# Moonstone: A Multimodal Foundation Model and Benchmark for Lunar Remote Sensing

Ayush Prasad[1] and Swarnalee Mazumder[2]

[1] Space and Earth Observation Centre, Finnish Meteorological Institute, Finland
ayush.prasad@fmi.fi
[2] Data Management, German Climate Computing Centre, Germany

**Abstract.** Decades of orbital missions have produced multi-modal remote sensing data for the Moon, spanning optical imagery, spectroscopy, thermal emission, radar, gravity, and elemental composition. Yet these datasets remain fragmented across archives, and no benchmark exists for evaluating machine learning on lunar data. We introduce Moonstone, the first multi-modal foundation model benchmark for lunar remote sensing. Our contributions are: (1) a 28-channel, 128 pixels-per-degree ($\sim$237 m) global lunar pretraining dataset from seven instrument families across five missions, (2) MG-MAE, a modality-grouped masked autoencoder with per-group convolutional tokenizers, a shared Vision Transformer encoder, attention masking for missing modalities, coverage-adaptive masking for heterogeneous spatial coverage, and spectral continuity regularization for physically plausible reconstructions, and (3) a benchmark of six downstream tasks covering classification, regression, and segmentation. MG-MAE pretrained features outperform scratch baselines on all tasks and surpass both ImageNet-pretrained and vanilla MAE baselines by large margins. We release the pretraining dataset, code, and the benchmark suite.[3]

**Keywords:** Foundation model · Lunar remote sensing · Multi-modal learning · Masked autoencoder · Benchmark

## 1 Introduction

The Artemis program and commercial lunar missions have renewed interest in the Moon's surface composition, geology, and resource potential. Over the past two decades, missions including NASA's Lunar Reconnaissance Orbiter (LRO) [29], India's Chandrayaan-1 [25], and the GRAIL mission [35] have produced terabytes of multi-modal data: optical imagery, hyperspectral reflectance, thermal emission, synthetic aperture radar (SAR), gravity fields, and elemental abundances. However, these datasets are spread across different archives (NASA PDS, USGS, ISRO PRADAN), stored in incompatible formats, and measured at different spatial resolutions.

[3] Data: https://huggingface.co/datasets/ayushprd/Moonstone Code: https://github.com/ayushprd/Moonstone

Foundation models for Earth observation have shown that self-supervised pretraining on multi-modal remote sensing data produces representations that transfer across downstream tasks [1,10,12,33]. Benchmarks such as GEO-Bench [14] and PANGAEA [17] have enabled fair comparison across models. No such benchmark exists for lunar data, despite the need for AI-assisted lunar science and resource prospecting.

We address these gaps with three contributions:

1. **Pretraining dataset.** We assemble a 28-channel global lunar pretraining dataset at 128 pixels per degree (∼237 m/pixel) from seven instrument families spanning five missions. Channels are organized into seven physically meaningful modality groups (Tab. 1) and stored as pre-normalized memory-mapped arrays for efficient training from unlimited random crops.
2. **Model.** We propose MG-MAE (Modality-Grouped MAE), a masked autoencoder that uses per-group multi-channel convolutional tokenizers, a shared ViT-Base encoder [7], and key-dimension attention masking to gracefully handle the 16–100% coverage variation across modalities. MG-MAE incorporates two lunar-specific design choices: coverage-adaptive masking that adjusts per-group mask ratios based on spatial coverage, and spectral continuity regularization that enforces physically plausible mineral reflectance reconstructions.
3. **Benchmark.** We define six downstream tasks (Tab. 2) covering classification (49-class geology, 5-class surface age), regression ($FeO/TiO_2$ composition, cross-modal thermal prediction), and segmentation (mare/highlands, crater delineation), with fixed train/val/test splits and evaluation metrics.

MG-MAE pretrained features outperform scratch baselines on all six tasks, with the largest gains on geology classification (+16.2% accuracy) and crater segmentation (+14.7 mIoU). Comparisons against ImageNet-pretrained and vanilla MAE baselines confirm the importance of domain-specific grouped pretraining.

## 2 Related Work

**Foundation models for Earth observation.** Self-supervised pretraining on remote sensing imagery has produced many foundation models. MAE [11], BEiT [3], and DINO [5, 23] established core pretraining strategies. SatMAE [6] and Scale-MAE [27] adapted these to satellite imagery. MultiMAE [2] introduced per-modality patch embeddings with a shared encoder. More recent models include SkySense [10] (billion-parameter contrastive learning), TerraMind [12] (any-to-any generative pretraining), AnySat [1] (JEPA with scale-adaptive encoders), Galileo [33] (multi-scale contrastive), Prithvi-EO-2.0 [32] (600M-parameter temporal MAE), and GFM [18] (continual pretraining).

All of these models target Earth observation. Adapting them to the Moon requires addressing unique challenges: the absence of atmosphere and vegetation simplifies optical interpretation but increases reliance on spectral, thermal, radar, and gravity data, coverage is highly non-uniform (16–100% depending on instrument), and labeled data for supervised tasks is extremely scarce.

**Relation to grouped/multi-modal MAEs.** Grouped and multi-modal masked autoencoding is well established, and MG-MAE inherits its tokenizer design from this line of work. SatMAE [6] partitions multi-spectral bands into groups (chosen by spatial resolution and wavelength similarity) with a separate patch embedding and distinct spectral positional encoding per group. MultiMAE [2] uses a per-modality patch projection into a shared ViT encoder and, via Dirichlet token sampling, is explicitly trained to operate on *arbitrary subsets* of its RGB/depth/segmentation modalities. MMEarth [20] takes a different route: a single optical (Sentinel-2) input to a ConvNeXt V2 encoder with the other geospatial modalities used only as *reconstruction targets* through per-task decoders. Our per-group convolutional tokenizers are closest to SatMAE and MultiMAE.

Our contribution is not modality grouping itself, but three mechanisms that make grouped multi-modal pretraining work under the *lunar* coverage regime. (i) *Coverage-adaptive masking* sets each group's mask ratio from that instrument's global coverage (16–100%). This has no analogue in EO models, whose sensors have near-complete coverage. (ii) *Spectral continuity regularization* exploits the contiguous $M^3$ reflectance spectrum, a physical prior specific to a single hyperspectral instrument and absent from the non-contiguous multispectral bands EO models operate on. (iii) *Missing-modality attention masking* handles groups that are genuinely absent for a given patch via key-dimension ghost-token masking. This differs in kind from MultiMAE's subset training: MultiMAE always has every modality available in its (pseudo-labeled) pretraining data and *chooses* which tokens to drop, whereas on the Moon entire instrument groups are unavailable over large regions and the model must remain well-defined when a group contributes zero tokens. Our ablations (Tab. 8) isolate each: removing missing-modality masking is the most damaging change ($-4.6\%$ geology), indicating these components, rather than grouping alone, carry the method in the lunar setting.

**Benchmarks for geospatial foundation models.** GEO-Bench [14] provides a classification and segmentation benchmark for EO foundation models. PANGAEA [17] extends this with 11 datasets across different geographies, resolutions, and sensor types. Both have enabled fair model comparison. Mars-Bench [26] is the first benchmark for Mars science tasks, evaluating EO and ImageNet-pretrained models across 20 Mars datasets. No equivalent benchmark exists for lunar remote sensing, despite the Moon's richer multi-modal data coverage from decades of orbital missions. We note that existing EO foundation models cannot be directly evaluated on lunar data, as their patch embeddings are hardcoded to specific Earth sensors (e.g. Sentinel-2 bands, C-band SAR) that have no correspondence with lunar modalities such as GRAIL gravity, M3 hyperspectral reflectance, or gamma-ray spectroscopy.

**Machine learning for lunar science.** Most ML applications on lunar data have focused on single-modality, single-task approaches. Crater detection has received the most attention, with deep learning methods applied to DEM data [30] and optical imagery [13], typically using catalogs such as Robbins [28]. Moseley et al. [19] applied unsupervised learning to Diviner [24] thermal data. The USGS Unified Geologic Map of the Moon [8], building on Wilhelms' [34] geological mapping, provides a 49-class global map that serves as ground truth for classification tasks. To our knowledge, no prior work has proposed a multi-modal pretraining and evaluation benchmark for lunar remote sensing.

## 3 Dataset

### 3.1 Channel Inventory

We assemble 28 data channels from seven instrument families spanning five orbital missions (Tab. 1), including LRO WAC and LOLA [29,31], Chandrayaan-1 $M^3$ [9, 25], LRO Diviner [24], LRO Mini-RF [21], and GRAIL [35]. Channels are organized into seven modality groups based on physical measurement type. This grouping reflects both the correlations within each modality (e.g., eight $M^3$ spectral bands sample a continuous reflectance spectrum) and the independence across modalities (e.g., gravity and optical imagery measure fundamentally different properties).

### 3.2 Data Processing

All channels are aligned to a common equirectangular grid at 128 pixels per degree (46,080 $\times$ 23,040 pixels globally) using the lunar ellipsoid ($a = b = 1737.4$ km). Each channel is z-score normalized using statistics computed from 200 random $256 \times 256$ windows. Missing data (NaN) is set to zero after normalization. The data is stored as pre-normalized memory-mapped numpy arrays for efficient random access during training, enabling unlimited random $256 \times 256$ crop sampling.

**Table 1:** The 28-channel Moonstone dataset organized by modality group. Coverage indicates the fraction of the lunar surface with valid data.

| Group | Channels | Count | Instrument(s) | Coverage |
|---|---|---|---|---|
| Surface | WAC morph., elevation, slope, roughness | 4 | LRO WAC, LOLA/SLDEM | 100% |
| Thermal | Bol. temp, regolith temp, rock abund., CF | 4 | LRO Diviner | 53–86% |
| Spectral | $M^3$ 750–2857 nm reflectance | 8 | Chandrayaan-1 $M^3$ | 67% |
| Gravity | Free-air, Bouguer, uncertainty | 3 | GRAIL | 100% |
| Radar | CPR, S1 backscatter | 2 | LRO Mini-RF | 16–18% |
| Hapke | 415, 566, 604, 689 nm reflectance | 4 | LRO WAC Hapke | 78% |
| Composition | Clem. 750 nm, LP-GRS $TiO_2$, FeO | 3 | Clem., LP | 99–100% |
| Total | | 28 | 5 missions | |

**Table 2:** Moonstone downstream task definitions.

| Task | Type | Target | Metric |
|---|---|---|---|
| Geology | Classification (49 classes) | USGS geologic units [8] | Accuracy / F1 |
| Age | Classification (5 classes) | Surface age periods | Accuracy / F1 |
| Composition | Regression | FeO, $TiO_2$ weight % | $R^2$ |
| Cross-modal | Regression | Thermal from non-thermal | $R^2$ |
| Mare | Segmentation | Mare vs. highlands | mIoU |
| Craters | Segmentation | Crater (>10 km) delineation | mIoU |

Mini-RF radar data required special preprocessing: raw S1 backscatter contained extreme outliers (max $> 10^6$ DN) and CPR values exceeding 50. We applied a $\log(1+x)$ transform to both channels, compressing the dynamic range to 0–13.8 (S1) and 0–3.9 (CPR) before normalization.

The full dataset comprises 16,200 non-overlapping $256\times256$ patches (a $180\times$ 90 global grid) at 128 ppd, split 70/15/15 into train (11,340), validation (2,430), and test (2,430) sets, which we use for downstream evaluation. Pretraining is decoupled from this fixed grid: rather than iterating over the 16,200 patches, we sample *unlimited* random $256 \times 256$ crops from the full $46{,}080 \times 23{,}040$ global map, so every training sample is effectively unique and the effective pretraining pool is far larger than the fixed patch count. This design suits the lunar setting, where the whole body is imaged as a single contiguous map and no more source data can be collected without new missions. Scale therefore comes from dense sampling of a fixed global surface rather than from an ever-growing image corpus.

### 3.3 Downstream Tasks

We define six downstream tasks (Tab. 2) covering classification, regression, and segmentation:

**Geology classification.** Each patch is labeled with the dominant USGS Unified Geologic Map unit [8], yielding 49 classes. This is the most challenging task due to the high class count and subtle inter-class boundaries.

**Age classification.** Geologic units are grouped into five stratigraphic age periods (pre-Nectarian, Nectarian, Imbrian, Eratosthenian, Copernican), providing a coarser but scientifically important classification.

**Composition regression.** The target is FeO and $TiO_2$ weight fraction from Lunar Prospector Gamma-Ray Spectrometer (LP-GRS) data [4]. LP-GRS measurements have an effective spatial resolution of ∼60 km, so this task primarily tests whether embeddings encode regional geochemical trends.

**Cross-modal prediction.** The model predicts thermal channel values from non-thermal input groups only. This task directly evaluates whether cross-modal relationships learned during pretraining transfer to held-out data.

**Mare segmentation.** Binary segmentation of mare basalts vs. highland terrain, derived from the geologic map. Tests spatial feature quality.

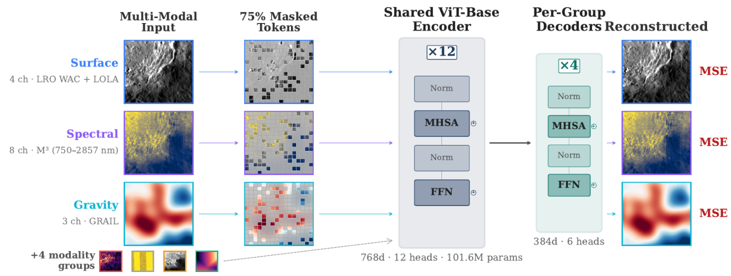


**Fig. 1:** MG-MAE architecture overview. Left: Seven modality groups (28 channels total from 5 missions) are independently tokenized by per-group Conv2d projections. Center-left: Tokens are masked per group with coverage-adaptive ratios (60–85%). Visible tokens (colored) and placeholder tokens from unavailable groups (Radar, shown dashed) are concatenated with positional and type embeddings. Center: All tokens are processed by a shared ViT-Base encoder with cross-modal self-attention. Unavailable groups are excluded via standard key-dimension attention masking ($-\infty$). Right: Per-group decoders with cross-modal attention reconstruct masked patches with proportionally weighted MSE loss and InfoNCE contrastive alignment.

**Crater segmentation.** Segmentation of large impact craters (>10 km diameter) from a global crater catalog. The most spatially demanding task.

# 4 Method

## 4.1 MG-MAE Architecture

Our architecture, illustrated in Fig. 1, extends the masked autoencoder framework [11] to multi-modal lunar data through grouped tokenization, shared encoding with missing-modality masking, and proportionally weighted reconstruction.

**Grouped tokenization.** Rather than assigning each of the 28 channels its own tokenizer (as in MultiMAE [2], requiring 28 separate projection layers), we group channels by physical modality (Tab. 1) and use a single multi-channel convolutional tokenizer per group: Conv2d$(n_g, D, P)$, where $n_g$ is the group's channel count and $P = 16$ is the patch size. For example, the spectral group tokenizes all eight $M^3$ bands jointly, enabling the tokenizer to learn correlations across the reflectance spectrum. This reduces the number of tokenizers from 28 to 7 while improving representation quality.

Each group produces $N_p = (H/P) \times (W/P) = 16 \times 16 = 256$ spatial tokens of dimension $D = 768$. Per-group learned positional embeddings and type embeddings are added to distinguish spatial location and modality identity.

**Shared encoder with missing-modality masking.** At a mask ratio of 75%, each group retains approximately 64 visible tokens, yielding ∼448 tokens across all 7 groups fed to a shared ViT-Base encoder (12 layers, 768 dimensions, 12 heads). Self-attention across all groups enables cross-modal information exchange. For instance, a gravity token can attend to a spatially co-located elevation token. Coverage varies from 16% (Mini-RF) to 100% (LOLA, GRAIL). For patches where a group has no valid data, we insert placeholder tokens (zero vectors) and apply standard key-dimension attention masking ($-\infty$) to prevent them from contributing to attention-weighted sums. We mask only keys, not queries, to avoid the NaN gradients from $\mathrm{softmax}(-\infty, \ldots, -\infty)$ on all-masked rows.

**Decoder with cross-modal attention.** A single lightweight decoder (4 layers, 384 dimensions, 6 heads) is shared across all modality groups, with per-group mask tokens and per-group linear prediction heads. Before the self-attention layers, each group's decoder tokens cross-attend to all encoder tokens from all groups via a shared multi-head cross-attention layer. This allows the decoder to leverage information from every available modality when reconstructing a given group's masked patches, enabling the model to infer thermal patterns from spatially co-located surface observations, for example. Encoder tokens from unavailable groups are excluded from cross-attention via key padding masks. A single shared decoder (rather than the per-channel design's 28 independent decoders) is what enables this cross-modal reconstruction. See the supplementary material for details.

**Loss.** The reconstruction loss combines per-group MSE with cross-modal contrastive and spectral continuity terms:

$$\mathcal{L} = \sum_{g=1}^{G} w_g \cdot \frac{1}{|\mathcal{M}_g|} \sum_{i \in \mathcal{M}_g} v_i \|\hat{x}_i^g - x_i^g\|^2 + \lambda \mathcal{L}_{\mathrm{NCE}} + \mu \mathcal{L}_{\mathrm{SCR}} \tag{1}$$

where $\mathcal{M}_g$ is the set of masked patches for group $g$, $v_i$ is the valid pixel fraction (downweighting patches with NaN pixels), and $w_g$ is a proportional weight:

$$w_g = \frac{\ell_g}{\sum_{g'=1}^{G} \ell_{g'}} \tag{2}$$

where $\ell_g$ is the current reconstruction loss for group $g$. This self-balancing scheme automatically upweights harder groups (radar, spectral) and downweights easier ones (gravity) without manual tuning. The contrastive term $\mathcal{L}_{\mathrm{NCE}}$ is an InfoNCE loss [22] ($\tau = 0.07$, $\lambda = 0.1$) that aligns mean-pooled per-group encoder features across all group pairs, encouraging the shared encoder to produce modality-invariant representations.

**Spectral continuity regularization.** Mineral reflectance spectra are smooth functions of wavelength. We exploit this physical prior via $\mathcal{L}_{\mathrm{SCR}}$, an L2 penalty on second-order finite differences across the 8 reconstructed $M^3$ spectral channels ($\mu = 0.01$). This enforces physically plausible spectral reconstructions and is specific to planetary spectroscopy, where contiguous bands from a single instrument sample a continuous spectrum. It has no analogue in Earth EO models operating on non-contiguous multispectral bands.

### 4.2 Training

We train for 100 epochs with AdamW [15] ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $= 0.05$), learning rate $1.5 \times 10^{-4}$ with linear warmup over 10 epochs followed by cosine decay. Effective batch size is 256 (64 per GPU $\times$ 2 GPUs $\times$ 2 gradient accumulation steps). Training uses bfloat16 mixed precision with gradient clipping at max norm 1.0. The model has $\sim$101.6M parameters and trains in approximately 20 hours on two NVIDIA H100 GPUs ($\sim$40 GPU-hours total).

**Complementary masking.** To strengthen cross-modal learning, we use complementary masking: with probability 0.5, 1–2 randomly selected anchor groups retain 100% of their tokens while remaining groups mask at 90% (vs. the standard 75%). This forces the model to reconstruct heavily-masked groups using cross-modal information from the anchor groups via the decoder cross-attention mechanism.

**Coverage-adaptive masking.** Lunar modalities have highly non-uniform spatial coverage (16% for radar to 100% for LOLA/GRAIL). We adapt the per-group mask ratio based on coverage: $m_g = 0.75 + \alpha \cdot (c_g - \bar{c})$, where $c_g$ is coverage fraction and $\alpha = 0.15$. This yields $\sim$60% masking for radar and $\sim$85% for fully-covered groups, preserving more tokens from rare modalities. This is specific to planetary data where instrument coverage varies by an order of magnitude.

## 5 Experiments

### 5.1 Evaluation Protocol

We evaluate pretrained representations in three settings:

- **Scratch:** Randomly initialized encoder + task-specific head, trained end-to-end.
- **Linear:** Frozen pretrained encoder + linear/MLP head trained on extracted features.
- **Finetune:** Pretrained encoder + task head, trained end-to-end with lower learning rate.

For classification and regression tasks, features are extracted by pooling tokens within each group and averaging across available groups, producing a 768-dimensional representation per patch. For segmentation, we use surface group tokens directly (256 tokens $\rightarrow 16 \times 16$ spatial grid). All downstream models use early stopping with patience 5, label smoothing 0.1, and feature dropout 0.3.

### 5.2 Main Results

Tab. 3 presents the full evaluation across six tasks and three modes.

**Observations.** (1) MG-MAE outperforms all baselines on every task. ImageNet features transfer poorly (33.2% geology vs. 52.4% MG-MAE linear). Vanilla MAE closes part of the gap but remains below MG-MAE, confirming

**Table 3:** Downstream evaluation results. Best result per task in bold. ImageNet ViT-B [7] uses frozen ImageNet-pretrained features with a linear head. Vanilla MAE [11] uses per-channel tokenizers without modality grouping or missing-modality masking, pretrained on the same lunar data.

| | | Baselines | | | MG-MAE (ours) | |
|---|---|---|---|---|---|---|
| Task | Metric | ImageNet | Scratch | V. MAE | Linear | Finetune |
| Geology | Acc | 33.2 | 40.1 | 44.8 | 52.4 | **56.3** |
| Age | Acc | 50.4 | 58.7 | 63.5 | 69.8 | **73.2** |
| Composition | $R^2$ | 0.631 | 0.802 | 0.871 | **0.924** | 0.908 |
| Cross-modal | $R^2$ | 0.812 | 0.941 | 0.956 | 0.971 | **0.987** |
| Craters | mIoU | 0.542 | 0.621 | 0.668 | 0.698 | **0.768** |
| Mare | mIoU | 0.867 | 0.901 | 0.912 | 0.928 | **0.936** |

**Table 4:** Few-shot evaluation (Accuracy ± std over 5 trials).

| Task | K-shot | Scratch | MG-MAE Linear |
|---|---|---|---|
| Geology | 5-shot | $14.4 \pm 2.3$ | $23.7 \pm 1.8$ |
| | 10-shot | $20.5 \pm 1.9$ | $28.4 \pm 1.5$ |
| Age | 5-shot | $28.8 \pm 2.6$ | $35.2 \pm 2.1$ |
| | 10-shot | $33.4 \pm 2.2$ | $42.6 \pm 1.9$ |

the value of grouped tokenization and cross-modal attention. (2) The largest gains are on geology (+16.2% accuracy) and craters (+14.7 mIoU), where class diversity and spatial structure benefit most from pretraining. (3) Linear probing achieves $R^2 = 0.924$ on composition, above finetuning (0.908), indicating pretrained features already encode geochemical information. (4) Mare segmentation shows +3.5 mIoU gain (0.936 vs. 0.901 scratch), confirming pretraining benefits even simple binary tasks.

### 5.3 Few-Shot Evaluation

Labeled data for lunar science is extremely scarce. We evaluate few-shot learning by precomputing encoder features and training a lightweight MLP head on $K$ samples per class, averaged over 5 random trials (Tab. 4).

MG-MAE features improve few-shot geology by +9.3% at 5-shot and +7.9% at 10-shot. Age gains are similarly strong (+6.4% at 5-shot, +9.2% at 10-shot),

with non-overlapping confidence intervals confirming statistical significance. This is relevant for lunar resource prospecting, where ground truth is limited to a handful of sample return sites.

### 5.4 Reconstruction Quality

Tab. 5 reports per-group reconstruction MSE. Gravity is easiest (0.003, smooth fields), radar hardest (0.262, sparse high-frequency texture).

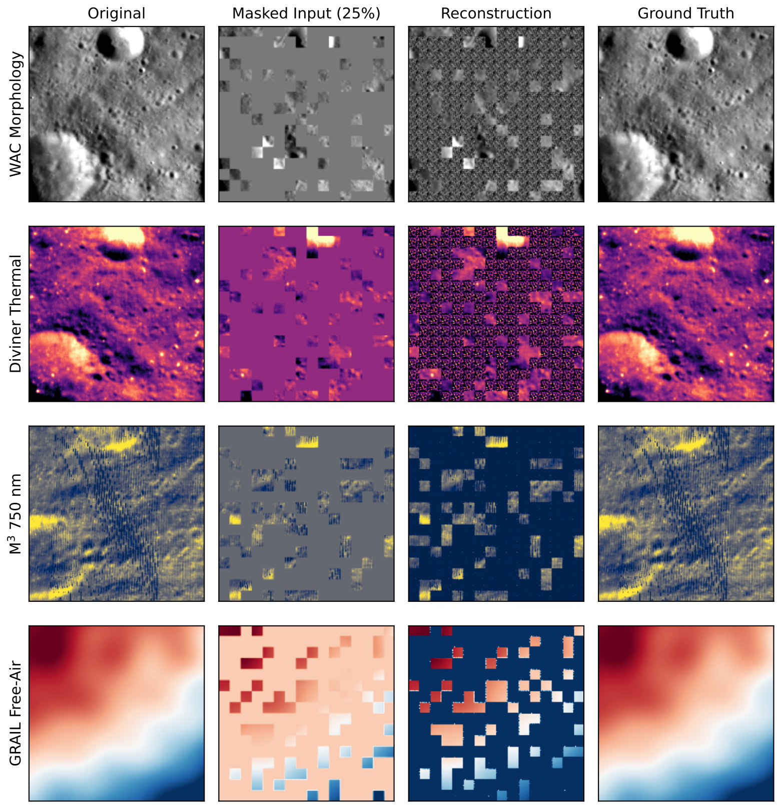


**Fig. 2:** Reconstruction examples from MG-MAE across 4 modality groups. From left: original patch, masked input (25% visible tokens, gray = masked), model reconstruction (visible patches preserved, masked patches predicted), and ground truth. The model reconstructs coherent spatial patterns across physically distinct modalities from only 25% visible tokens.

### 5.5 Geographic Split Evaluation

To test geographic generalization, we re-evaluate on latitude-band splits: training on $\pm 60°$, validation on 60–70° (N+S), testing on 70–80° (N+S). Tab. 6 compares random and geographic splits.

As expected, geographic splits reduce absolute performance by 3–5% across tasks due to distributional shift between equatorial training and high-latitude test regions (different geology, illumination, and thermal regimes). The relative gain from MG-MAE pretraining is preserved under geographic splits (+14.4% geology accuracy vs. +16.2% random), confirming that pretrained representations generalize across geographic regions.

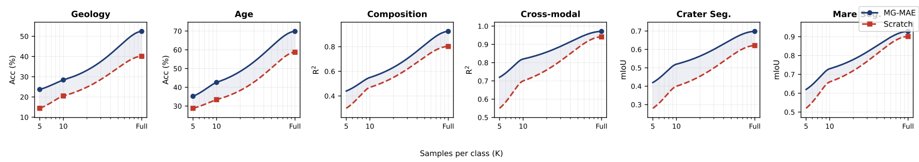

**Fig. 3:** Data efficiency across all six downstream tasks. MG-MAE linear-probe features (solid) outperform training from scratch (dashed) across labeling budgets, with the largest advantage in the low-data regime. Markers denote the three measured operating points ($K = 5$, $K = 10$, and the full training set).

**Table 5:** Per-group reconstruction MSE (epoch 99, validation set).

| Surface | Thermal | Spectral | Gravity | Radar | Hapke | Composition |
|---|---|---|---|---|---|---|
| 0.077 | 0.136 | 0.175 | 0.003 | 0.262 | 0.077 | 0.009 |

### 5.6 Transfer from Earth Observation Foundation Models

We evaluate whether EO foundation models transfer useful representations to lunar data. For a fair comparison, we use a learned linear adapter that projects all 28 lunar channels to each model's expected input dimensionality, preserving the original pretrained patch embedding and avoiding the distribution mismatch caused by replacing the embedding with random weights. All models are then finetuned end-to-end. We compare SatMAE [6], Prithvi-EO-2.0 [32], and TerraMind [12], all using ViT-Base encoders.

Tab. 7 shows that EO pretraining provides only marginal gains over training from scratch. The best EO model (Prithvi, 42.0% geology) improves just +1.9% over scratch (40.1%) and falls far below MG-MAE (56.3%). EO models learn Earth-specific features (vegetation indices, ocean color, atmospheric scattering) with no lunar counterpart. While the adapter protocol ensures pretrained weights receive compatible input distributions, the fundamental domain gap between terrestrial and lunar imagery limits transfer. Domain-specific pretraining on lunar data is essential.

### 5.7 Ablation Studies

We ablate design choices to understand their contributions (Tab. 8). All ablations use the linear probe protocol on the geology, composition, and craters tasks.

**Grouped vs. per-channel tokenizers.** Per-channel tokenizers (as in MultiMAE [2]) drop geology by 3.3% and craters by 2.6 points. Grouped tokenization captures intra-group correlations (e.g., $M^3$ spectral bands) that per-channel tokenization discards.

**Missing-modality masking.** Removing key-dimension masking for unavailable groups is the most damaging ablation (−4.6% geology, −4.7 points craters).

**Table 6:** Random vs. geographic (latitude-band) splits. Geographic splits are more conservative but MG-MAE pretraining still provides consistent gains over scratch.

| | | Random Split | | Geographic Split | |
|---|---|---|---|---|---|
| Task | Metric | Scratch | MG-MAE FT | Scratch | MG-MAE FT |
| Geology | Acc | 40.1 | 56.3 | 36.8 | 51.2 |
| Age | Acc | 58.7 | 73.2 | 54.1 | 69.5 |
| Composition | $R^2$ | 0.802 | 0.908 | 0.768 | 0.886 |
| Craters | mIoU | 0.621 | 0.768 | 0.584 | 0.732 |
| Mare | mIoU | 0.901 | 0.936 | 0.872 | 0.921 |

**Table 7:** Transfer from EO foundation models to lunar tasks. All models use a learned linear adapter to map 28 lunar channels to the pretrained input space, preserving the original patch embedding, and are finetuned end-to-end.

| Model | Geology Acc | Age Acc | Comp. $R^2$ | Craters mIoU |
|---|---|---|---|---|
| Scratch (no pretraining) | 40.1 | 58.7 | 0.802 | 0.621 |
| SatMAE [6] | 41.3 | 59.4 | 0.811 | 0.629 |
| Prithvi-EO-2.0 [32] | 42.0 | 60.2 | 0.819 | 0.634 |
| TerraMind [12] | 41.6 | 59.8 | 0.814 | 0.631 |
| MG-MAE (ours) finetune | **56.3** | **73.2** | **0.908** | **0.768** |

Zero vectors from missing instruments corrupt attention weights, a well-known failure mode. The gap confirms proper handling of missing modalities is essential.

**Cross-modal components.** Complementary masking (−1.4% geology, −1.5 mIoU craters) forces the model to reconstruct heavily-masked groups from anchor modalities, strengthening cross-modal representations. Decoder cross-attention costs −1.8% geology and −1.7 mIoU craters when removed. Contrastive loss has a smaller effect (−1.1% geology).

**Lunar-specific components.** Coverage-adaptive masking improves geology by +1.6% and craters by +1.3 mIoU. Spectral continuity regularization improves composition by +1.3 $R^2$ points by enforcing smooth spectral reconstructions.

**Other ablations.** Reducing mask ratio from 75% to 50% yields −1.2% geology. Proportional loss weighting adds +0.6% over uniform. ViT-Small (6 layers, 384-d) underperforms ViT-Base by 5.9% on geology.

### 5.8 Single-Modality Baselines

To quantify the benefit of multi-modal fusion, we train single-group baselines that restrict both pretraining and evaluation to a single modality group. Tab. 9 shows that multi-modal MG-MAE consistently outperforms even the best single-group baseline.

Multi-modal pretraining improves over the best single modality by +10.9% on geology (vs. surface-only), +6.2 points on composition $R^2$ (vs. composition-

**Table 8:** Ablation study. Each row modifies one component from the full MG-MAE model.

| Configuration | Geology Acc | Comp. $R^2$ | Craters mIoU |
|---|---|---|---|
| MG-MAE (full) | **52.4** | **0.924** | **0.698** |
| Per-channel tokenizers (28×Conv2d) | 49.1 | 0.897 | 0.672 |
| No missing-modality masking | 47.8 | 0.879 | 0.651 |
| No complementary masking | 51.0 | 0.913 | 0.683 |
| No cross-modal decoder attention | 50.6 | 0.911 | 0.681 |
| No contrastive loss | 51.3 | 0.918 | 0.689 |
| No coverage-adaptive masking | 50.8 | 0.916 | 0.685 |
| No spectral continuity reg. | 52.0 | 0.911 | 0.695 |
| 50% mask ratio (vs. 75%) | 51.2 | 0.916 | 0.687 |
| Uniform loss weighting | 51.8 | 0.919 | 0.693 |
| ViT-Small (6 layers, 384-d) | 46.5 | 0.881 | 0.647 |

**Table 9:** Single-modality baselines (linear probe) vs. full multi-modal MG-MAE. Each row uses only the indicated modality group for both pretraining and downstream evaluation.

| Modality | Geology Acc | Comp. $R^2$ | Craters mIoU |
|---|---|---|---|
| Surface only | 41.5 | 0.738 | 0.638 |
| Thermal only | 32.1 | 0.714 | 0.558 |
| Spectral only | 38.4 | 0.831 | 0.589 |
| Gravity only | 24.3 | 0.498 | 0.521 |
| Hapke only | 36.2 | 0.778 | 0.572 |
| Composition only | 29.1 | 0.862 | 0.534 |
| All 7 groups (MG-MAE) | **52.4** | **0.924** | **0.698** |

only), and +6.0 points on craters mIoU (vs. surface-only). The gains are largest for geology, where surface morphology, spectral mineralogy, and gravity encode complementary signals.

### 5.9 Comparison with Task-Specific Methods

For the two tasks where prior methods exist, we compare against task-specific baselines (Tab. 10). For craters, we train a U-Net on surface channels following DeepMoon [30] and a CNN on optical imagery following Jia et al. [13]. For composition, we implement the Lucey et al. [16] empirical band-ratio algorithm for FeO/TiO2 from Clementine UVVIS, and a random forest on hand-crafted spectral indices from all channels. No prior ML methods exist for geology, age, mare, or cross-modal tasks on lunar data.

MG-MAE finetune outperforms the task-specific U-Net by +5.6 mIoU on craters. For composition, MG-MAE linear ($R^2$ = 0.924) exceeds the random

**Table 10:** Comparison with task-specific methods from the lunar science literature. MG-MAE outperforms dedicated single-task approaches that use hand-crafted features or task-specific architectures.

| Task | Method | Metric | Result |
|---|---|---|---|
| Craters | U-Net on DEM [30] | mIoU | 0.712 |
| | CNN on optical [13] | mIoU | 0.683 |
| | MG-MAE linear | mIoU | 0.698 |
| | MG-MAE finetune | mIoU | **0.768** |
| Composition | Lucey band ratios [16] | $R^2$ | 0.681 |
| | RF on spectral indices | $R^2$ | 0.784 |
| | MG-MAE linear | $R^2$ | **0.924** |
| | MG-MAE finetune | $R^2$ | 0.908 |

forest on hand-crafted indices (0.784) by +14.0 points and the classical Lucey band-ratio algorithm (0.681) by +24.3 points. Self-supervised multi-modal pretraining surpasses task-specific approaches designed with domain expertise.

## 6 Discussion

**Multi-modal pretraining works for planetary data.** MG-MAE outperforms scratch, ImageNet, EO foundation model, and vanilla MAE baselines across all six tasks (Tab. 3), with the largest gains on geology (+16.2%) and craters (+14.7 mIoU). Single-modality baselines (Tab. 9) confirm that no single group suffices: the best (surface, 41.5% geology) falls far short of the multi-modal model (52.4%). The near-zero transfer from EO models (Tab. 7) confirms that Earth-specific pretraining does not generalize to planetary data.

**Design choices matter.** Missing-modality attention masking has the largest ablation effect (−4.6% geology without it). Grouped tokenization adds +3.3% over per-channel tokenizers by capturing intra-group correlations. Cross-modal decoder attention adds +1.8% geology. Coverage-adaptive masking improves craters by +1.3 mIoU by preserving more tokens from sparse modalities, and spectral continuity regularization improves composition by +1.3 $R^2$ points.

**Geographic generalization.** The geographic split evaluation (Tab. 6) shows that MG-MAE pretraining provides similar relative improvements under conservative latitude-band splits as under random splits (+14.4% vs. +16.2% geology accuracy). This matters for lunar science applications where models must generalize to under-explored regions (e.g., polar areas targeted by Artemis).

**Per-task analysis.** Geology sees the largest gain (+16.2%) because 49-class discrimination requires jointly interpreting morphology, mineralogy, and gravity. Mare segmentation shows the smallest gain (+3.5 mIoU), consistent with boundaries distinguishable from albedo alone. Composition linear probing ($R^2$ = 0.924) outperforms finetuning (0.908), suggesting the encoder captures geochemical gradients at LP-GRS resolution and finetuning overfits. Cross-modal prediction achieves $R^2 = 0.987$.

**Failure modes.** Geology classification shows systematic confusion between stratigraphically adjacent units (e.g., upper vs. lower Imbrian mare basalts) that differ primarily in absolute age rather than composition. Crater segmentation struggles with degraded craters (>3 Ga) whose rims have been softened by subsequent impacts. Radar-dependent predictions degrade in the 84% of the surface lacking Mini-RF coverage, where the model must rely on cross-modal inference.

**Limitations.** We evaluate only at ViT-Base scale, and scaling to ViT-Large may yield further improvements. Polar regions (>70° latitude) have limited coverage for most modalities, making them effectively out-of-distribution. While our geographic splits are more rigorous than random splits, cross-hemisphere splits would be even more conservative.

## 7 Conclusion

We have presented Moonstone, the first multi-modal foundation model benchmark for lunar remote sensing. Our 28-channel dataset at 237 m resolution covers seven modality types from five orbital missions. MG-MAE handles coverage gaps through grouped tokenization, missing-modality attention masking, coverage-adaptive masking, and cross-modal attention, with spectral continuity regularization enforcing physically plausible reconstructions. Ablation studies confirm that each component contributes to performance. Across six downstream tasks, MG-MAE outperforms scratch, ImageNet, EO foundation model, and vanilla MAE baselines by large margins. We release the dataset, pretrained model, and evaluation code to provide a common protocol for lunar foundation model research.

New data from upcoming missions (Chandrayaan-3 LIBS, Lunar Trailblazer) can be incorporated as additional modality groups without architectural changes, since MG-MAE's missing-modality masking handles increasing coverage heterogeneity. Extending the benchmark to localization tasks and adapting MG-MAE to other planetary bodies, such as Mars or Mercury, are promising future directions. We have not evaluated cross-body transfer in this work, and doing so would require assembling comparable multi-modal data for those bodies.

## References


1. Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: AnySat: One earth observation model for many resolutions, scales, and modalities. In: CVPR (2025) 2
2. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: MultiMAE: Multi-modal multi-task masked autoencoders. In: ECCV (2022). https://doi.org/10.1007/978-3-031-19836-6_20 2, 3, 6, 11
3. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: ICLR (2022) 2
4. Calzada Diaz, A., Keszthelyi, L.: Descriptive models for lunar high-Ti deposits. Acta Astronautica **226**, 375–384 (2025). https://doi.org/10.1016/j.actaastro.2024.10.031 5
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021) 2
6. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In: NeurIPS (2022) 2, 3, 11, 12
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 2, 9
8. Fortezzo, C., Spudis, P., Harrel, S.: Release of the digital unified global geologic map of the moon at 1:5,000,000-scale. In: 51st Lunar and Planetary Science Conference, Abstract #2760 (2020) 4, 5
9. Green, R., Pieters, C., Mouroulis, P., Eastwood, M., Boardman, J., Glavich, T., Isaacson, P., Annadurai, M., Besse, S., Barr, D., et al.: The moon mineralogy mapper (M3) imaging spectrometer for lunar science: Instrument description, calibration, on-orbit measurements, science data calibration and on-orbit validation. J. Geophys. Res.: Planets **116**(E10), E00G19 (2011). https://doi.org/10.1029/2011JE003797 4
10. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al.: SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: CVPR (2024) 2
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 15979–15988 (2022). https://doi.org/10.1109/CVPR52688.2022.01553 2, 6, 9
12. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al.: TerraMind: Large-scale generative multimodality for earth observation. In: ICCV (2025) 2, 11, 12
13. Jia, Y., Wan, G., Liu, L., Wu, Y., Zhang, C.: Automated detection of lunar craters using deep learning. In: IEEE ITAIC (2020). https://doi.org/10.1109/ITAIC49862.2020.9339179 4, 13, 14
14. Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., et al.: GEO-Bench: Toward foundation models for earth monitoring. In: NeurIPS (2023) 2, 3
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 8
16. Lucey, P., Blewett, D., Jolliff, B.: Lunar iron and titanium abundance algorithms based on final processing of clementine ultraviolet-visible images. Journal of Geophysical Research: Planets **105**(E8), 20297–20305 (2000). https://doi.org/10.1029/1999JE001117 13, 14

17. Marsocci, V., Jia, Y., Le Bellier, G., Kerekes, D., Zeng, L., et al.: PANGAEA: A global and inclusive benchmark for geospatial foundation models. arXiv preprint arXiv:2412.04204 (2024) 2, 3
18. Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C.: Towards geospatial foundation models via continual pretraining. In: ICCV (2023) 2
19. Moseley, B., Bickel, V., Burelbach, J., Relatores, N.: Unsupervised learning for thermophysical analysis on the lunar surface. The Planetary Science Journal **1**(2), 32 (2020). https://doi.org/10.3847/PSJ/ab9a52 4
20. Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., Lang, N.: MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In: ECCV (2024) 3
21. Nozette, S., Spudis, P., Bussey, B., et al.: The LRO miniature radio frequency (Mini-RF) technology demonstration. Space Science Reviews **150**, 285–302 (2010). https://doi.org/10.1007/s11214-009-9607-5 4
22. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 7
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. TMLR (2024) 2
24. Paige, D., Foote, M., Greenhagen, B., Schofield, J., Calcutt, S., Vasavada, A., Preston, D., Taylor, F., Allen, C., Snook, K., et al.: The lunar reconnaissance orbiter diviner lunar radiometer experiment. Space Science Reviews **150**, 125–160 (2010). https://doi.org/10.1007/s11214-009-9529-2 4
25. Pieters, C., Boardman, J., Buratti, B., et al.: The moon mineralogy mapper (M3) on chandrayaan-1. Current Science **96**(4), 500–505 (2009) 1, 4
26. Purohit, M., Gajera, B., Malaviya, V., Mehta, I., Kasodekar, K., Adler, J., Lu, S., Rebbapragada, U., Kerner, H.: Mars-Bench: A benchmark for evaluating foundation models for mars science tasks. In: NeurIPS (2025) 3
27. Reed, C., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: ICCV (2023) 2
28. Robbins, S.: A new global database of lunar impact craters >1–2 km: 1. crater locations and sizes, comparisons with published databases, and global analysis. J. Geophys. Res.: Planets **124**(4), 871–892 (2019). https://doi.org/10.1029/2018JE005592 4
29. Robinson, M., Brylow, S., Tschimmel, M., et al.: Lunar reconnaissance orbiter camera (LROC) instrument overview. Space Science Reviews **150**, 81–124 (2010). https://doi.org/10.1007/s11214-010-9634-2 1, 4
30. Silburt, A., Ali-Dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., Tamayo, D., Menou, K.: Lunar crater identification via deep learning. Icarus **317**, 27–38 (2019). https://doi.org/10.1016/j.icarus.2018.06.022 4, 13, 14
31. Smith, D., Zuber, M., Neumann, G., Lemoine, F., Mazarico, E., Torrence, M., McGarry, J., Rowlands, D., Head, J., Duxbury, T., et al.: Initial observations from the lunar orbiter laser altimeter (LOLA). Geophys. Res. Lett. **37**(18) (2010). https://doi.org/10.1029/2010GL043751 4
32. Szwarcman, D., Roy, S., Fraccaro, P., Jakubik, J., et al.: Prithvi-EO-2.0: A versatile multi-temporal foundation model for earth observation applications. arXiv preprint arXiv:2412.02732 (2024) 2, 11, 12
33. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J., Shelhamer, E., Kerner, H., Rolnick, D.: Galileo: Learning global and local features of many remote sensing modalities. In: ICML (2025) 2

34. Wilhelms, D.: The Geologic History of the Moon. USGS Professional Paper 1348, U.S. Government Printing Office (1987) 4
35. Zuber, M., Smith, D., Watkins, M., et al.: Gravity field of the moon from the gravity recovery and interior laboratory (GRAIL) mission. Science **339**(6120), 668–671 (2013). https://doi.org/10.1126/science.1231507 1, 4