10 based on relevance, clarity, and accuracy.

4.3. Qualitative Analysis

To complement quantitative LLM-based scoring, we conduct qualitative analysis to better understand model behavior across question categories. We analyze performance of the base LLaVA model, the alignment-only model (Stage 1), and the fully adapted model (Stage 2) to illustrate differences. Figure 5 presents example responses for a selected lunar region.

Category-Level Performance. Table 1 reports relative performance across detailed, conversational, and reasoning question types. LLaVA-LE Stage 2 achieves an average overall score of 0.921 (averaged across GPT and Gemini judges), representing a **3.3**$\times$ **improvement** over Base LLaVA (0.278) and a **2.1**$\times$ **improvement** over LLaVA-LE Stage 1 (0.443). These gains are consistent across all categories, with Stage 2 scoring 0.922 on Detailed ($\sim$3.4$\times$ over Base LLaVA), 0.698 on Conversation ($\sim$2.7$\times$), and most notably **1.070 on Reasoning** ($\sim$**3.6**$\times$) — *exceeding the judge's own reference score* — demonstrating that Stage 2 instruction tuning is particularly effective at unlocking complex reasoning capabilities that neither Base LLaVA nor Stage 1 can approach. The consistent 2$\times$ margin of Stage 2 over Stage 1 across all categories further highlights the critical role of instruction tuning in transforming a capable vision-language backbone into a model that can engage in detailed, conversational, and reasoning-heavy visual question answering.

To ensure that the inferred subsurface interpretations were scientifically plausible, we manually cross-checked representative examples against independently available gravity and slope data for the corresponding regions. Specifically, we consulted Bouguer gravity anomaly maps derived from the GRAIL mission [18, 56] and terrain slope products computed from LOLA elevation data [40]. This inspection supports that the model's geological reasoning aligns with established lunar geophysical observations.

5. Conclusions

Recent advances in vision–language models have enabled powerful multimodal reasoning, yet their application to planetary science remains limited by the lack of large-scale datasets that pair planetary imagery with scientifically meaningful descriptions. In this work, we address this gap by introducing LUCID, a large-scale multimodal lunar dataset consisting of **96k** high-resolution panchromatic images paired with scientifically grounded captions, along with **LUCID-VQA**, a complementary benchmark containing **81k** question–answer pairs. Unlike prior efforts that rely on synthetic or limited datasets, LUCID is curated from real lunar mission products and incorporates geophysical slope and gravity information to ensure captions are grounded in geological observations. Leveraging this dataset, we develop **LLaVA-LE**, a domain-adapted vision–language model trained through a two-stage curriculum for domain concept alignment and instruction-tuned visual question answering. Experimental results show that **LLaVA-LE** substantially outperforms the base LLaVA model, achieving a **3.3× overall** improvement and strong gains in reasoning performance under both GPT and Gemini evaluation frameworks. These findings highlight the importance of domain-specific multimodal supervision for enabling vision–language models in scientific domains. By releasing the dataset, benchmarks, and model weights, we aim to facilitate future research at the intersection of computer vision, multimodal learning, and planetary science.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[3] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. 3

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3

[5] Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian conference on computer vision*, pages 2404–2420, 2024. 3

[6] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. 3

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90 https://vicuna.lmsys.org, 2023. 3

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 8

[9] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023. 3

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023. 3

[11] Matthew Foutter, Daniele Gammelli, Justin Kruger, Ethan Foss, Praneet Bhoj, Tommaso Guffanti, Simone D'Amico, and Marco Pavone. Space-llava: a vision-language model adapted to extraterrestrial applications. *arXiv preprint arXiv:2408.05924*, 2024. 2

[12] Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*, 2023. 3

[13] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023. 3

[14] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. 3

[15] Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, pages 1–39, 2025. 3

[16] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 3

[17] Shuaifeng Jiao, Zhiwen Zeng, Zhuoqun Su, Xieyuanli Chen, Zongtan Zhou, and Huimin Lu. Luseg: Efficient negative and positive obstacles segmentation via contrast-driven multi-modal feature fusion on the lunar. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5962–5969. IEEE, 2025. 2

[18] Alex S Konopliv, Ryan S Park, Dah-Ning Yuan, Sami W Asmar, Michael M Watkins, James G Williams, Eugene Fahnestock, Gerhard Kruizinga, Meegyeong Paik, Dmitry Strekalov, et al. High-resolution lunar gravity fields from the grail primary and extended missions. *Geophysical Research Letters*, 41(5):1452–1458, 2014. 8

[19] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 3, 7

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified

vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[23] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023. 3

[24] Yanda Li, Chi Zhang, Gang Yu, Wanqi Yang, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Enhanced visual instruction tuning with synthesized image-dialogue data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14512–14531, 2024. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3, 5, 7

[26] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023. 3

[27] Jiayi Liu, Qianyu Zhang, Xue Wan, Shengyang Zhang, Yaolin Tian, Haodong Han, Yutao Zhao, Baichuan Liu, Zeyuan Zhao, and Xubo Luo. Lusnar: A lunar segmentation, navigation and reconstruction dataset based on muti-sensor for autonomous exploration. *arXiv preprint arXiv:2407.06512*, 2024. 2

[28] Marcus Märtens, Kevin Farries, John Culton, and Tat-Jun Chin. Synthetic lunar terrain: A multimodal open dataset for training and evaluating neuromorphic vision algorithms. *arXiv preprint arXiv:2408.16971*, 2024. 2

[29] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3

[30] Henry J Moore. *Lunar remote sensing and measurements*. Number 1046. US Government Printing Office, 1980. 4

[31] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36:25081–25094, 2023. 3

[32] Amanda L Nahm, Thomas R Watters, Catherine L Johnson, Maria E Banks, Carolyn H van der Bogert, Renee C Weber, and Jeffrey C Andrews-Hanna. Tectonics of the moon. *Reviews in Mineralogy and Geochemistry*, 89(1):691–727, 2023. 4

[33] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, 2022. 3, 8

[34] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3, 5

[36] Mark S Robinson, SM Brylow, M ea Tschimmel, D Humm, SJ Lawrence, PC Thomas, Brett W Denevi, E Bowman-Cisneros, J Zerr, MA Ravine, et al. Lunar reconnaissance orbiter camera (lroc) instrument overview. *Space science reviews*, 150(1):81–124, 2010. 4

[37] Mark S Robinson, SM Brylow, M ea Tschimmel, D Humm, SJ Lawrence, PC Thomas, Brett W Denevi, E Bowman-Cisneros, J Zerr, MA Ravine, et al. Lunar reconnaissance orbiter camera (lroc) instrument overview. *SFpace science reviews*, 150(1):81–124, 2010. 3, 5

[38] DAVID H Scott. The geologic significance of some lunar gravity anomalies. In *In: Lunar Science Conference, 5th, Houston, Tex., March 18-22, 1974, Proceedings. Volume 3.(A75-39540 19-91) New York, Pergamon Press, Inc., 1974, p. 3025-3036.*, pages 3025–3036, 1974. 4

[39] VV Shevchenko, PC Pinet, SD Shevrel, Y Dadu, Ya Lu, TP Skobeleva, OI Kvaratskhelia, and K Rosemberg. Modern slope processes on the moon. *Solar System Research*, 46(1): 1–17, 2012. 4

[40] David E Smith, Maria T Zuber, Glenn B Jackson, John F Cavanaugh, Gregory A Neumann, Haris Riris, Xiaoli Sun, Ronald S Zellar, Craig Coltharp, Joseph Connelly, et al. The lunar orbiter laser altimeter investigation on the lunar reconnaissance orbiter mission. *Space science reviews*, 150(1): 209–241, 2010. 3, 8

[41] David E Smith, Maria T Zuber, Glenn B Jackson, John F Cavanaugh, Gregory A Neumann, Haris Riris, Xiaoli Sun, Ronald S Zellar, Craig Coltharp, Joseph Connelly, et al. The lunar orbiter laser altimeter investigation on the lunar reconnaissance orbiter mission. *Space science reviews*, 150(1): 209–241, 2010. 4

[42] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 11–23, 2023. 3

[43] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5034–5042, 2024. 3

[44] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 3

[45] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023. 3

[46] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023. 3

[47] Ying Wang, Xiaozhong Ding, Jian Chen, Kunying Han, Chenglong Shi, Ming Jin, Liwei Liu, Xinbao Liu, and Jiayin Deng. Interpretation of geological features and volcanic activity in the tsiolkovsky region of the moon. *Remote Sensing*, 16(6):1000, 2024. 4

[48] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023. 3

[49] Bang Yang, Yue Yu, Yuexian Zou, and Tong Zhang. Pclmed: Champion solution for imageclefmedical 2024 caption prediction challenge via medical vision-language foundation models. In *CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France*, 2024. 3

[50] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024. 3

[51] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685, 2023. 3

[52] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[53] Sheng Zhang, Jianzhong Liu, Gregory Michael, Kai Zhu, Danhong Lei, Jingyi Zhang, Jingwen Liu, and Man Ren. Detecting lunar linear structures based on multimodal semantic segmentation: the case of sinuous rilles. *Remote Sensing*, 16 (9):1602, 2024. 4

[54] Feizhong Zhou, Xingyue Liu, Qiao Zeng, Zhuhan Li, and Hanguang Xiao. Sigphi-med: A lightweight vision-language assistant for biomedicine. *Journal of Biomedical Informatics*, page 104849, 2025. 3

[55] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2304.10691*, 2023. 3

[56] Maria T Zuber, David E Smith, Michael M Watkins, Sami W Asmar, Alexander S Konopliv, Frank G Lemoine, H Jay Melosh, Gregory A Neumann, Roger J Phillips, Sean C Solomon, et al. Gravity field of the moon from the gravity recovery and interior laboratory (grail) mission. *Science*, 339(6120):668–671, 2013. 3, 8