MULTIMODAL-MULTIRESOLUTION FOUNDATION MODEL FOR LUNAR REMOTE SENSING

Paolo Fraccaro[†2], Gabby Nyirjesy[†2], Daniela Szwarcman[†2], Himanshu Patil[†1], Vishal Gaur[†1], Rohit Lal[†1], Rachel A. Slank[†3], Geoffrey Dawson[†2], Hiyam Debary[†2], Michael K. Barker[†4,*], Andrew Annex[†5,6], Vishnu Viswanathan[4,7], Zachary Morse[4,8], Ethan I. Schaefer[5], Nikolaos Dionelis[1], Ankur Kumar[1], Campbell D. Watson[2], Manil Maskey[10], Rebekah I. Dawson-Rigas[9], Juan Bernabé-Moreno[2], Rahul Ramachandran[10], and Sujit Roy[1,10,*]

[1]The University of Alabama in Huntsville, Huntsville, AL, USA
[2]IBM Research, Yorktown Heights, NY, USA
[3]University Space Research Association (USRA), Science and Technology Institute, Huntsville, AL, USA
[4]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[5]SETI Institute, Mountain View, CA, USA
[6]NASA Ames Research Center, Moffett Field, CA, USA
[7]University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA
[8]Howard University, Washington, DC, USA
[9]NASA Headquarters, Washington, DC, USA
[10]NASA Marshall Space Flight Center, Huntsville, AL, USA
[*]Corresponding authors: Sujit Roy (sujit.roy@nasa.gov) and Paolo Fraccaro (paolo.fraccaro@ibm.com)
[†]Equal contribution to this work

ABSTRACT

We present a multimodal foundation model for lunar remote sensing, pretrained from scratch on SomBench, a co-registered corpus of nearly two million tile bundles spanning 11 modalities at two spatial scales (i.e., 1 m/pixel and 100 m/pixel). The dataset is geographically partitioned to prevent leakage across train, validation, and test splits. The model adapts the TerraMind masked-token architecture to the lunar setting with two extensions: per-tile acquisition geometry (i.e., illumination angles, solar-frame anchors, tile footprint) is provided as explicit contextual information rather than requiring the model to infer it implicitly from image content; and tiles anchored at meter and hundred-meter scale are trained jointly in a mixed-batch regime so that a single set of weights covers both resolution families. A FlexiViT patch-embedding scheme allows the same pretrained checkpoint to be reused at different working patch sizes without backbone retraining, and modality-wise tokenization lets downstream users drop or add modalities at fine-tuning time. Qualitative generation experiments indicate that the model captures meaningful cross-modal correspondences (i.e., terrain derivatives from elevation, illumination-consistent reflectance from geometry, and coarse spatial pattern across co-registered modality sets). Evaluated across four downstream benchmarks (i.e., crater detection at 100 m/pixel and NAC 1 m/pixel imagery, irregular mare patch segmentation, and polar ice prospectivity regression) using TerraTorch, the pretrained model matches or outperforms ImageNet-pretrained baselines and an architecturally identical random-initialization control. On multimodal ice prospectivity regression at the poles, the pretrained model variants (i.e., full finetuning, Low-Rank Adaptation (LoRA) and frozen) outperform all baselines by margins that exceed run-to-run variability, with the randomly initialized version of our model that beats all but one baselines. This indicates that part of the gain comes from our architecture's native token-level multimodal processing rather than from pretraining alone. Label efficiency is particularly pronounced for crater detection: at 50% training data on crater detection at 100 m/pixel,

the pretrained model already exceeds the strongest ImageNet baseline trained on the full dataset. Among adaptation strategies, LoRA matches or surpasses full fine-tuning on crater detection and irregular mare patches segmentation, using a small fraction of the trainable parameters with lower run-to-run variance. Full fine-tuning remains best on ice-prospectivity regression, with LoRA that however still beats all baselines. We release the pretrained checkpoint, fine-tuning code, and benchmark datasets to support reproducibility and community reuse.

***K*eywords** lunar remote sensing · foundation model · multimodal pretraining · crater detection · ice prospectivity

1 Introduction

The lunar surface is observed through a heterogeneous collection of orbital instruments rather than a single canonical image stream. The Lunar Reconnaissance Orbiter Camera (LROC) provides wide-angle Wide Angle Camera (WAC) and meter-scale Narrow Angle Camera (NAC) imagery [1, 2]; the Lunar Orbiter Laser Altimeter (LOLA) measures topography [3]; Diviner Lunar Radiometer Experiment (Diviner) records reflected and emitted radiation across the thermal spectrum [4]; Mini-RF supplies radar observations [5, 6]; the Kaguya/SELENE Multiband Imager characterizes visible and near-infrared mineralogy [7]; Lunar Prospector Neutron Spectrometer provides hydrogen abundance maps for the lunar poles [8, 9]; and GRAIL (Gravity Recovery and Interior Laboratory) maps variations in the lunar gravity field [10]. These measurements encode complementary aspects of morphology, composition, thermophysics, and subsurface structure, but their native spatial resolutions span approximately four orders of magnitude. Terrain products alone range from near-global 100 m WAC stereo topography [11] and the approximately 60 m SLDEM2015 (SELENE Digital Elevation Model) product [12] to meter-scale NAC stereo elevation models [13, 14]. Consequently, learning a reusable lunar representation is not simply a matter of scaling natural-image pretraining: it requires a model that can align different physical variables, tolerate absent observations, and preserve useful structure while moving between regional and local spatial scales.

Masked autoencoding offers an attractive basis for such a model because it converts abundant unlabeled imagery into a scalable reconstruction objective [15], and MultiMAE extends this principle across modalities [16]. Nevertheless, existing geospatial adaptations [17, 18, 19] transfer imperfectly to the Moon. First, raw-pixel reconstruction targets are noisy, spending model capacity on high-frequency detail that carries little semantic content — a poor trade on low-signal, illumination-dominated lunar imagery. Apparent surface appearance on the Moon is governed more by acquisition and illumination geometry than by intrinsic surface variation, so a model left to disentangle illumination from pixels alone spends capacity recovering a quantity that is already recorded for every tile. Second, because native resolutions span four orders of magnitude, a recipe that trains one resolution family at a time cannot learn how regional context and meter-scale detail constrain one another.

We address these issues with the NASA-IBM Lunar Foundation Model (NASA-IBM LFM), pretrained from scratch on SomBench [20]: nearly two million co-registered tile bundles organized as two data tracks. The WAC-anchored track covers optical, ultraviolet, terrain, slope and aspect; the NAC-anchored track pairs panchromatic meter-scale imagery with co-registered 3 m terrain, slope, and aspect. Both tracks also include contextual information for illumination geometry, thermal, radar, mineralogical, gravity, and various geologic products. Tiles are partitioned by Lunar Transverse Mercator (LTM) zones rather than independently, so geographically overlapping observations cannot straddle the training, validation, and test splits (Section 3). The corpus is deliberately broad enough to support both regional reasoning and meter-scale analysis of features such as small craters and irregular mare patches, whose scientific catalogs illustrate the breadth and difficulty of lunar mapping [21, 22, 23].

Our starting point is TerraMind [24], a best-in-class multimodal recipe developed for Earth observation, which we retrain from scratch rather than fine-tune from its Earth-observation checkpoint. Two lunar-specific extensions follow directly from the second and third problems above. Per-tile acquisition geometry — illumination angles, solar-frame anchors, and tile footprint — is promoted to its own sequence-tokenized encoder input, so the dominant confound on lunar appearance is handed to the model rather than inferred by it. Furthermore, the two anchored tracks are trained jointly at their native resolutions in a single mixed-batch loop, so one set of weights covers both scales instead of one optimization stream per resolution family. Modality-wise tokenization and FlexiViT patch-embedding resizing [25] then make the resulting

checkpoint reusable under input conditions it never saw during pretraining. Section 4 gives the full architecture and pretraining recipe.

Our contributions are:

- **Geometry-conditioned, mixed-resolution pretraining.** We show that supplying acquisition geometry as encoder tokens and training both resolution families in one loop are compatible with a standard multimodal masked-token objective at scale.
- **A flexible checkpoint for downstream task adaptation.** The same weights adapt to patch grids other than the pretraining one, accept modality subsets that differ from the pretraining mixture, and support Low-Rank Adaptation (LoRA) applied to the pretrained encoder.
- **Evidence of importance of lunar pretraining** Across four benchmarks (i.e., crater detection at WAC and NAC scale, irregular mare patch segmentation, and polar ice prospectivity regression, all run through a common fine-tuning framework [26]) the pretrained model matches or exceeds ImageNet-pretrained baselines and an architecture-matched random-init control.

The remainder of the paper is organized as follows. Section 2 compares the model against masked image modeling, available lunar foundation models, and prior machine learning on lunar observations. Section 3 describes the pretraining corpus and the four benchmarks. Section 4 presents the architecture, the pretraining recipe, and the fine-tuning protocol. Section 5 first probes what pretraining has learned through multimodal generation, then reports transfer performance against ImageNet-pretrained baselines and a random-initialization control.

2 Related Work

Masked image modeling and reconstruction targets. Masked autoencoders learn visual representations by reconstructing randomly hidden patches with an asymmetric encoder–decoder architecture [15]. MultiMAE generalizes this objective to multiple input and output domains, sampling a limited visible-token budget across modalities to encourage cross-modal prediction [16].

Discrete-target formulations replace pixels with tokenized codes: BEiT reconstructs discrete visual codes from a pretrained tokenizer [27]; MaskFeat reconstructs handcrafted local descriptors [28]; and Feature-guided MAE augments pixel targets with spectral and spatial descriptors to counteract the pixel-fidelity bias in noisy SAR imagery [29], which is a bias that also afflicts low-signal lunar imagery. TerraMind, inspired by 4M [30], sits in this lineage, using a VQ-VAE tokenizers to produce per-modality discrete codes as the reconstruction target and predicting a categorical distribution over the resulting vocabulary at masked positions [24]. We adopt this tokenization scheme unchanged.

Lunar foundation models. Compared to the rich literature on Earth observation foundation models, prior lunar foundation-model efforts remain limited. LunarFM [31] is trained on curated global WAC-resolution lunar products, emphasizing multimodal representation learning at regional and global scales. A TerraMind-style unified transformer for lunar NAC imagery [32] restricts pretraining to high-resolution generation and shape-from-shading on a handful of Apollo-site NAC scenes, targeting monocular DEM (Digital Elevation Model) prediction rather than a reusable multimodal representation. Our model differs from these because it brings together three aspects: pretraining is globally distributed rather than site-limited, spans both resolution families rather than one, and targets a reusable representation rather than a single reconstruction task or a curated map snapshot.

Lunar observations and machine learning. Impact-crater mapping is a central downstream application. The Robbins catalog contains more than two million lunar craters and provides a global reference for crater location and size [21]. Prior learning-based methods include residual U-Net detection from lunar DEMs [33] and a super-resolution-assisted YOLO architecture aimed at improving small-crater detection under limited resolution and varying illumination [34]. We evaluate crater transfer at both WAC and NAC scales. Irregular mare patches (IMPs) form a distinct fine-boundary segmentation problem. Their complex morphologies and debated origin have motivated an updated catalog and geologic classification [23], making them a useful test of whether meter-scale pretraining captures more than crater morphology.

Terrain reconstruction and geometric consistency. High-resolution lunar terrain has traditionally been obtained from stereo observations or shape-from-shading. Monocular methods combine image shading

with a coarse DEM constraint to produce pixel-scale elevation [35]. Learning-based approaches predict large-area NAC topography constrained by SLDEM2015 [36], combine a CNN estimate with shape-from-shading refinement [37], or use adversarial learning to generate high-resolution DEMs from monocular images and low-resolution terrain [38]. ELunarDTMNet introduces hierarchical multiscale modeling for single-view lunar DTM (Digital Terrain Model) reconstruction and evaluates cross-dataset generalization [39]. Beyond the Moon, DEM super-resolution methods improve local and global terrain features through feature enhancement, residual fusion, attention, and topography-aware transformers [40, 41, 42]. In our model, DTM, slope, and aspect enter the pretraining objective as three separate co-ingested modalities, each with its own tokenizer and its own masked-token reconstruction target, and the shared encoder is left to learn their inter-relationship from the co-registered pretraining data rather than through an explicit differential constraint.

3 Datasets

To train and evaluate NASA-IBM LFM we use the pretraining and benchmarking datasets from SomBench [20], which we built and released as a companion dataset publication. Full details are in the original paper; we summarize the corpus here for the reader's convenience.

3.1 Pretraining dataset

Motivation Lunar remote-sensing archives are large but instrument-siloed: each product ships in its own projection, resolution, and coverage pattern, and most lunar science tasks remain label-poor. We therefore constructed, to our knowledge, the largest co-registered multimodal lunar corpus: approximately two million multimodal tile bundles spanning more than thirty lunar data products at two spatial scales. Its design targets three capabilities that no existing lunar dataset provides jointly. First, pixel-aligned modality bundles make cross-modal structure learnable from spatially corresponding observations and support any-to-any modality completion. Second, two spatial regimes enable multi-resolution representation learning across a $100\times$ scale gap. These pivot around regional geology at Wide Angle Camera (WAC) scale and meter-scale geomorphology at Narrow Angle Camera (NAC) scale, both with geometry acquisition metadata. Third, Geographic partitioning makes downstream evaluation spatially leakage-free.

Data Sources The optical anchors are fully calibrated and map projected Lunar Reconnaissance Orbiter Camera (LROC) experiment data records (EDRs). The low-resolution track is anchored to 54,080 WAC observation pairs, each providing five-band visible reflectance (415–689 nm) at 100 m/px and two-band ultraviolet reflectance (321–360 nm) at 500 m/px. WAC was chosen as the regional anchor because it offers near-global coverage at a resolution that preserves regional geological context, with overlapping observations exposing the same terrain under different illumination. The high-resolution track is anchored to 1,095 NAC frames at $\sim$1 m/px, restricted to the NAC PHO (photometry) sites and the orthorectified frames used to produce the LRO 3 m stereo Digital Terrain Models (DTMs). This ensures that every NAC tile is paired with a co-registered 3 m terrain model rather than 60 m topography alone.

Topography enters at both scales: the LOLA–Kaguya merged SLDEM2015 model (60 m) for the WAC track and NAC-stereo DTMs (3 m) for the NAC one. For each elevation model slope and aspect are also computed at their native resolution, with aspect stored as a sine–cosine pair to avoid the angular wrap-around. This sine–cosine representation is what we will refer to as *aspect* throughout the remainder of the paper. As a photometrically controlled reflectance layer we use the WAC 643 nm normalized-reflectance product (100 m), which is available for every WAC anchor. Beyond these, the corpus attaches $\sim$20 further co-registered data sources spanning across geologic mapping, Diviner thermophysics, LOLA products (roughness, permanently shadowed regions, illumination, 1064 nm normal albedo), Mini-RF radar, Kaguya MI/SP mineralogy and space weathering, GRAIL gravity disturbance, and Lunar Prospector hydrogen abundance (Table 1). Every tile also carries per-product acquisition metadata (solar incidence, emission, phase, and azimuth angles, sub-solar and tile-center coordinates, and ground sampling distance) joined from the LROC EDR geodatabase; these fields are available at inference whenever the image itself is, so they introduce no label leakage.

Dataset Construction Pipeline Tiles are generated by an image-anchored pipeline in which each optical observation defines the spatial extent of its samples (Figure 1): a sliding window over the anchor canvas creates 512×512 px tiles, and all auxiliary products are attached at each tile's exact geographic bounds. Construction operates in equal-area map projections—per-strip Lunar Transverse Mercator (LTM) zones

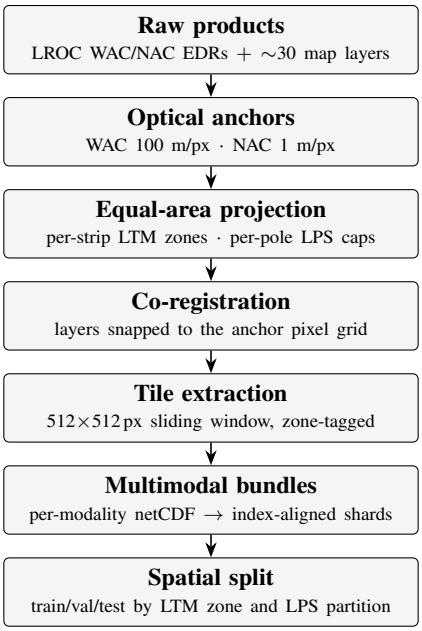


Figure 1: Dataset construction. Each optical observation defines the sample extent; auxiliary products are co-registered onto the anchor grid, extracted as multimodal tile bundles, and partitioned geographically.

and per-pole Lunar Polar Stereographic (LPS) caps—with every source product pre-warped once per zone, so tile extraction never re-projects the anchor imagery itself. Each tile is tagged with its zone, and tiles that straddle zone boundaries are dropped, which later guarantees clean split boundaries (Section 3.1).

Auxiliary data sources are re-gridded onto a per-channel resolution snapped to the anchor grid, so data sources of a bundle share corner pixels despite native resolutions spanning four orders of magnitude. Continuous modalities are bilinearly resampled, categorical products (e.g., the geologic map) use nearest-neighbor interpolation, and multi-band products are stacked along a band dimension. Missing observations are preserved explicitly rather than imputed: data exceeding a missing-pixel threshold are recorded as absent. Bundles are stored as compressed per-modality numpy files and packaged into index-aligned WebDataset shards for high-throughput multimodal loading.

Spatial Splits Random tile-level splits are not suitable for our dataset. The sliding window creates overlapping tiles, and repeated EDR coverage observes the same ground patch in many tile bundles, so a random partition would place near-duplicates on both sides of the evaluation boundary. The corpus is therefore split geographically. Every tile carries its LTM zone or LPS cap; whole zones are assigned to train, validation, or test (≈80/10/10), the polar caps are spatially partitioned with the same guarantee, and tiles that straddle zone boundaries were already dropped at construction time. Zone assignments are shared between the WAC and NAC tracks, so a ground patch admitted to both views is routed to the same split at both scales. Train and validation drive pretraining and checkpoint selection; the test partition is reserved for downstream fine-tuning and evaluation (Section 5).

Statistics Table 1 summarizes the corpus composition. The final corpus contains 963,609 WAC bundles and 1,000,113 NAC bundles, built from 54,080 WAC pairs and 1,095 NAC frames and totalling 43 million modality files (∼53 TB compressed; 51.4 TB WAC, 1.6 TB NAC). All pretraining modalities are present for essentially every tile of their track (null fractions $\leq$0.3%), while coverage of the wider corpus varies by instrument footprint—from near-global layers (geologic map, T_{bol}, gravity) through mid-latitude products (Mini-RF, Kaguya MI) to polar-only layers (PSR, average illumination, ice-stability depth), which exist for ∼6,900 polar tiles per track.

3.2 Benchmarking datasets

Table 2 shows a summary of the SomBench benchmark datasets we used to validate our model. A description for each of the different task is provided below.

Table 1: Corpus composition. Channels marked "multi" are multi-band stacks (e.g., 24 Diviner T_{bol} phases plus the acquisition-matched phase). The **Pretrain** column gives the representation each layer receives during pretraining: *dense* layers are patch-embedded and tokenized as full images, forming the nine image-like modalities; *scalar* layers are averaged over the tile footprint and contribute a single discrete token each, together forming the static-map context modality. With the per-tile optical metadata, this gives the 11 pretraining modalities.

Modality	Key	Native res.	Ch.	Source	Pretrain
WAC family (low-resolution track, 51.2 km tiles)					
Visible reflectance	`vis`	100 m	5	LROC WAC	Dense
Ultraviolet reflectance	`uv`	500 m	2	LROC WAC	Dense
Topography	`dtm`	60 m	1	SLDEM2015	Dense
Slope	`slope`	60 m	1	SLDEM2015	Dense
Aspect	`aspect`	60 m	2	SLDEM2015	Dense
NAC family (high-resolution track, 512 m tiles)					
Panchromatic imagery	`nac`	1 m	1	LROC NAC	Dense
Topography	`dtm_3m`	3 m	1	NAC-stereo DTM	Dense
Slope	`slope_3m`	3 m	1	NAC-stereo DTM	Dense
Aspect	`aspect_3m`	3 m	2	NAC-stereo DTM	Dense
Additional corpus layers (attached to both tracks as a tile-level average, where available)					
Norm. 643 nm reflectance	`nr643`	100 m	1	LROC WAC	Scalar
Regolith temp. anomaly	`TREG`	240 m	1	Diviner	Scalar
Bolometric temp. (phases)	`TBOL`	15 km	multi	Diviner	Scalar
Rock abundance	`ROCK_ABUND.`	240 m	1	Diviner	Scalar
H-parameter	`HPAR`	240 m	1	Diviner	Scalar
Ice-stability depth	`DICE`	240 m	1	Diviner	Scalar
Roughness	`ROUGHNESS`	1 km	1	LOLA	Scalar
PSR / illumination	`PSR`, `AVG_ILLUM`	20–120 m	1	LOLA	Scalar
1064 nm albedo	`ALBEDO`	1 km	1	LOLA	Scalar
Radar CPR / reflectivity	`MINIRF_*`	90 m	1	Mini-RF	Scalar
MI mineralogy / SW-Fe	`MI_*`, `SW_FE`	60 m–1 km	multi	Kaguya	Scalar
Polar mineralogy	`SP_MINER.`	1 km	multi	Kaguya SP	Scalar
TiO_2 / morph mosaic / refl.	`TIO2`, `WAC_*`	100–500 m	1	LROC WAC	Scalar
Gravity disturbance	`GRAVITY`	20 km	1	GRAIL	Scalar
Hydrogen abundance	`HYDROGEN`	15 km	1	Lunar Prospector	Scalar

Table 2: Overview of SomBench benchmark datasets used for the validation of our model.

Name	Science Theme	Image Size (px)	Resolution (m/px)	Target	Train	Val	Test
Object Detection							
Robbins craters	Impact processes	512×512	100	1 class	800	100	100
NAC hand labeled	Impact processes	256×256	1	1 class	645	59	62
Segmentation							
Irregular Mare Patches	Volcanic History	256×256	1	1 class	100	20	10
Regression							
Lunar ice prospectivity	Polar Volatiles	256×256	240	0–1	108	23	25

Impact Processes

Impact craters are the most ubiquitous lunar landform and span orders of magnitude in scale across all terrain types, making them a natural ML benchmark for detection, segmentation, and size–frequency analysis. Crater catalogs also underpin downstream applications from relative-age dating to geologic mapping and landing-site hazard assessment.

Robbins Crater Catalog The Robbins catalog [21] is a manually annotated global database of over 2 million lunar impact craters, intended as a near-complete census of diameters $\geq$1–2,km, digitized

from LROC WAC global mosaics with topographic cross-checks under challenging illumination. The rim polygons are converted to per-tile bounding-box annotations. This is used to create a context-scale object-detection benchmark of 1,000 WAC visible tiles drawn from the test split of the pretraining dataset, filtered to solar-incidence angles between 60° and 80° for favorable illumination.

NAC Hand Labeled datasets A meter-scale crater benchmark is built from LROC NAC imagery (1–5 m/pixel) spanning six NAC PHO sites (Highlands, King Ejecta, Apollo 15 S-IVB, Apollo 17, Reiner Gamma, and March 17 Impact Crater) selected under relief-enhancing illumination conditions (incidence angles of ~50–80°). Using co-registered 3 m/pixel DTMs alongside the NAC images, craters are manually labeled within 4–6 fixed 1024×1024 px study areas per site and digitized as circles using OpenCraterTool [43]. The annotations are then mapped to 256×256 NAC patches and packaged in COCO format, yielding 766 patches with 97,104 crater annotations. Dataset splits are enforced at both the site and study-area levels to prevent data leakage. Sites annotated at 5 m/pixel are visibly blurrier than those acquired at higher spatial resolution.

Volcanic History Irregular mare patches (IMPs) complement impact-feature benchmarks for assessing volcanic history. They are rare, morphologically diverse landforms that are typically distinguished by subtle boundary expressions rather than strong visual signatures. The benchmark was constructed starting from 2,623 polygon annotations [23] paired with corresponding NAC images. In particular, a subset of annotations was manually adjusted to correct misalignment caused by NAC pointing uncertainty (typically on the order of tens of meters in uncontrolled frames), and approximately 100 polygons exceeding the 256 m patch size were excluded. These limitations highlight the potential value of future benchmark datasets derived from controlled NAC products, such as NAC PHO.

Polar Volatiles To support polar-volatiles benchmarking, we test NASA-IBM LFM at incorporating the polar map stack of [44], packaged as eight static layers within 10° latitude of each pole at 240 m/pixel. The layers combine thermophysical constraints (ice stability depth, maximum surface temperature), illumination state (permanently shadowed regions (PSRs), computed via horizon/visibility modeling), terrain controls (slope, aspect, curvature), and two spatial-context features capturing cold-trap "halo" effects (distance to PSRs, PSR density). The regression target is the final continuous ice-prospectivity map that estimates where water ice is most likely within the upper ∼1 m of regolith [44]. This was originally calculated as knowledge-driven fuzzy overlay of the eight inputs [44]. Topographic derivatives are computed from the DTM and thermal layers from a numerical thermal model calibrated to Diviner observations. This compact, physically motivated stack enables standardized evaluation of polar terrain classification and ice-prospectivity prediction under extreme illumination regimes, with consistent spatial extent, resolution, and projection conventions.

4 Method

NASA-IBM LFM's architecture is shown in Figure 2. It follows the TerraMind [24] design, which formulates multimodal learning as a token-prediction problem. Heterogeneous modalities are mapped into a shared embedding space and processed by an encoder-decoder Transformer. During pretraining, the model learns correlations across and within modalities by predicting sampled target tokens from sampled multimodal inputs.

4.1 Modalities and representations

As described in Section 3, we pretrained NASA-IBM LFM on two families of multimodal samples centered on LROC NAC and WAC optical observations, respectively. Each training sample belongs to either a NAC-centered or a WAC-centered modality set and does not combine observations from the two. Within each modality set, corresponding regions from the available dense modalities describe *the same geographic region*.

The model processes dense image-like modalities and sequence-like contextual modalities. The dense modalities comprise the optical observation (WAC or NAC), DTM-derived topographic modalities (elevation, slope, and aspect), and UV (the latter being available only in the WAC-centered set). The sequence-like modalities comprise per-tile optical metadata and contextual variables derived from static maps. Table 3 details the fields considered and the type of information they provide. The three input processing branches shown in Figure 2 (center-left) are:

- *Pixel-level image input*: Each dense modality is divided into 16×16-pixel patches. Learnable linear projection layers map the pixel values of each patch to the shared encoder embedding dimension, providing the model with fine-grained representations of the dense modalities.
- *Token-level image input*: The dense modalities are also processed by modality-specific VQ-VAE tokenizers trained on the pretraining corpus. Each tokenizer encodes a 16×16-pixel region into a latent representation and quantizes it into a discrete code using Finite Scalar Quantization (FSQ) [45]. The resulting codes are mapped to the shared encoder embedding dimension through modality-specific learnable embedding layers.
- *Sequence-level input*: The eight optical metadata variables and the 28 static-map-derived context variables are encoded as sequences of discrete tokens. Optical metadata variables are provided as per-tile scalar values, while static-map-derived variables are averaged over the tile footprint to obtain one scalar per field. Each value is assigned to a field-specific interval and represented by a string that identifies both the field and the interval. For example, using longitude bins of width 0.25°, a center longitude of 87.49° is assigned to the interval [87.25°,87.50°) and represented as `C_LON=87.25->87.50`. A text tokenizer then converts the resulting string for each variable into a token ID. Similar to the image case, modality-specific learnable embedding layers map the token sequences to the shared encoder embedding dimension.

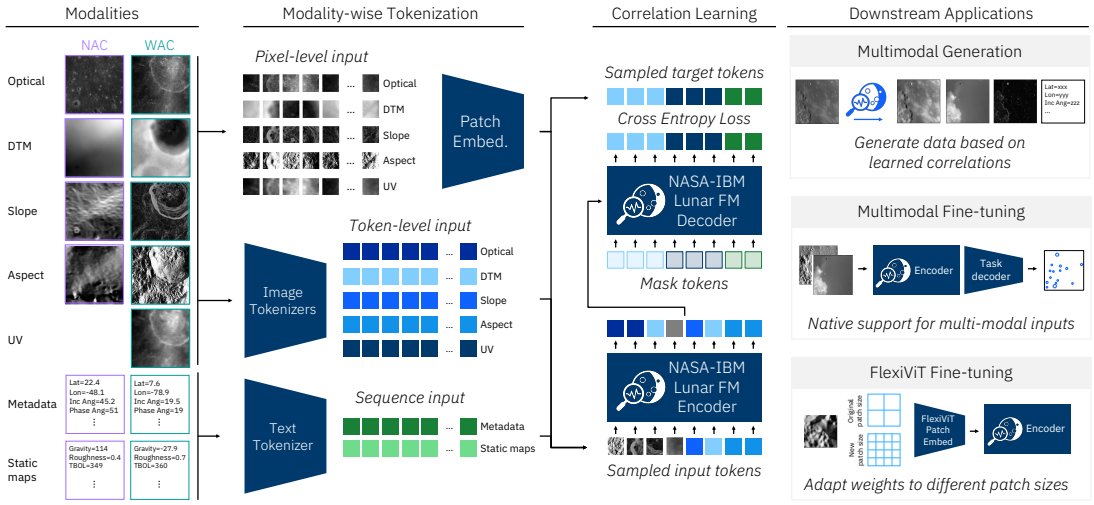


Figure 2: Model overview. *Left:* input modalities include low- and high-resolution images (LROC WAC and NAC), DTM-derived topographic variables (DTM, slope, aspect), per-tile optical metadata, and static context maps. *Center-left:* modality-wise tokenization in parallel branches. Image modalities are processed by patch-embedding layers (*Pixel-level input*) and VQ-VAE tokenizers (*Token-level input*); metadata and static maps are encoded with a text tokenizer (*Sequence input*). *Center-right:* correlation learning follow the TerraMind training procedure. A subset of the modality tokens is sampled as encoder inputs, while a separate subset is sampled as prediction targets. The encoder processes the input tokens, and the decoder predicts the target tokens under a cross-entropy objective. *Right:* The pretrained model can be used for multimodal generation and multimodal fine-tuning with native support for multiple input modalities. The added FlexiViT support enables fine-tuning at patch sizes different from those used in pretraining.

4.2 Multimodal pretraining framework

Pretraining follows the TerraMind masked-token objective [24] (Figure 2, center-right). The encoder can jointly process tokens and patches from the different input modalities. A learnable modality embedding is added to each token or patch, together with a sine-cosine positional embedding: 2D for dense modalities and 1D for sequence modalities. To predict the selected target tokens, the decoder receives the encoder output together with mask tokens for dense modalities or left-shifted tokens for sequence modalities.

At each training step, we draw per-modality token-budget fractions for the input and target sets from separate Dirichlet distributions over the modalities.

Table 3: Context variables derived from static maps and optical metadata. For each tile, the corresponding map is averaged over the tile footprint and quantized into a single token.

Category	Variables	Description
Optical metadata	EM_ANG INC_ANG PHASE_ANG SS_GROUND_AZIMUTH SS_LAT SS_LON C_LON C_LAT	- Emission, incidence, and phase angles describing viewing geometry - sub-solar azimuth, latitude, and longitude anchoring the solar reference frame - tile center coordinates
Photometric properties	WAC_NORM_REF_[321,360,415,566,604,643,689] ALBEDO	Photometrically normalized WAC reflectance products and broadband albedo.
Illumination and thermophysical	AVG_ILLUM TBOL_closest TBOL_POLES_closest HPAR TREG DICE	Quantities describing illumination conditions and thermal environment.
Composition and maturity proxies	SP_MINERALOGY_feo SP_MINERALOGY_high_calcium_pyroxene SP_MINERALOGY_low_calcium_pyroxene SP_MINERALOGY_nanophase_iron SP_MINERALOGY_olivine SP_MINERALOGY_omat SP_MINERALOGY_plagioclase SW_FE_mpfe SW_FE_npfe SW_FE_smfe TIO2	Composition and maturity proxies from Kaguya-derived spectral mineralogy, space-weathering iron partitioning, and titanium.
Volatile-related	HYDROGEN	Hydrogen abundance proxy.
Physical and geophysical	GRAVITY ROCK_ABUNDANCE ROUGHNESS	Physical surface characteristics and geophysical products.

$$f(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{m=1}^{M} p_m^{\alpha_m - 1}, \qquad p_m \geq 0, \qquad \sum_{m=1}^{M} p_m = 1, \tag{1}$$

where M is the number of eligible modalities in the sample, $B(\boldsymbol{\alpha})$ is the multivariate beta function, $\alpha = (\alpha_1, \ldots, \alpha_M)$ is the concentration vector, and π_m determine the fractions of the budgets assigned to modality m. Eligibility depends on the sample: NAC- and WAC-centered samples contain different modality sets, and only modalities present in that sample can be selected as corresponding inputs or targets.

These fractions are converted into per-modality token counts that can round to zero (absence of modality) if too small. With the per-modality budgets defined, we then sample the tokens within modalities: for dense modalities, patches or tokens are sampled uniformly, and for sequence modalities we use span-masking. This sampling mechanism of allocating input and target tokens exposes the model to a broad range of within- and cross-modal prediction tasks.

Although the inputs can include tokens and pixel-level patches, the prediction targets are always discrete tokens. Predicting the discrete tokens in a finite vocabulary can be seen as a classification problem over that vocabulary. The decoder outputs a categorical distribution at each target position, and the loss is the standard cross-entropy against the ground-truth codes.

4.3 Model design considerations

Acquisition geometry as explicit context. Standard vision transformers encode spatial arrangement of image patches through positional embeddings but do not necessarily represent acquisition conditions explicitly. This is particularly relevant for lunar optical data, where image characteristics vary with illumination and viewing geometry. Incidence, emission, and phase angles characterize the photometric geometry, while latitude and local solar geometry provide context for illumination, shadowing, and surface temperature. Since this information is available for each tile, we provide it to NASA-IBM LFM as inputs. Explicit conditioning allows the model to account for known acquisition and environmental context when learning representations of the lunar surface.

Mixed low- and high-resolution inputs. Each tile is anchored to one of two native optical grids (LROC NAC at $\sim$1 m per pixel or LROC WAC at $\sim$100 m per pixel) with every non-optical modality (i.e. DTM and derived modalities) resampled to its tile's anchor grid. Rather than forcing the corpus onto a single common grid, we let the data-parallel setup mix the two: each worker independently samples either a NAC- or a WAC-anchored batch at each step, so a typical global step carries gradients from both resolution regimes and combines them through the standard all-reduce. Per-modality generated token losses are averaged into one scalar per step that drives a single backward pass, with no separate optimization stream per anchor type. This joint training strategy is intended to expose the model parameters to both resolution regimes while retaining their complementary spatial characteristics.

FlexiViT patch-size adaptation. Downstream lunar tasks vary widely in the spatial scale of the phenomena of interest (kilometer-scale craters, meter-scale hazards, sub-tile polar terrain classes). To let a single pretrained checkpoint serve tasks at different working scales, the model adopts the FlexiViT [25] patch-embedding scheme: the patch-embedding weights learned at pretraining time can be resized on the fly to a different patch grid at fine-tuning time, without re-training the backbone.

4.4 Pretraining Setup

The pretraining procedure comprises three stages: (1) training the modality-specific tokenizers, (2) using the trained tokenizers to convert the pretraining data into discrete tokens, and (3) training NASA-IBM LFM on the resulting tokenized data. Tokenizing the data in advance enables efficient multimodal pretraining by avoiding repeated tokenizer inference. From stage (2) onward, we use 256$\times$256 images (i.e., the model input size) obtained by cropping the corners of the original 512$\times$512 SomBench samples. Corner cropping ensures alignment between the pre-computed tokenized data and the raw modalities, which would not be guaranteed under random cropping without re-running tokenization for each sample at each pretraining step. In the following paragraphs, we describe the hyperparameter configurations used for tokenizer training and NASA-IBM LFM pretraining.

Tokenizers We train nine modality-specific tokenizers considering the same encoder-decoder architecture: a ViT-B encoder with a patch size of 16$\times$16 and a U-Net-patched decoder. Latent representations are quantized using FSQ [45] with a single codebook and per-dimension levels of (8, 8, 8, 6, 5), yielding a five-dimensional quantized latent representation. The generative decoder is trained using a denoising diffusion probabilistic model (DDPM) with 1,000 diffusion timesteps and a linear beta schedule.

Optimization is performed using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 0.05, together with a cosine learning-rate schedule that decays to zero and fp16 mixed precision. The training objective combines pixel-wise mean squared error with a perceptual loss computed from intermediate activations of an ImageNet pretrained ViT-Tiny model. The number of GPUs, training epochs, warmup duration, and learning rate vary by modality and is detailed in the Appendix.

Training configuration. The encoder is ViT-B with 768 hidden dimensions, 12 encoder layers, and 12 attention heads; the decoder is a Transformer with 12 layers and shared width. Training is carried out on 16 H100 GPUs for 150k steps at a peak learning rate of 1e-4 with a cosine schedule and 15k warm-up steps. Global batch size is 1,536 tiles. Mixed-precision training uses bf16, with the final pretraining run that took 1.1k GPU-hours. The pretraining configuration is released as a yaml file alongside the model weights, and the pretraining data is exactly as released in SomBench [20].

4.5 Finetuning Setup

All finetuning experiments are implemented using TerraTorch [26], an open-source PyTorch-based framework for geospatial foundation-model adaptation. TerraTorch provides a unified interface for assembling encoder–decoder pipelines from interchangeable backbones and task-specific decoder heads (UNet for dense prediction, Faster/Mask R-CNN for detection and instance segmentation, and a ViTDet-style regression head for dense regression), configured via a single YAML file without custom training code. On top of TerraTorch we build a custom module that packages our benchmark datasets and backbone wrappers, adds support for registering new input modalities into the pretrained backbone at fine-tuning time, and implements FlexiViT patch-embedding interpolation (so the pretrained checkpoint can be re-used at working patch scales other than that used in the pretraining). Within each downstream task the data loaders, splits, augmentations, loss, and evaluation metric are held fixed across all backbones so that only the encoder weights and initialization differ. Each configuration is repeated for five random seeds (see per task details below), and we report mean and standard deviation across them. Full hyperparameter configurations are released alongside the finetuning code on GitHub.

We compare our model against a set of widely-used baselines spanning the two dominant architectural families (convolutional and transformer based) for geospatial fine-tuning. For the convolutional architectures we include ResNet-50 [46], whose skip-connection design remains a well-understood supervised baseline, and the modernized ConvNeXt-Base [47] and its ConvNeXt-V2 successor [48], which incorporate Vision-Transformer design principles while retaining spatial locality. For transformer architectures we include ViT-Base pretrained under the MAE self-supervised objective [49, 15], together with the hybrid local-global attention variants Swin V2-Base [50] and DaViT-Base [51]. All backbones are initialized from public ImageNet-1k or ImageNet-22k pretrained weights loaded through standard registries. To isolate the contribution of pretraining from that of architecture, we additionally run a from-scratch ResNet-50 control and a randomly initialized copy of our model. For IMP segmentation, which uses a segmentation specific decoder, we also include two task specific architectures assembled through the standard `segmentation_models_pytorch` factory: DeepLabV3+ with a ResNet-50 backbone [52] and SegFormer with a MiT-B2 encoder [53].

The optimizer recipe is deliberately not shared across backbone families. Our pretrained model uses AdamW with layer-wise learning-rate decay (LLRD) and a lower encoder learning rate than head learning rate. This aims at avoiding that the representations learned during pretraining are pushed away in the early epochs of fine-tuning. All ImageNet-pretrained baselines and the from-scratch controls (including the random-init copy of our model) use AdamW with a flat learning rate and weight-decay values commonly used in the natural-image finetuning literature for their respective architectures. For impact processes and volcanic history tasks we additionally evaluate a parameter-efficient LoRA variant of our model, applying rank-16 adapters (α =32) to the attention and MLP linear layers of the pretrained encoder while keeping the remaining encoder weights frozen; the decoder head is trained normally. The task-specific values used per benchmark are given in the paragraphs below.

Impact Processes finetuning configuration. All backbones are fine-tuned with a Faster R-CNN head. For our ViT-based encoder we assemble a Multilayer Simple Feature Pyramid neck that reshapes encoder tokens from four selected layers (indices 2, 5, 8, 11 of 12, zero-indexed) into a spatial grid and interpolates them into a four-level FPN before the RCNN heads. Optimization uses AdamW with a cosine schedule after 500 warm-up steps and a floor of $\eta_{\min}$ = 1e-6. Our pretrained backbone uses LLRD with backbone and head learning rates both set to 5e-5, layer decay 0.65, encoder weight decay 0.05–0.1 and head weight decay 1e-3, together with stochastic depth (drop_path = 0.1), following the recipe standard for MAE-pretrained ViTs. The random-init copy of our model uses a flat learning rate (1e-4 on both backbone and head, layer decay 1.0, no stochastic depth), since there are no pretrained features to preserve. All ImageNet baselines use the same flat 1e-4 learning rate on both backbone and head. Anchor sizes are set per-dataset to match crater size distributions (NAC: sizes ranging from 4 to 96 px across four pyramid levels, aspect ratios [0.7, 1.0, 1.4] reflecting near-circular geometry). RPN and ROI sampling budgets are increased (RPN 512 samples/image, ROI 1024 samples/image, 600 detections/image) to accommodate the density of craters per tile. Training data is augmented with D4 group augmentation (rotations and flips) applied jointly to images and bounding boxes; images are 256×256 (NAC, 1 m/px) or 512×512 (WAC, 100 m/px). For WAC we subsample to 50% of the dataset to assess label efficiency. All models are trained with batch size 4 for up to 100 epochs with early stopping on validation mAP (patience 15), and best checkpoints are selected by validation mAP.

Volcanic History finetuning configuration. The IMP segmentation benchmark comprises 100 / 20 / 10 image–mask pairs (train / validation / test) of 256×256 single-channel NAC tiles, and is framed as a binary semantic-segmentation task (background vs. Irregular Mare Patch). All backbones use a UNet decoder trained with Dice loss and an auxiliary segmentation head (weight 1.0), with the exception of DeepLabV3+/ResNet-50 and SegFormer/MiT-B2, which use their native segmentation heads and a single unified learning rate. Inputs are standardized with per-dataset statistics computed over the training split after masking no-data pixels, and augmented with D4 group transforms and random brightness/contrast jitter. Every configuration is trained with batch size 32 for a fixed budget of 600 epochs using AdamW with a cosine schedule and 500 warm-up steps; our pretrained backbone uses LLRD 0.85 with backbone learning rate 1e-4 and head learning rate 3e-4, while the ImageNet baselines and from-scratch controls use a flat learning rate of 1e-4. Experiments are repeated for five random seeds.

Polar Volatiles finetuning configuration. All ice-prospectivity models share a ViTDet-style Simple Feature Pyramid (SFP) → torchvision FPN → sum-fuse decoder → 4-stage GroupNorm+GELU deep head → 1×1 regression conv. For our ViT-B backbone (patch size 8, 32×32 token grid, 12 blocks) we tap four encoder blocks (2, 5, 8, 11) and feed each into a distinct SFP scale head (upsample 4× / 2× / identity / MaxPool, yielding strides 4/8/16/32). Each modality has its own pretrained `PatchedInputAdapter` and tokens are concatenated along the sequence axis, preserving per-modality pretraining. The ImageNet baselines, which cannot do late fusion natively, share the same FPN, decoder and head but use early channel-concatenation of the nine input bands at the stem. This is done given their lack of ability to process multimodal data natively. The FPN unifies all pyramid levels to 256 channels; the decoder bilinearly upsamples every level to the finest scale, sums them, and further upsamples to the full 256×256 input resolution before a 4-conv GroupNorm head. Optimization uses AdamW ($\beta = (0.9, 0.98)$, $\varepsilon =$ 1e-6) with LLRD 0.85 on our pretrained backbone and a flat learning rate on every baseline. Batch size is 4, warm-up over 130 steps (≈10 epochs) is followed by cosine decay to 1e-6; weight decay is 0.05 and gradient norm is clipped at 0.5. The training loss is MSE. Inputs are per-channel z-score standardized and augmented with the D4 group. Every model trains at both backbone and head learning rate 3e-4 for 100 epochs.

For this task, we run two experiments: a full eight-modality comparison of our model (pretrained and randomly initialized) against the ImageNet baselines, repeated across five random seeds; and an "add-one" modality-count ablation for our model and ConvNeXt-Base that starts from the two pretrained modalities {aspect, slope} and progressively adds one modality at a time in domain-scientist priority order (DICE → TMAX → LPSR → LPSR_DEN → LPSR_DIS → CUR), yielding configurations $m = 2 \dots 8$. This is repeated across three random seeds. Every ablation uses the same architecture, learning-rate schedule and training length as its baseline row, keeping the comparison controlled to the axis being varied.

5 Results

5.1 Multimodal correlations exploration

Before turning to downstream fine-tuning, we ask a simpler question of the pretrained model: has it learned the cross-modal correspondences that its objective is designed to capture? A masked-token generative model can be probed for this qualitatively by asking it to generate a subset of modalities conditioned on the remaining ones. We stress upfront that the goal of the model is not to serve as a lunar generative product, and that generated fields should not be interpreted as scientific-grade outputs; the examples in this section are qualitative sanity checks on the pretraining objective, not calibrated predictions. A systematic evaluation of generation quality is out of scope for this paper; the downstream evaluations in Section 5.2 test the representation on tasks that have well-defined ground truth and standardized metrics. Here we report several examples of chained generation. In this setting, when multiple outputs are required, each output modality is generated individually and sequentially, with the output of each step used as input to the next. This approach promotes greater coherence across modalities, although errors introduced in earlier steps may propagate to subsequent generations. Additional examples are provided in the Appendix.

DTM to slope and aspect. The simplest correlation to check is between elevation and its analytical derivatives. Figure 3 and Figure 4 show the model generating WAC-scale and NAC-scale slope and aspect maps conditioned on the corresponding DTM. Slope and aspect are near-deterministic functions of elevation, so we expect the model to have learned this mapping. Qualitatively, this seems the case: generated fields track the target closely at both scales, and the derivative structure (crater rims, ridge orientations, shadowed slopes) is preserved.

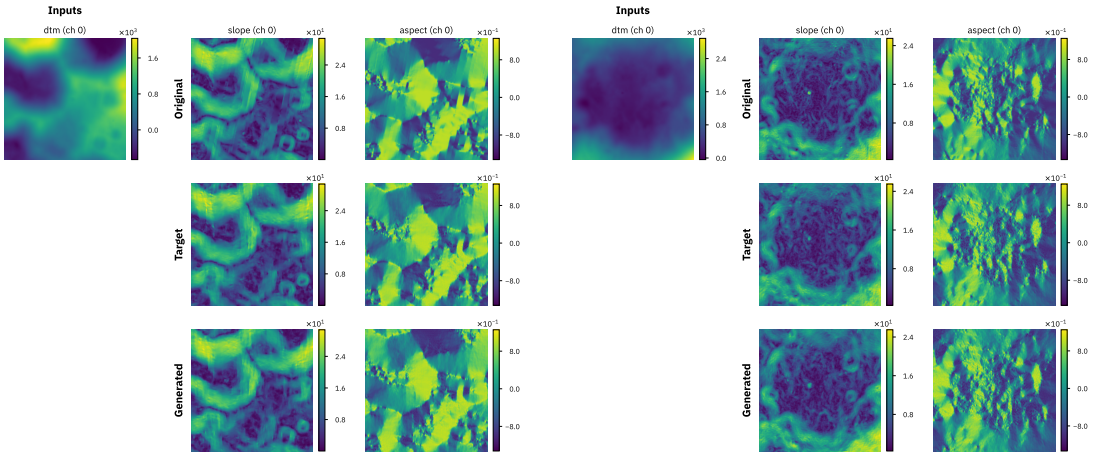


Figure 3: Example of generation from DTM to slope and aspect for the WAC family.

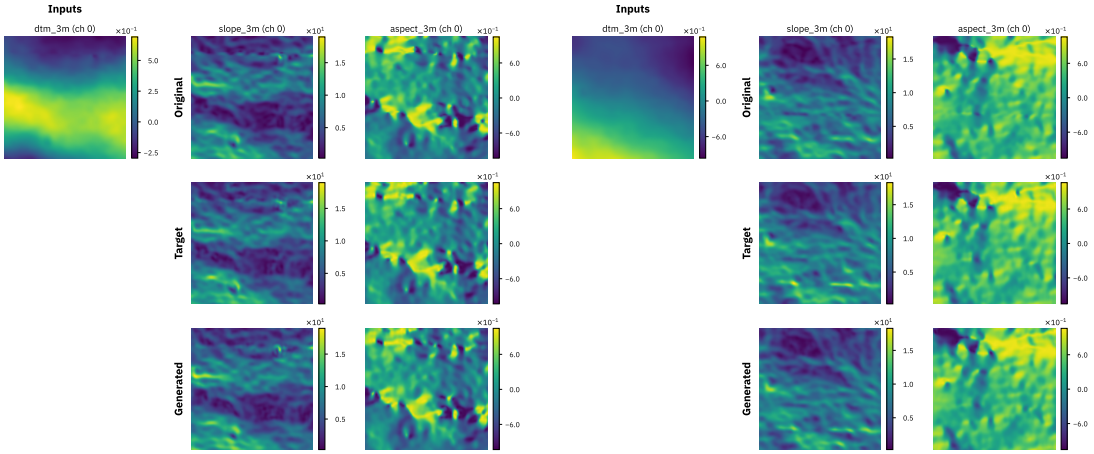


Figure 4: Example of generation from DTM to slope and aspect for the NAC family.

DTM and geometry to visible reflectance. A stronger test is whether the model can render the same terrain under different acquisition geometries. Figure 5 shows two WAC visible-reflectance generations at the same geographic location but with different metadata — one at moderately high incidence, one at a much more grazing incidence. In the first case, the generation is visually close to the observed image, with the difference-histogram concentrating near zero. In the second case, the generated field does not reproduce the target pixel-for-pixel. Nevertheless, the illumination pattern the model produces is qualitatively consistent with the metadata it was given: shadowed regions align with the sun direction implied by the sub-solar coordinates, and the illumination gradient follows the specified geometry. This is the behavior we hoped the sequence-tokenized metadata would elicit, with the conditioning on geometry rather than treating pixel content as the sole signal. However, we caution that this is a single-example qualitative demonstration and not a systematic evaluation of geometric fidelity.

Chained vis-to-all generation. Figure 6 and Figure 7 show the harder regime of chained generation across all modalities from a single visible input. In the first example, spatial patterns are recovered across DTM, slope, aspect, and UV, and several static-map scalars fall close to their target values. In the second example, we illustrate a limitation we often observe: the DTM *pattern* is recovered (crater outlines and ridges appear in the right places) but the absolute elevation offset is not, so the generated topography has the correct shape at a shifted value. Similar behavior appears in the static-map bank. Some fields (e.g. gravity, roughness, TREG) are close to the target range while others drift more. We observe similar patterns for the metadata output, where absolute latitude and longitude coordinates are frequently off by tens of degrees. These failures are exactly the ones we would expect from a model that has learned local

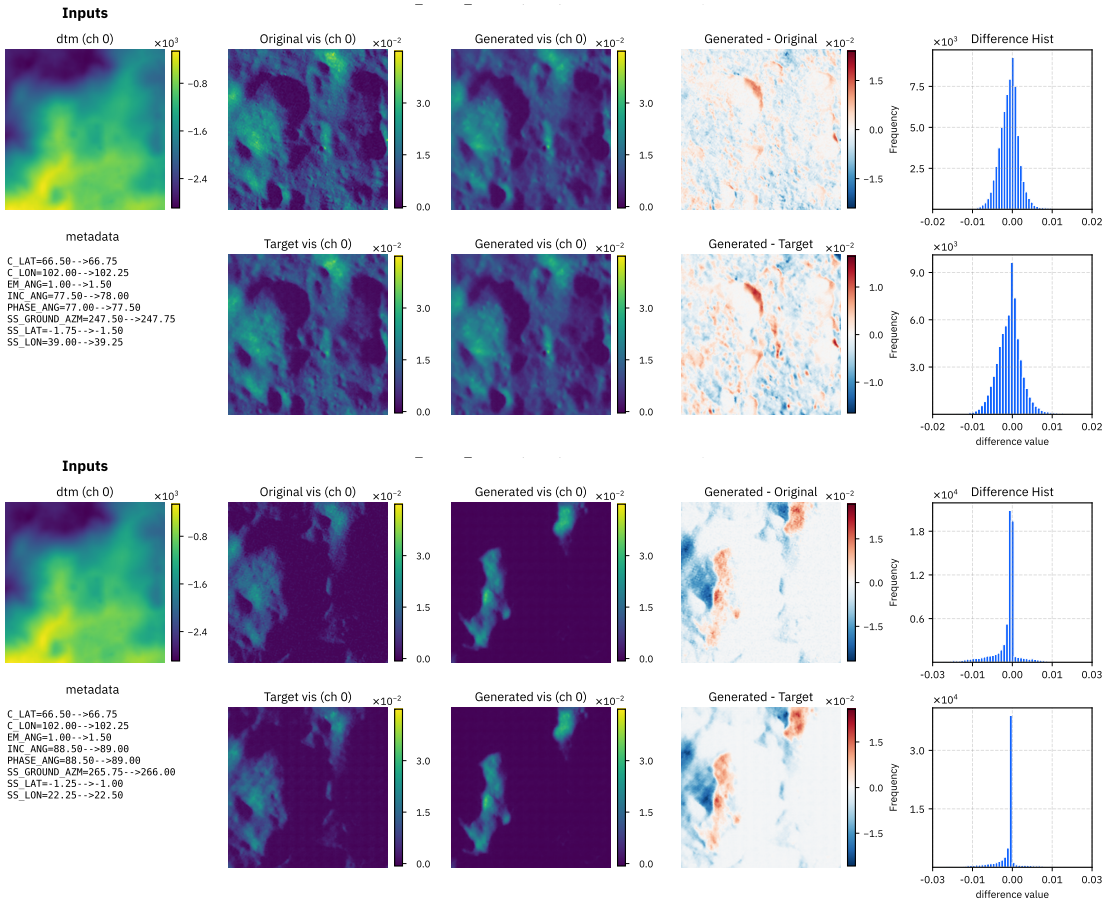


Figure 5: Example of generation from DTM and metadata to WAC for the same location.

structure from pixel-level correlations but has no direct route to a global geodetic reference frame, and they underline that the model is a representation for downstream tasks rather than a substitute for physical instruments or geodetic solutions.

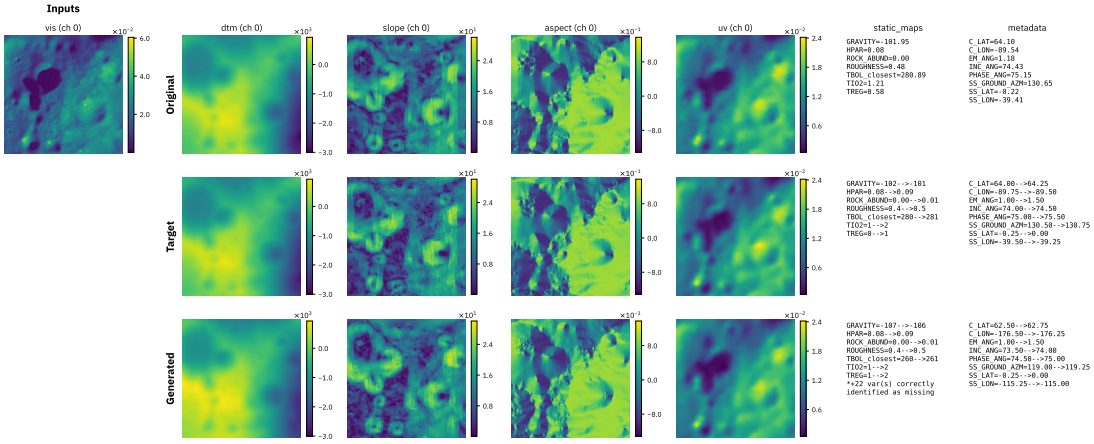


Figure 6: First example of generation from WAC to all other modalities.

Taken together, these qualitative examples indicate that pretraining has induced the cross-modal correspondences the objective was designed to capture (i.e., terrain derivatives from elevation, illumination-consistent

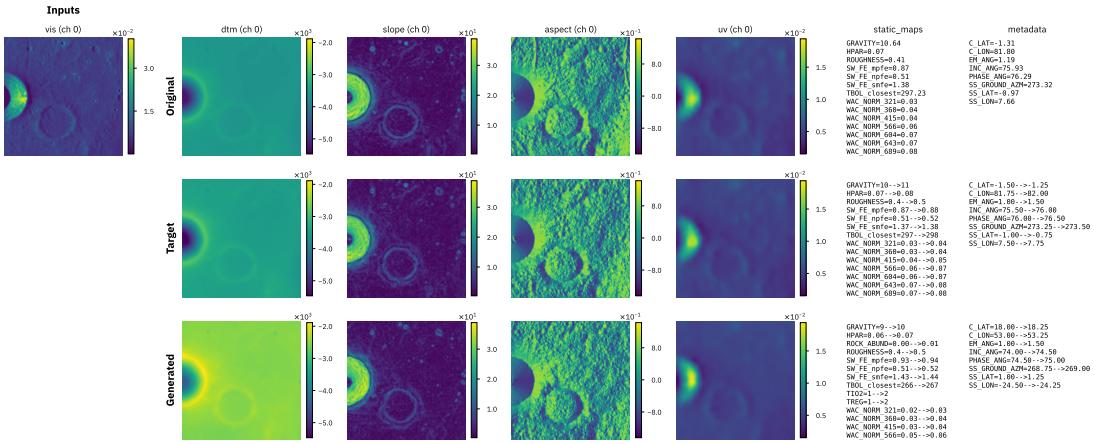


Figure 7: Second example of generation from WAC to all other modalities.

reflectance from geometry, and coarse spatial pattern across a wide set of co-registered layers), while also making the model's limits legible. The quantitative case for the pretraining recipe is made in Section 5.2, on downstream tasks with objective evaluation metrics.

5.2 Finetuning

Impact Processes

Robbins Crater Catalog Table 4 reports bounding-box mAP on the WAC context-scale crater detection benchmark at two training-data fractions. NASA-IBM LFM (pretrained) leads in every metric column at both fractions: at 50% training data, full fine-tuning achieves mAP 0.2541±0.0018 and LoRA achieves mAP 0.2539±0.0014, both ahead of the strongest ImageNet baseline (SwinV2-B at mAP 0.2313±0.0027) and of the randomly initialized NASA-IBM LFM (mAP 0.2197± 0.0027). At 100% training data the LoRA variant tops every column (mAP 0.2581±0.0017, AP@50 0.6183±0.0034, AP@75 0.2263±0.0027), while full fine-tuning follows closely at mAP 0.2537±0.0034. Frozen-backbone fine-tuning underperforms the baselines. The pretrained NASA-IBM LFM variants trained on 50% of the data already match or exceed SwinV2-B trained on the full set, suggesting labeling efficiency improvement from lunar-domain pretraining. The ranking is stable across data fractions, and the performance gap is substantially larger than the run-to-run variation. This indicates a meaningful advantage of NASA-IBM LFM on this benchmark.

NAC Hand Labeled dataset Table 5 reports mAP on the meter-scale NAC crater detection benchmark, which generally shows low scores across the board. NASA-IBM LFM finetuned with LoRA achieves 0.1543±0.0098 mAP, immediately after the best performing model (SwinV2-B) at 0.1552±0.0086 mAP. DaViT-B performs best for AP@50 and SwinV2-B shows top performance for AP@75, with NASA-IBM LFM following closely. The random-init LFM (mAP 0.1274±0.0151) is one of the weakest results. The frozen variant (mAP 0.1295±0.0038) performs similarly to random init on this task, indicating that some encoder adaptation is needed at meter scale for this task. Variability across the different seeds is wide on this task, with performance models at the top that are within each other's standard deviation. This suggests comparable performance among those, rather than a clear winner.

Figure 8 presents examples of predictions for a sample for the Robbins crater catalog (top) and NAC Hand Labeled dataset (bottom) for NASA-IBM LFM and SwinV2-B. The predictions for the two models look similar. However the images show some differences across the two datasets. The Robbins catalog presents less dense and more clearly defined labels compared to the NAC Hand Labeled dataset. Furthermore, part of the NAC benchmark was also annotated at 5 m/pixels, and those sites — including the one shown in the bottom row — are visibly blurrier. These two differences might explain the lower scores for all models in the NAC Hand Labeled dataset.

Volcanic history

Table 4: Crater detection on LRO WAC (context-scale) as a function of training-data fraction. Test-set bounding-box mean average precision. Values are mean ± standard deviation over five random seeds. **Bold** indicates the best value per column; underline the second best; differences smaller than the reported seed spread should not be read as a ranking.

Model	mAP	AP@50	AP@75
50% training data			
ResNet-50 (ImageNet)	0.1993 ± 0.0007	0.4960 ± 0.0036	0.1572 ± 0.0038
ResNet-50 (random)	0.1873 ± 0.0011	0.4663 ± 0.0029	0.1472 ± 0.0041
ViT-B MAE (ImageNet)	0.2167 ± 0.0022	0.5656 ± 0.0092	0.1670 ± 0.0042
ConvNeXt-B (IN22k)	0.2266 ± 0.0030	0.5684 ± 0.0030	0.1840 ± 0.0040
ConvNeXtV2-B (IN22k)	0.2303 ± 0.0034	0.5845 ± 0.0036	0.1829 ± 0.0087
SwinV2-B (ImageNet)	0.2313 ± 0.0027	0.5849 ± 0.0072	0.1862 ± 0.0041
DaViT-B (ImageNet)	0.2250 ± 0.0037	0.5655 ± 0.0086	0.1781 ± 0.0099
NASA-IBM LFM (ps8, random)	0.2197 ± 0.0027	0.5433 ± 0.0072	0.1781 ± 0.0038
NASA-IBM LFM (ps8)	$\mathbf{0.2541 \pm 0.0018}$	0.6100 ± 0.0032	0.2213 ± 0.0028
NASA-IBM LFM (ps8, LoRA)	0.2539 ± 0.0014	$\mathbf{0.6103 \pm 0.0012}$	$\mathbf{0.2214 \pm 0.0030}$
NASA-IBM LFM (ps8, frozen)	0.1617 ± 0.0026	0.3962 ± 0.0052	0.1081 ± 0.0040
100% training data			
ResNet-50 (ImageNet)	0.2148 ± 0.0018	0.5288 ± 0.0055	0.1762 ± 0.0038
ResNet-50 (random)	0.2077 ± 0.0016	0.5131 ± 0.0017	0.1698 ± 0.0035
ViT-B MAE (ImageNet)	0.2259 ± 0.0054	0.5834 ± 0.0151	0.1785 ± 0.0078
ConvNeXt-B (IN22k)	0.2308 ± 0.0067	0.5830 ± 0.0054	0.1890 ± 0.0137
ConvNeXtV2-B (IN22k)	0.2396 ± 0.0037	0.6025 ± 0.0056	0.2014 ± 0.0080
SwinV2-B (ImageNet)	0.2420 ± 0.0047	0.6020 ± 0.0061	0.2028 ± 0.0086
DaViT-B (ImageNet)	0.2347 ± 0.0029	0.5981 ± 0.0051	0.1912 ± 0.0053
NASA-IBM LFM (ps8, random)	0.2289 ± 0.0037	0.5595 ± 0.0080	0.1893 ± 0.0073
NASA-IBM LFM (ps8)	0.2537 ± 0.0034	0.6156 ± 0.0024	0.2174 ± 0.0072
NASA-IBM LFM (ps8, LoRA)	$\mathbf{0.2581 \pm 0.0017}$	$\mathbf{0.6183 \pm 0.0034}$	$\mathbf{0.2263 \pm 0.0027}$
NASA-IBM LFM (ps8, frozen)	0.1871 ± 0.0024	0.4583 ± 0.0055	0.1317 ± 0.0038

Table 5: Crater detection on LRO NAC (meter-scale) at 100% training data. Test-set bounding-box mean average precision. Values are mean ± standard deviation over five random seeds. **Bold** indicates the best value per column; underline the second best; differences smaller than the reported seed spread should not be read as a ranking.

Model	mAP	AP@50	AP@75
ResNet-50 (ImageNet)	0.1411 ± 0.0036	0.4374 ± 0.0196	0.0895 ± 0.0055
ResNet-50 (random)	0.1296 ± 0.0034	0.4159 ± 0.0171	0.0787 ± 0.0057
ViT-B MAE (ImageNet)	0.1236 ± 0.0119	0.3798 ± 0.0460	0.0703 ± 0.0109
ConvNeXt-B (IN22k)	0.1464 ± 0.0060	0.4400 ± 0.0249	0.0933 ± 0.0039
ConvNeXtV2-B (IN22k)	0.1345 ± 0.0085	0.3949 ± 0.0545	0.0783 ± 0.0107
SwinV2-B (ImageNet)	$\mathbf{0.1552 \pm 0.0086}$	0.4586 ± 0.0308	$\mathbf{0.1090 \pm 0.0146}$
DaViT-B (ImageNet)	0.1501 ± 0.0096	$\mathbf{0.4620 \pm 0.0265}$	0.0983 ± 0.0080
NASA-IBM LFM (ps8, random)	0.1274 ± 0.0151	0.3632 ± 0.0439	0.0818 ± 0.0139
NASA-IBM LFM (ps8)	0.1460 ± 0.0264	0.4565 ± 0.0732	0.0953 ± 0.0318
NASA-IBM LFM (ps8, LoRA)	0.1543 ± 0.0098	0.4416 ± 0.0496	0.1072 ± 0.0116
NASA-IBM LFM (ps8, frozen)	0.1295 ± 0.0038	0.4098 ± 0.0121	0.0756 ± 0.0040

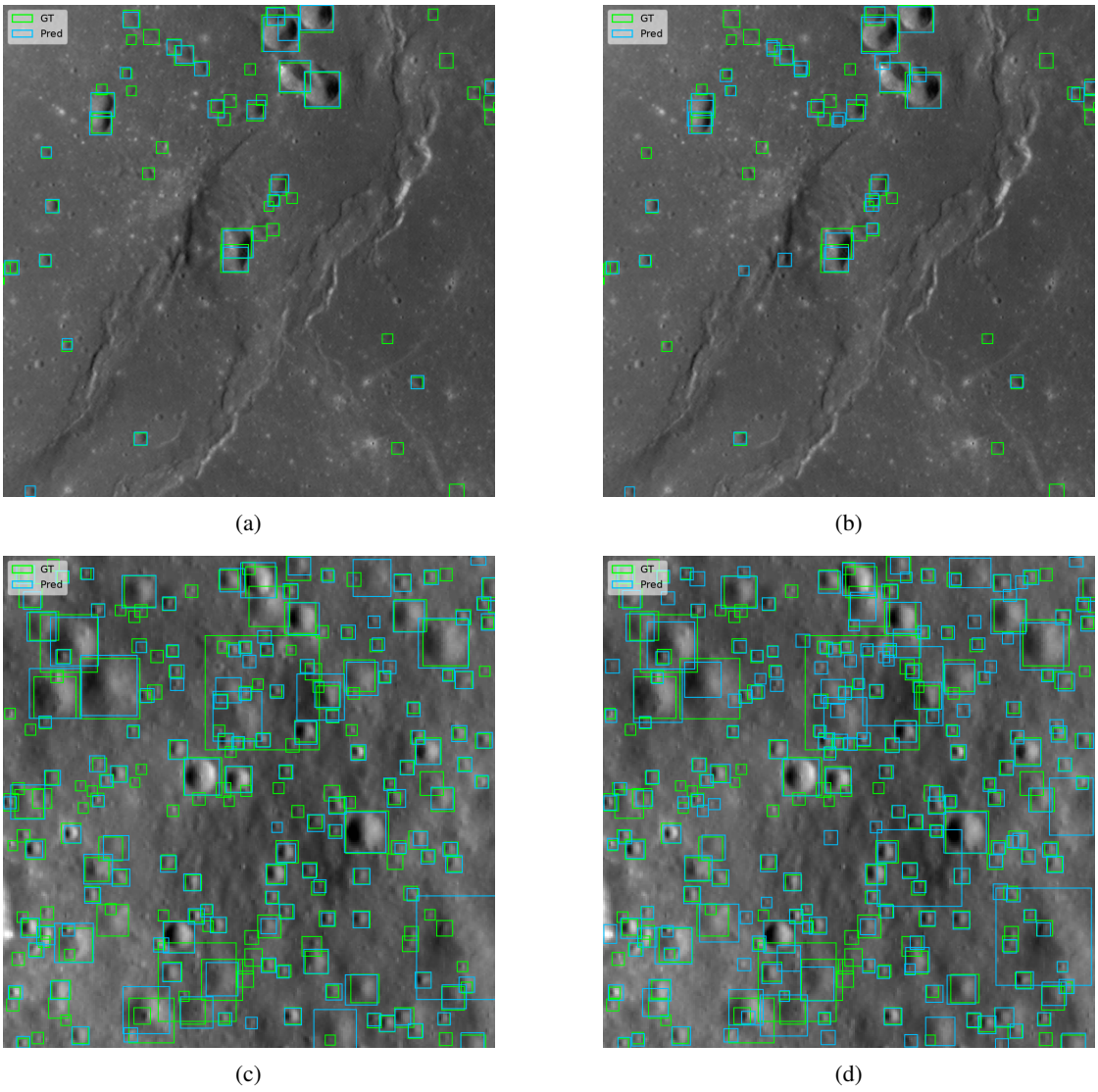


Figure 8: Example of predictions for the WAC Robbins Crater Catalog (first row) and the NAC Hand Labeled Dataset (second row). The first column (*a* and *c*) shows predictions over ground truth for NASA-IBM LFM, with the second column (*b* and *d*) showing the same for SwinV2-B. Green: ground truth; blue: predictions at confidence ≥ 0.5.

Table 6 reports foreground IoU_1 and F1_1 on the IMP segmentation benchmark. The NASA-IBM LFM variants occupy the top positions, with the frozen-backbone configuration that achieves the best IoU_1 (0.5709±0.0114) and F1 (0.7268±0.0093), followed by full fine-tuning (IoU_1 0.5693±0.0114, F1 0.7255±0.0093). These are slightly ahead of the strongest ImageNet baseline, such as ConvNeXt-V2-B (IoU_1 0.5687±0.0181 and F1_1 0.7249±0.0146). NASA-IBM LFM LoRA remains competitive also for this task (IoU_1 0.5593±0.0137, F1 0.7173±0.0113), while the random-init control shows poor performance (IoU_1 0.3142±0.0746), confirming that pretraining is essential for this low-data segmentation task. While a ranking can be established, the performance of the top models falls within the variability observed across runs, indicating that they achieve comparable results.

Figure 9 shows representative qualitative predictions on four IMP examples. Predictions are broadly similar across models, with NASA-IBM LFM generally identifying slightly larger IMP regions than the baselines, which might lead to more false positives. In the second example, this behavior enables our model to capture IMP regions that are not detected by the baseline models.

Table 6: Irregular Mare Patch (IMP) segmentation. IMP delineation is a binary dense-prediction task, so we report test-set metrics on the positive (IMP) class only: intersection-over-union (IoU_1) and F1 score ($F1_1$). Values are mean ± standard deviation over five random seeds. **Bold** indicates the best mean per column and underline the second best; differences smaller than the reported seed spread should not be read as a ranking.

Model	IoU_1	$F1_1$
ResNet-50 (ImageNet)	0.4731 ± 0.0114	0.6423 ± 0.0105
ResNet-50 (random)	0.3496 ± 0.0323	0.5175 ± 0.0355
ViT-B MAE (ImageNet)	0.5541 ± 0.0369	0.7125 ± 0.0298
ConvNeXt-B (IN22k)	0.5374 ± 0.0197	0.6989 ± 0.0168
ConvNeXtV2-B (IN22k)	0.5687 ± 0.0181	0.7249 ± 0.0146
SwinV2-B (ImageNet)	0.5555 ± 0.0159	0.7141 ± 0.0132
DaViT-B (ImageNet)	0.5623 ± 0.0055	0.7198 ± 0.0045
DeepLabV3+ - ResNet-50	0.5154 ± 0.0180	0.6801 ± 0.0157
SegFormer-MiT-B2	0.5475 ± 0.0099	0.7076 ± 0.0083
NASA-IBM LFM (ps8, random)	0.3142 ± 0.0746	0.4742 ± 0.0892
NASA-IBM LFM (ps8)	0.5693 ± 0.0114	0.7255 ± 0.0093
NASA-IBM LFM (ps8, LoRA)	0.5593 ± 0.0137	0.7173 ± 0.0113
NASA-IBM LFM (ps8, frozen)	**0.5709 ± 0.0114**	**0.7268 ± 0.0093**

Polar volatiles

Table 7 reports RMSE, MAE, and R^2 on the full eight-modality ice prospectivity regression benchmark. NASA-IBM LFM (pretrained, full fine-tuning) achieves the lowest RMSE (0.0293±0.0013, $R^2 = 0.9884 \pm 0.0010$), followed by the LoRA variant (RMSE 0.0330±0.0005) and the frozen variant (RMSE 0.0357±0.0007). The best ImageNet-pretrained baseline, SwinV2-B, reaches RMSE 0.0377±0.0004; ConvNeXt-Base follows at RMSE 0.0442 ±0.0008. The pretrained NASA-IBM LFM variants outperform all ImageNet baselines and the random-init control (RMSE 0.0397± 0.0004) across every metric, with a consistent gap well beyond variability across runs. Table 8 reports an ablation in which we start from the two modalities seen during pretraining (aspect and slope) and progressively add polar-stack inputs that neither model ever saw during pretraining. We compare NASA-IBM LFM against ConvNeXt-B, as one of the strongest convolutional baselines. NASA-IBM LFM obtains lower RMSE at every modality count, and the margin is most striking in the sparse-input regime: with only three modalities NASA-IBM LFM reaches RMSE 0.0434±0.0006, matching ConvNeXt-B using the full eight-modality stack (0.0437±0.0002). At m=2 both collapse to RMSE ≈0.2, indicating that slope and aspect alone are insufficient for this task regardless of pretraining.

Table 7: Polar ice prospectivity regression on the full eight-modality polar stack (aspect, slope, DICE, TMAX, LPSR, LPSR_DEN, LPSR_DIS, CUR). Test-set metrics against the prospectivity map in [44]: RMSE and MAE lower-is-better, R^2 higher-is-better. Values are mean ± standard deviation over five random seeds. **Bold** indicates the best value per column; underline the second best.

Model	RMSE ↓	MAE ↓	R^2 ↑
ResNet-50 (ImageNet)	0.0788 ± 0.0004	0.0568 ± 0.0004	0.9162 ± 0.0009
ResNet-50 (random)	0.0695 ± 0.0007	0.0495 ± 0.0005	0.9347 ± 0.0013
ViT-B MAE (ImageNet)	0.0917 ± 0.0009	0.0674 ± 0.0007	0.8865 ± 0.0021
ConvNeXt-B (IN22k)	0.0442 ± 0.0008	0.0301 ± 0.0005	0.9737 ± 0.0009
ConvNeXtV2-B (IN22k)	0.0439 ± 0.0008	0.0298 ± 0.0006	0.9739 ± 0.0010
SwinV2-B (ImageNet)	0.0377 ± 0.0004	0.0256 ± 0.0003	0.9808 ± 0.0004
DaViT-B (ImageNet)	0.0469 ± 0.0007	0.0321 ± 0.0005	0.9704 ± 0.0009
NASA-IBM LFM (ps8, random)	0.0397 ± 0.0004	0.0271 ± 0.0003	0.9787 ± 0.0005
NASA-IBM LFM (ps8)	**0.0293 ± 0.0013**	**0.0197 ± 0.0009**	**0.9884 ± 0.0010**
NASA-IBM LFM (ps8, LoRA)	0.0330 ± 0.0005	0.0224 ± 0.0004	0.9853 ± 0.0004
NASA-IBM LFM (ps8, frozen)	0.0357 ± 0.0007	0.0246 ± 0.0005	0.9828 ± 0.0007

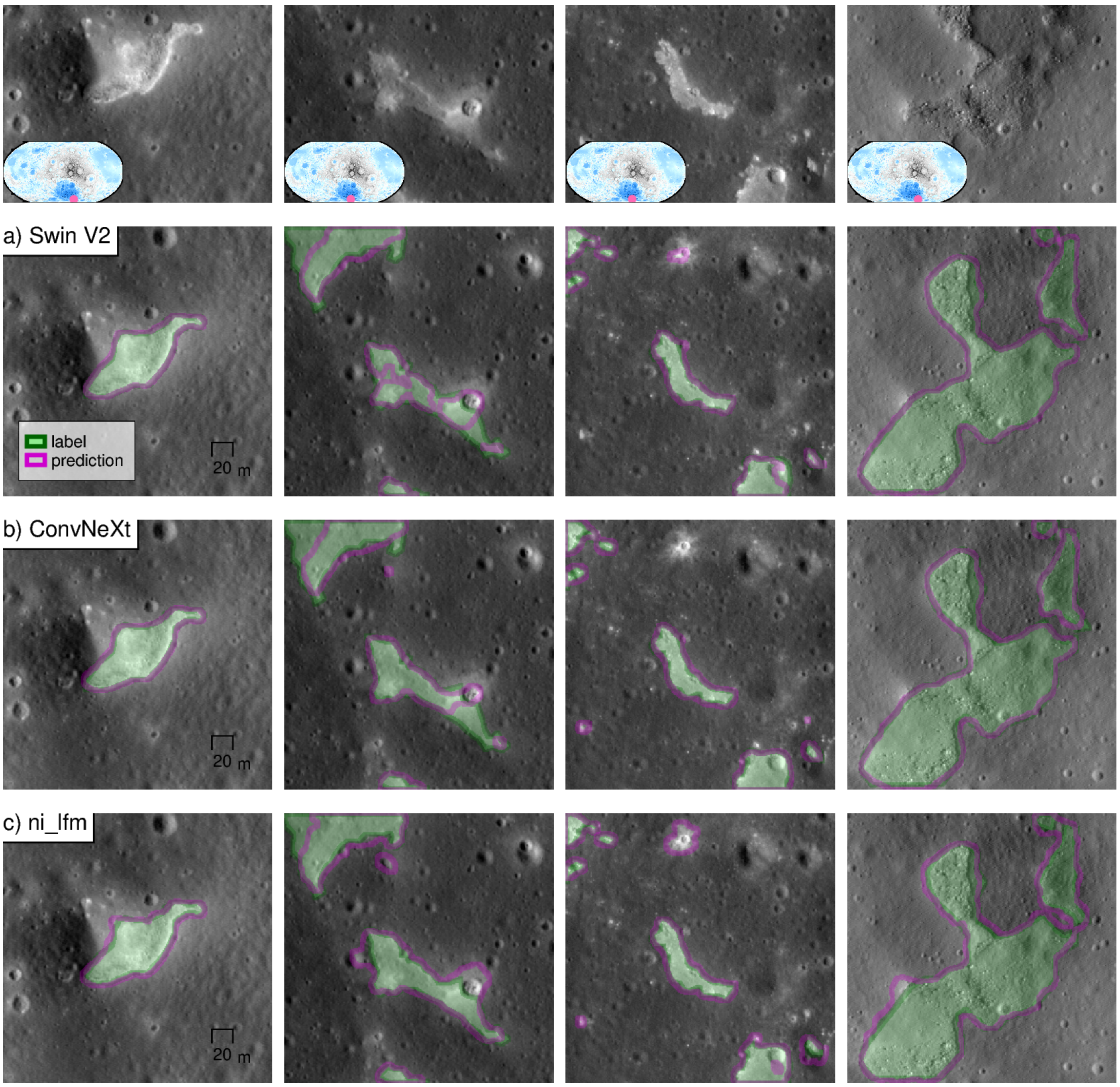


Figure 9: Qualitative IMP segmentation predictions. Our model and the baselines produce broadly comparable outputs; in the second example, the pretrained NASA-IBM LFM identifies IMP extents that none of the baselines detect. Green overlays mark ground-truth IMP extents; colored overlays show model predictions. The colors used in the maps in the top row are for visualization purposes only and do not convey any specific meaning.

Figure 10 shows some predictions for polar volatiles across the south and north pole. The NASA-IBM LFM is able to more accurately preserve details across the different examples compared to ConvNext, even when only a limited number of modalities is used.

6 Conclusion

We have presented a multimodal lunar foundation model built by adapting the TerraMind recipe to the lunar setting and pretraining from scratch on SomBench, a co-registered corpus of nearly two million tile bundles spanning 11 modalities across two spatial scales. Beyond the different data used, two lunar-specific extensions distinguish our model from its Earth-observation ancestor: per-tile acquisition geometry is provided as explicit context input, removing a dominant confound from the pixel-level signal; and NAC- and WAC-anchored tiles are trained jointly in a mixed-batch regime, so a single set of weights absorbs both meter-scale geomorphology and hundred-meter-scale regional context. A FlexiViT patch-embedding

Table 8: Polar ice prospectivity regression, modality-count ablation. Each column corresponds to results from models trained on a fixed set of polar input variables, starting from the two pretrained modalities (aspect, slope) at m=2 and incrementally adding polar-stack modalities not seen during pretraining, in the order DICE, TMAX, LPSR, LPSR_DEN, LPSR_DIS, and finally CUR at m=8. This isolates the contribution of each additional benchmark-only modality on top of the pretrained backbone. Values are test-set RMSE (lower is better). Values are mean ± standard deviation over three random seeds. **Bold** indicates the best value per column.

Model	Number of input modalities m						
	2 (as, sl)	3 +DICE	4 +TMAX	5 +LPSR	6 +DEN	7 +DIS	8 +CUR
ConvNeXt-B (IN22k)	0.2083 (±0.0086)	0.0614 (±0.0007)	0.0522 (±0.0007)	0.0513 (±0.0004)	0.0509 (±0.0005)	0.0517 (±0.0007)	0.0437 (±0.0002)
NASA-IBM LFM (ps8)	**0.1981** (±0.0063)	**0.0434** (±0.0006)	**0.0412** (±0.0009)	**0.0393** (±0.0002)	**0.0403** (±0.0010)	**0.0398** (±0.0004)	**0.0288** (±0.0011)

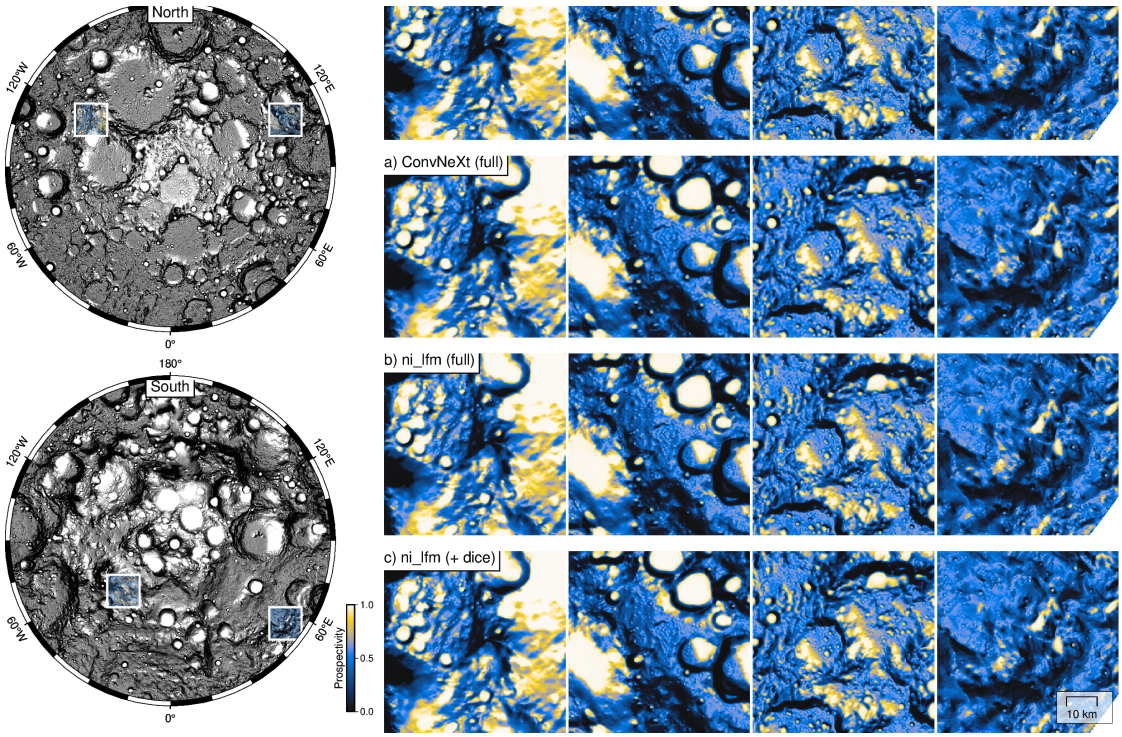


Figure 10: Qualitative ice prospectivity predictions on test patches. NASA-IBM LFM recovers PSR-boundary sharpness and fine spatial detail that the ImageNet-pretrained baseline smooths out. The last row shows our model prediction using only slope, aspect, and DICE, illustrating competitive performance even with a drastically reduced input set.

scheme then lets the same pretrained checkpoint serve downstream tasks at multiple working patch grids without any backbone retraining.

Across four downstream benchmarks the pretrained model matches or exceeds ImageNet-pretrained baselines while also improving over an architecturally identical random-initialization control. The gains are largest where labeled data is scarce: at 50% training data on the WAC crater benchmark, our pretrained variants already exceed the strongest ImageNet baseline trained on the full set. Furthermore, polar ice prospectivity shows the widest margin of all (RMSE 0.0293±0.0013 against 0.0377±0.0004 for the best baseline), well beyond run-to-run variation. This margin reflects two distinct advantages. Our encoder assigns each modality its own pretrained patch-embedding adapter and concatenates the resulting tokens along the sequence axis, whereas the baselines can only accept the eight layers stacked as channels at a

shared stem. The randomly initialized copy of our model isolates the two effects: it shares the tokenization scheme without lunar pretraining and already reaches RMSE 0.0397± 0.0004, ahead of five of the seven ImageNet-pretrained baselines, while pretraining accounts for the remaining gap of 0.0293±0.0013. Our modality-count ablation also shows that the pretraining advantage for ice prospectivity is realized with as few as three input modalities (aspect, slope, and DICE), without requiring the full eight-modality stack. On NAC crater detection and IMP segmentation NASA-IBM LFM places among the leading entries, but the margins are smaller than the seed spread. Therefore, performance of NASA-IBM LFM and baselines can be seen as comparable.

A practical insight can be derived from the adaptation comparison performed in our experiments. LoRA is competitive with full fine-tuning throughout, and on crater detection it is the better choice outright: it tops every column of the WAC benchmark at full training data while also giving the best NASA-IBM LFM mAP on NAC. This is achieved by training a small fraction of the encoder parameters, which also leads to narrower seed spread than full fine-tuning. Full fine-tuning retains an edge only on the two smallest benchmarks, IMP segmentation and ice prospectivity, where the extra capacity to adapt appears to pay off. Freezing the encoder entirely is the least transferable option: it is the strongest configuration on IMP, where 100 training tiles make the reduced parameter count a useful regularizer, but falls below every baseline on crater detection. Based on our experiments, practitioners adapting NASA-IBM LFM to a new lunar task could treat LoRA as the sensible default, reserving full fine-tuning for small-sample regimes and validating a frozen encoder before relying on it.

Alongside quantitative fine-tuning results, the multimodal generation experiments confirm that pretraining has induced the cross-modal correspondences the objective was designed to capture: terrain derivatives from elevation, illumination-consistent reflectance from geometry, and coarse spatial pattern across the full co-registered modality set. These qualitative experiments also make the model's limits apparent: absolute elevation offsets and geodetic coordinates are not reliably recovered, as well as some static maps. This reinforces that the model is intended as a reusable representation for downstream tasks rather than a substitute for physical instruments.

Several limitations bound the present work. The controlled ablations needed to isolate the contribution of geometry tokenization and mixed-resolution training from lunar-domain pretraining as a whole remain to be run. Despite SomBench providing high quality labeled data, some of the benchmarks have limited samples in the test set. However, given the thorough process that went into creating such benchmarks, we believe they still represent an important validation instrument for lunar ML models. Finally, the generation experiments reveal that the model does not maintain a geodetic reference frame, a gap that is relevant to applications requiring absolute positioning. However, in principle, this could be learned during finetuning.

Future work will address these gaps through targeted ablations, extension to the full modality stack, and evaluation on additional planetary bodies where instrument constellations share structural similarities to the lunar case. We release the pretrained checkpoint, fine-tuning code, and dataset as open resources to support reproducibility and community reuse.

Acknowledgments

This work was supported by the National Aeronautics and Space Administration under Award No. 80MSFC25M0084.

Data and Code availability

The data is available through SomBench [20]. All the finetuning code to reproduce our results is available via GitHub.

Appendix

Tokenizer training details

Table 9 shows details about the training configuration for each tokenizer.

Additional generation results

Table 9: Per-run training configuration for the nine domain-specific lunar tokenizers. All runs share a per-GPU batch size of 32.

Tokenizer	Modality	GPUs	Global Batch	Epochs	Warmup Epochs	Peak LR
NAC						
`nac_aspect_3m`	Aspect	16	512	100	2	3.06e-6
`nac_dtm_3m`	DTM	16	512	100	2	3.06e-6
`nac`	Panchromatic	16	512	100	2	3.06e-6
`nac_slope_3m`	Slope	16	512	60	2	3.06e-6
WAC						
`wac_aspect`	Aspect	8	256	100	1	1.57e-6
`wac_dtm`	DTM	16	512	100	1	3.06e-6
`wac_slope`	Slope	4	128	30	1	1.57e-6
`wac_uv`	UV reflectance	4	128	50	1	1.57e-6
`wac_vis`	Visible imagery	20	640	100	2	2.50e-5

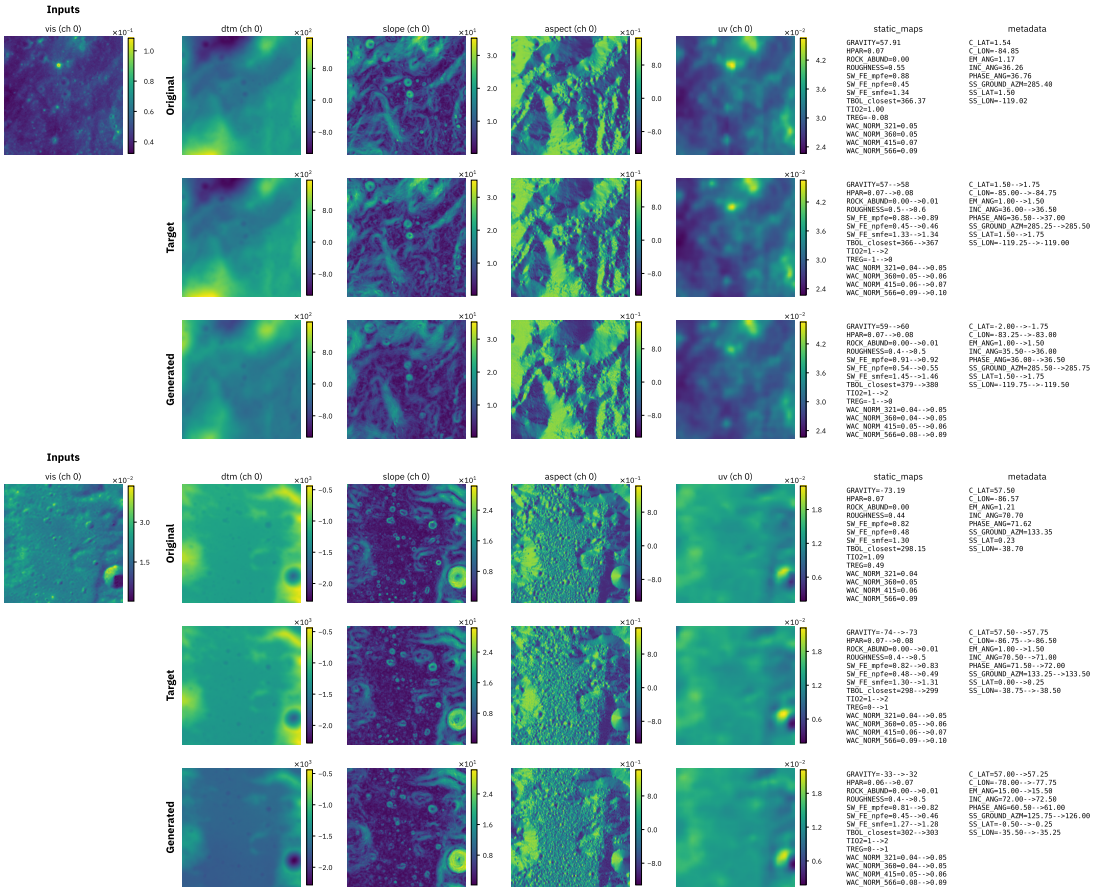


Figure 11: Additional generation examples from WAC to all other modalities.

References

[1] M. S. Robinson, S. M. Brylow, M. Tschimmel, D. Humm, S. J. Lawrence, P. C. Thomas, B. W. Denevi, E. Bowman-Cisneros, J. Zerr, M. A. Ravine, M. A. Caplinger, F. T. Ghaemi, J. A. Schaffner, M. C. Malin, P. Mahanti, A. Bartels, J. Anderson, T. N. Tran, E. M. Eliason, A. S. McEwen, E. Turtle, B. L. Jolliff, and H. Hiesinger. Lunar reconnaissance orbiter camera (lroc) instrument overview. *Space Science Reviews*, 2010.

[2] M. S. Robinson, S. M. Brylow, M. Tschimmel, D. Humm, S. J. Lawrence, P. C. Thomas, B. W. Denevi, E. Bowman-Cisneros, J. Zerr, and M. A. Ravine. Lunar reconnaissance orbiter camera: First results. *Science*, 2010.

[3] D. E. Smith, M. T. Zuber, G. A. Neumann, F. G. Lemoine, E. Mazarico, M. H. Torrence, J. F. McGarry, D. D. Rowlands, J. W. Head, T. H. Duxbury, O. Aharonson, P. G. Lucey, M. S. Robinson, O. S. Barnouin, J. F. Cavanaugh, X. Sun, P. Liiva, D. D. Mao, J. C. Smith, and A. E. Bartels. The lunar orbiter laser altimeter investigation on the lunar reconnaissance orbiter mission. *Space Science Reviews*, 2010.

[4] D. A. Paige, M. C. Foote, B. T. Greenhagen, J. T. Schofield, S. Calcutt, A. R. Vasavada, D. J. Preston, F. W. Taylor, C. C. Allen, K. J. Snook, B. M. Jakosky, B. C. Murray, L. A. Soderblom, B. Jau, S. Loring, J. Bulharowski, N. Bowles, I. R. Thomas, M. T. Sullivan, C. Avis, E. M. De Jong, W. Hartford, and D. J. McCleese. The lunar reconnaissance orbiter diviner lunar radiometer experiment. *Space Science Reviews*, 2009.

[5] S. Nozette, P. Spudis, B. Bussey, R. Jensen, K. Raney, H. Winters, C. Lichtenberg, W. Marinelli, J. Crusan, M. Gates, and M. Robinson. The lunar reconnaissance orbiter miniature radio frequency (mini-rf) technology demonstration. *Space Science Reviews*, 2010.

[6] C. I. Fassett, A. M. Bramson, J. T. S. Cahill, C. D. Harris, G. A. Morgan, C. D. Neish, C. Nypaver, G. W. Patterson, E. G. Rivera-Valentín, P. A. Taylor, and B. J. Thomson. Improved orthorectification and empirical reduction of topographic effects in monostatic mini-rf s-band observations of the moon. *The Planetary Science Journal*, 2024.

[7] M. Ohtake, J. Haruyama, T. Matsunaga, Y. Yokota, T. Morota, C. Honda, and LISM team. Performance and scientific objectives of the selene (kaguya) multiband imager. *Earth, Planets and Space*, 2008.

[8] WC Feldman, BL Barraclough, KR Fuller, DJ Lawrence, S Maurice, MC Miller, TH Prettyman, and AB Binder. The lunar prospector gamma-ray and neutron spectrometers. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 422(1-3):562–566, 1999.

[9] David J. Lawrence, Patrick N. Peplowski, Jack T. Wilson, and Richard C. Elphic. Global hydrogen abundances on the lunar surface. *Journal of Geophysical Research: Planets*, 127(7):e2022JE007197, 2022. e2022JE007197 2022JE007197.

[10] M. T. Zuber, D. E. Smith, D. H. Lehman, T. L. Hoffman, S. W. Asmar, and M. M. Watkins. Gravity field of the moon from the gravity recovery and interior laboratory (grail) mission. *Science*, 2013.

[11] F. Scholten, J. Oberst, Klaus-Dieter Matz, T. Roatsch, M. Wählisch, E. J. Speyerer, and M. S. Robinson. Gld100: The near-global lunar 100 m raster dtm from lroc wac stereo image data. *Journal of Geophysical Research: Planets*, 2012.

[12] M. K. Barker, E. Mazarico, G. A. Neumann, M. T. Zuber, J. Haruyama, and David E. Smith. A new lunar digital elevation model from the lunar orbiter laser altimeter and selene terrain camera. *Icarus*, 2016.

[13] K. N. Burns, E. J. Speyerer, M. S. Robinson, T. Tran, M. R. Rosiek, B. A. Archinal, E. Howington-Kraus, M. R. Edmundson, D. A. Weller, and C. I. Humm. Digital elevation models and derived products from lroc nac stereo observations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2012.

[14] M. R. Henriksen, M. R. Manheim, K. N. Burns, P. Seymour, E. J. Speyerer, A. Deran, A. K. Boyd, E. Howington-Kraus, M. R. Rosiek, B. A. Archinal, and M. S. Robinson. Extracting accurate and precise topography from lroc narrow angle camera stereo observations. *Icarus*, 2016.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *European Conference on Computer Vision*, 2022.

[17] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *arXiv (Cornell University)*, 2022.

[18] Anthony T. Fuller, Koreen Millard, and James R. Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *arXiv*, 2023.

[19] Johannes Jakubik, Sujit Roy, Chris Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Augusto Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dário Augusto Borges Oliveira, Michał Muszyński, Ankur Kumar, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *arXiv (Cornell University)*, 2023.

[20] NASA-IBM AI4Science. SomBench: Lunar multimodal corpus, benchmarks, and foundation model. Hugging Face collection, `https://huggingface.co/collections/nasa-ibm-ai4science/sombench`, 2026. Corpus, benchmark datasets, pretrained checkpoints, and fine-tuning code; to be released upon publication.

[21] S. J. Robbins. A new global database of lunar impact craters >1–2 km: 1. crater locations and sizes, comparisons with published databases, and global analysis. *Journal of Geophysical Research: Planets*, 2019.

[22] Le Qiao, J. W. Head, Zongcheng Ling, and Lionel Wilson. Lunar irregular mare patches: Classification, characteristics, geologic settings, updated catalog, origin, and outstanding questions. *Journal of Geophysical Research: Planets*, 2020.

[23] HI Hargitai and P Brož. Clusters of irregular patches on the moon: A new gis-based catalog. *Icarus*, 429:116439, 2025.

[24] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7383–7394. IEEE, 2025.

[25] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14496–14506. IEEE, 2023.

[26] Carlos Gomes, B. Blumenstiel, João Lucas de Sousa Almeida, Pedro Henrique Amorim Oliveira, P. Fraccaro, Francesc Marti Escofet, Daniela Szwarcman, Naomi Simumba, Romeo Kienzler, and Bianca Zadrozny. Terratorch: The geospatial foundation models toolkit. *IEEE International Geoscience and Remote Sensing Symposium*, 2025.

[27] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

[28] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[29] Yi Wang, Hugo Hernández Hernández, Conrad M. Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[30] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

[31] Marc Girona-Mata, Jakob Gawlikowski, Sumit Goski, Gautier Bardi de Fourtou, Valentin T Bickel, Ben Moseley, Abigail Calzada-Diaz, Sylvester Kaczmarek, and Raúl Ramos-Pollán. Lunarfm: A shared multimodal representation of the moon's surface. *arXiv preprint arXiv:2607.22408*, 2026.

[32] Tom Sander, Moritz Tenthoff, Kay Wohlfarth, and Christian Wöhler. The moon's many faces: A single unified transformer for multimodal lunar reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 236:363–379, 2026.

[33] Song Wang, Zizhu Fan, Zhengming Li, Hong Zhang, and Wei Chao. An effective lunar crater recognition algorithm based on convolutional neural network. *Remote Sensing*, 2020.

[34] Riccardo La Grassa, G. Cremonese, Ignazio Gallo, Cristina Re, and Elena Martellato. Yololens: A deep learning model based on super-resolution to enhance the crater detection of the planetary surfaces. *Remote Sensing*, 2023.

[35] Bo Wu, Wai Chung Liu, A. Grumpe, and Christian Wöhler. Construction of pixel-level resolution dems from monocular images by shape and albedo from shading constrained with low-resolution dem. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017.

[36] Hao Chen, Xuanyu Hu, Philipp Gläser, Haifeng Xiao, Zhen Ye, Hanyue Zhang, Xiaohua Tong, and Jürgen Oberst. Cnn-based large area pixel-resolution topography retrieval from single-view lroc nac images constrained with sldem. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022.

[37] Hao Chen, Xiaojun Hu, and Jürgen Oberst. Pixel-resolution dtm generation for the lunar surface based on a combined deep learning and shape-from-shading (sfs) approach. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022.

[38] Yang Liu, Yexin Wang, Kaichang Di, Man Peng, Wenhui Wan, and Zhaoqin Liu. A generative adversarial network for pixel-scale lunar dem generation from high-resolution monocular imagery and low-resolution dem. *Remote Sensing*, 2022.

[39] Hao Chen, Philipp Gläser, Xuanyu Hu, Konrad Willner, Yongjie Zheng, Friedrich Damme, Lorenzo Bruzzone, and Jürgen Oberst. Elunardtmnet: Efficient reconstruction of high-resolution lunar dtm from single-view orbiter images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[40] Xiaochuan Ma. Feature-enhanced deep learning network for digital elevation model super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[41] Guodong Chen, Yumin Chen, John P. Wilson, Annan Zhou, Yuejun Chen, and Heng Su. An enhanced residual feature fusion network integrated with a terrain weight module for digital elevation model super-resolution. *Remote Sensing*, 2023.

[42] Yi Wang, Shichao Jin, Zekun Yang, H. Guan, Yu Ren, K. Cheng, Xiaoqian Zhao, Xiaoqiang Liu, Mengxi Chen, Yu Liu, and Qinghua Guo. Ttsr: A transformer-based topography neural network for digital elevation model super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[43] T. Heyer, W. Iqbal, A. Oetting, H. Hiesinger, C. H. van der Bogert, and N. Schmedemann. A comparative analysis of global lunar crater catalogs using OpenCraterTool - An open source tool to determine and compare crater size-frequency measurements. *Planetary and Space Science*, 231:105687, July 2023.

[44] Joshua Coyan, Matt Siegler, Jose Martinez-Comacho, Ross Beyer, and Mark Shirley. Prospectivity modeling of the nasa viper landing site at mons mouton near the lunar south pole. *The Planetary Science Journal*, 6(5):105, 2025.

[45] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pages 51772–51783, 2024.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[48] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit,

and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[50] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[51] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. DaViT: Dual attention vision transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

[52] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[53] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.