Landsat30-AU: A Vision-Language Dataset for Australian Landsat Imagery

Sai Ma[1], **Zhuang Li**[2 *], **John A. Taylor**[1]

[1]Australian National University, Australia
[2] Royal Melbourne Institute of Technology, Australia
sai.ma@anu.edu.au, zhuang.li@rmit.edu.au, john.taylor@anu.edu.au

Abstract

Vision language models (VLMs) that enable natural language interaction with satellite imagery can democratize Earth observation by accelerating expert workflows, making data accessible to non-specialists, and enabling planet-scale automation. However, existing datasets focus mainly on short-term, high-resolution imagery from a limited number of satellites, overlooking low-resolution, multi-satellite, long-term archives, such as Landsat, that are essential for affordable and bias-robust global monitoring. We address this gap with Landsat30-AU, a large-scale vision-language dataset built from 30-meter resolution imagery collected by four Landsat satellites (5, 7, 8, and 9) over Australia, spanning more than 36 years. The dataset includes two components: Landsat30-AU-Cap, containing 196,262 image-caption pairs, and Landsat30-AU-VQA, comprising 17,725 human-verified visual question answering (VQA) samples across eight remote sensing domains. Both datasets are curated through a bootstrapped pipeline that leverages generic VLMs with iterative refinement and human verification to ensure quality. Our evaluation of eight VLMs on our benchmark reveals that off-the-shelf models struggle to understand satellite imagery. The open-source remote-sensing VLM EarthDial achieves only **0.07 SPIDEr** in captioning and a VQA accuracy of **0.48**, highlighting the limitations of current approaches. Encouragingly, lightweight fine-tuning of Qwen2.5-VL-7B on LANDSAT30-AU improves captioning performance from **0.11** to **0.31 SPIDEr** and boosts VQA accuracy from **0.74** to **0.87**.

Code — https://github.com/papersubmit1/landsat30-au

1 Introduction

For over fifty years, the *Landsat* program has provided a globally consistent, open-access archive of optical satellite imagery at 30-meter ground-sample distance (GSD) (Wulder et al. 2022). Since 1972, eight Landsat satellites have been launched, each equipped with distinct sensors and band configurations, resulting in varying appearances of standard red-green-blue composites across missions (U.S. Geological Survey 2025). The upcoming *Landsat Next* series will significantly increase daily acquisition volume, from 900 GB (750 scenes) to 8.2 TB (2,220 scenes), through expanded spectral coverage and improved sensor technology (U.S. Geological Survey 2024; NASA Landsat Science 2024). Meanwhile, vision-language models (VLMs) have shown impressive capabilities in managing and interpreting large-scale Earth observation data, especially with high-resolution sources such as Sentinel-2 imagery (Kuckreja et al. 2024; Zhang et al. 2024; Bazi et al. 2024; Yuan et al. 2024). Following these trends, VLMs offer a timely opportunity as natural-language interfaces for long-term, cost-effective, planet-scale analysis using the growing Landsat archive.

Progress is nevertheless constrained by the absence of large-scale image-text corpora that match Landsat's unique regime. Most existing remote-sensing datasets (i) *focus on sub-meter commercial imagery*, which encourages captions centered on fine-grained objects, such as cars, rooftops, or road markings, that are invisible at 30 m resolution, and often come with restrictive licensing costs that hinder global-scale applications (Qu et al. 2016; Ge et al. 2025); (ii) *cover only one or two Landsat satellites*, preventing VLMs from learning the radiometric and band-layout differences that span the full eight-mission Landsat program, and thus limiting their robustness to sensor shifts; and (iii) *include Landsat imagery with only a short temporal span*, depriving models of exposure to long-term seasonal patterns, land-cover change, and climate-driven dynamics critical for temporal generalization. For example, the datasets that *do* incorporate Landsat imagery fall short: EARTHDIAL includes 1.6 million image patches, but only from Landsat 8 (Soni et al. 2025), while SSL4EO-L provides five million multi-temporal patches across several missions, yet lacks the associated textual supervision necessary for vision-language alignment (Stewart et al. 2023). As a result, there is still no dataset that delivers the long-term, multi-satellite, and resolution-aware supervision needed to develop VLMs for scalable and bias-robust Earth monitoring.

Generating high-quality text annotations for remote sensing images also remains a significant challenge. Manual captioning by domain experts ensures high accuracy (Qu et al. 2016; Zhan, Xiong, and Yuan 2023) but does not scale. Crowdsourcing or automatic alternatives, such as OpenStreetMap (OSM) (OpenStreetMap Contributors 2025) tags or web alt-text (Muhtar et al. 2024; Zavras et al. 2025), offer scalability but suffer from two key issues: (i) spa-

*corresponding author

tial mismatch, where many labeled objects (e.g., `clinic`, `cemetery`) are too small to be resolved in 30 m Landsat imagery, and (ii) temporal misalignment, where the metadata may describe a scene years before or after the satellite image was acquired, leading to outdated associations.

To address these limitations, we present **LANDSAT30-AU**, the first large-scale vision-language dataset constructed entirely from 30-meter resolution imagery captured by four Landsat missions (5, 7, 8, and 9) across Australia, spanning from 1988 to 2024. It consists of two parts: (i) LANDSAT30-AU-CAP, containing $196,262$ image-caption pairs, and (ii) LANDSAT30-AU-VQA, comprising $17,725$ human-verified visual question answering (VQA) examples covering eight common remote sensing tasks. To address the challenges of scale and label quality, we develop a semi-automatic bootstrapped pipeline that extends the methodologies of LHRS-ALIGN (Muhtar et al. 2024) and VRS-BENCH (Li, Ding, and Elhoseiny 2024). In this pipeline, generic VLMs generate initial drafts guided by coarse, noisy metadata-spatially and temporally aligned information such as land-cover maps and crowdsourced OSM tags. Successive VLM-assisted refinement steps polish these drafts, and human reviewers remove any text that is visually ungrounded or temporally mismatched. By pairing multi-satellite, multi-decadal Landsat scenes with reliable language supervision, LANDSAT30-AU provides the first solid foundation for training and evaluating VLMs aimed at affordable, long-term Earth monitoring.

Our findings highlight a substantial gap between the capabilities of generic VLMs and the demands of long-term, low-resolution satellite imagery. Off-the-shelf VLMs perform poorly on Landsat-style data. For instance, the open-source VLM EarthDial achieves a captioning score of **0.07 SPIDEr** and an overall VQA accuracy of **0.48**, with particularly low scores of **0.23** on Agro-Phenology Reasoning and **0.10** on Cloud-Occlusion Assessment. However, after lightweight fine-tuning of the Qwen2.5-VL-7B model on our LANDSAT30-AU dataset, performance improves significantly, with captioning scores rising from **0.11** to **0.31 SPIDEr** and VQA accuracy increasing from **0.74** to **0.87**. These results suggest that scalable, cost-effective Earth monitoring with VLMs is feasible, but only when using data that captures Landsat's unique resolution, sensor diversity, and temporal depth.

The main contributions of our work are as follows:

- **LANDSAT30-AU dataset.** A large-scale, open-source vision-language dataset for the Landsat program featuring **30m resolution images**. It includes $196,262$ image-caption pairs and $17,725$ human-verified VQA samples, covering **four Landsat missions** from **1988 to 2024**.
- **Bootstrapped curation pipeline.** A semi-automatic data generation framework that leverages spatially and temporally aligned but noisy metadata (e.g., land-cover maps, OSM tags), generic VLM prompting, iterative refinement, and human verification to produce high-quality captioning and VQA annotations.
- **Comprehensive evaluation.** Benchmarks on eight generic VLMs reveal substantial limitations in both captioning and VQA, especially in spatial reasoning and counting, while fine-tuning on LANDSAT30-AU leads to significant improvements across tasks.

2 Related Work

Generic Vision-Language Datasets

Large-scale image-text datasets play an important role in the development of VLMs. Pioneering VLM datasets like FLICKR30K (Plummer et al. 2016) and MS COCO (Lin et al. 2015) relied on costly human annotation. The SBU CAPTIONED PHOTO DATASET (Ordonez, Kulkarni, and Berg 2011) and CONCEPTUAL CAPTIONS 3M (Sharma et al. 2018) expanded the scale of VLM datasets to several million image-text pairs by using web images and their associated alt-text. Using a similar approach and adding quality control from machine learning models, VLM datasets like CONCEPTUAL 12M (Changpinyo et al. 2021), LAION-5B (Schuhmann et al. 2022), and COYO-700M (Byeon et al. 2022) further increased the scale to hundreds of millions or even billions of pairs. The success of models like CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) demonstrated that even large-scale datasets with noisy information can significantly contribute to VLM development. Many researchers are working to improve dataset quality by using advanced VLMs, like BLIP (Li et al. 2022) and InstructBLIP (Dai et al. 2023), to refine noisy data and generate higher-quality annotations. Furthermore, models such as LLaVA (Liu et al. 2023) and MiniGPT-4 (Zhu et al. 2023) generate synthetic captions to build large-scale training datasets and reduce dataset costs.

Remote-Sensing Vision-Language Datasets

The evolution of remote sensing VLMs has mirrored that of the general VLM community. Datasets like UCM-CAPTIONS and SYDNEY-CAPTIONS (Lu et al. 2018) consisted of only a few hundred images with domain expert labels. To increase dataset scale, NWPU-CAPTIONS (Cheng et al. 2022) and RSICD (Rosario and Noever 2023) retrieved imagery and metadata from commercial map services and then used crowdsourcing to edit the captions. With imagery from open-source satellite platforms, SKYSCRIPT (Wang et al. 2023) and OPENSENTINELMAP (Johnson, Treible, and Crispell 2022) utilized open-source tags from free map services to create captions. However, this approach introduces temporal misalignments between static map tags and dynamic landcover. Following the success of synthetic datasets in general VLMs, remote sensing projects such as RS5M (Zhang et al. 2024), SkySenseGPT (Luo et al. 2024), ChatEarthNet (Yuan et al. 2024), GIT-10M (Liu et al. 2025), and RS-LLaVA (Bazi et al. 2024) have scaled to tens of millions of image-text pairs by using prompted LLMs to synthesize instructions. Meanwhile, task-specific VQA benchmarks such as RSIVQA (Lobry et al. 2020) continue to reveal VLM weaknesses in counting, spatial reasoning, and domain transfer. Despite recent progress, the historical Landsat archive remains underutilized in VLM research. For example, the recent EARTHDIAL (Soni et al. 2025), despite its multi-sensory approach, includes only imagery from

Landsat 8. LANDSAT30-AU addresses this gap by providing images from four Landsat sensors that span from 1988 to 2024, thereby enabling long-term, continental-scale studies with an open-source VLM dataset.

3 Dataset Construction

To tackle the challenges posed by low spatial resolution, sensor diversity, and noisy metadata, we implement a three-stage, human-in-the-loop pipeline (Fig. 1) that steadily produces reliable, resolution-aware textual annotations for Landsat imagery. The stages are: (1) preparing imagery and auxiliary metadata, (2) fine-tuning generic VLMs on Landsat-specific tasks, and (3) generating captions and VQA items through multi-stage refinement.

Stage 1: Imagery and Metadata Preparation

Landsat imagery. We source atmospherically and geometrically corrected imagery from the Digital Earth Australia (DEA) Analysis Ready Data (ARD) archive (Geoscience Australia 2024). We use Bands 4 (Red), 3 (Green), and 2 (Blue) to generate over 400,000 256×256 pixel RGB tiles at 30-meter GSD.

OpenStreetMap tags. OpenStreetMap (OSM) is a crowdsourced geospatial database containing fine-grained vector annotations such as `clinic`, `road`, and `footpath`. For each tile, we extract OSM tags located within its footprint and map them to coarser, Landsat-visible categories using a predefined lookup table (e.g., `clinic` $\rightarrow$ `urban fabric`). These tags provide supplemental semantic cues when the associated objects are large enough to be resolved at 30 m GSD.

Land cover reference. The DEA Land Cover product (Geoscience Australia 2025) provides annually updated, pixel-level classifications (e.g., artificial surfaces, natural bare, water) derived from satellite observations. We extract the dominant land cover class for six fixed spatial regions within each image: top-left, top-right, bottom-left, bottom-right, center, and entire tile. These structured region-level labels support downstream tasks such as region classification and guided captioning.

Stage 2: Fine-tuning VLMs for Landsat Tasks

Generic VLMs are not calibrated for 30 m imagery or Landsat's mission-specific colour shifts. We therefore divide the adaptation process into three lightweight modules: *region classification*, *caption generation*, and *caption review*, and fine-tune each using a small, manually verified subset.

Region classification. Following ChatEarthNet (Yuan et al. 2024), each 256×256 tile is partitioned into six zones: top-left, top-right, bottom-left, bottom-right, center, and entire tile. For each zone, we assign the dominant land-cover class based on the DEA annual land-cover raster (Geoscience Australia 2025), using a taxonomy of coarse land-cover types (e.g., cropland, forest, water, urban).

We manually validate 2,722 such tile-region label sets and fine-tune GPT-4o (gpt-4o-2024-08-06) from OpenAI (OpenAI 2024) on this task. For correctness, we use Subset Accuracy (Godbole and Sarawagi 2004). For set similarity and per-label quality, we employ the Jaccard Index, Precision, Recall, and F1-score. Ranking performance is measured with Label-Ranking Average Precision (LRAP) (Elisseeff and Weston 2001) and nDCG (Järvelin and Kekäläinen 2002). To incorporate error rates, we report (1 - Hamming Loss) and (1 - Ranking Loss), ensuring higher values are consistently better across all metrics. The fine-tuned model achieves Subset Accuracy 0.28 and Jaccard 0.63, outperforming a Qwen2.5-VL-7B (Qwen) (Bai et al. 2025) baseline (Table 1a).

Image captioning. We curated 1,005 image-caption pairs whose text explicitly referenced objects visible at 30 m and aligned with the corresponding acquisition date. All generated image-caption pairs underwent manual review. As shown in Fig. 2a, we used free high-resolution mapping services to verify the presence of key objects. The caption identifying a `golf course` was retained because its presence was confirmed during verification.

We fine-tuned GPT-4.1 (gpt-4.1-2025-04-14) from OpenAI (OpenAI 2025) on this seed dataset, resulting in captions with broader semantic coverage and improved factual grounding. Quantitatively, SPIDEr increased from 0.44 to 0.52, indicating better alignment with reference semantics, while 1-CHAIR-s rose from 0.44 to 0.47, reflecting fewer hallucinated object mentions. The average caption length also increased from 149 to 161 tokens, suggesting greater descriptive depth (Table 1b).

Caption review. From our initial manual review pass, we collect 9,440 image-caption labelled *keep* or *delete*. Qwen2.5-VL-7B is fine-tuned for three epochs on these labels and thereafter prunes any sentence that is visually unsupported or temporally inconsistent, providing an automated hallucination filter for Stage 3.

Together, these three fine-tuned components supply region structure, domain-specific captioning, and scalable quality control, forming the backbone of the multi-stage caption and VQA generation pipeline.

Stage 3: Multi-Stage Caption and VQA Generation

Stage 3 uses the fine-tuned modules from Stage 2 to produce large-scale, quality-controlled annotations (Fig. 1). It involves two tasks: caption refinement and VQA generation. The caption refinement task combines model-generated drafts with VLM-based verification to ensure resolution-awareness and factual consistency, while the VQA generation task incorporates human verification to ensure answer accuracy and to increase the difficulty and diversity of the questions and options.

Caption refinement. For each image tile, the captioning model (fine-tuned GPT-4.1) first generates an *Initial* caption conditioned on region labels, OSM tags, and the image. We then prompt Qwen2.5-VL-7B to augment the caption with missing objects and spatial relations, resulting in an *Extra* version. Next, the caption reviewer module prunes hallucinated or temporally inconsistent content, producing the final

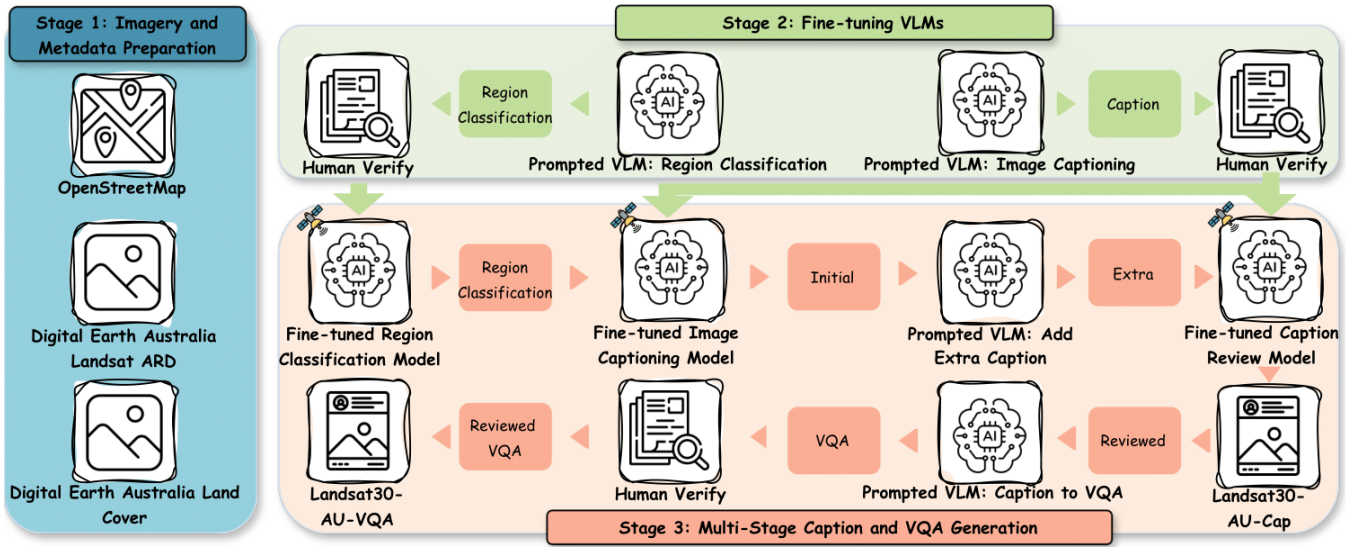


Figure 1: Overview of the LANDSAT30-AU dataset construction pipeline. Stage 1: Sources Landsat imagery and collects metadata. Stage 2: Adapts VLMs into specialized modules for region classification, caption generation, and review. Stage 3: Produces large-scale annotations via iterative VLM refinement and human verification.

Model	Subset Acc↑	Jaccard↑	Precision↑	Recall↑	F1↑	LRAP↑	nDCG↑	1-hamming-loss↑	1-ranking-loss↑
GPT-4o	**0.278**⋆	**0.630**⋆	**0.768**	**0.722**	**0.727**⋆	**0.826**⋆	**0.917**	**0.783**⋆	**0.705**
Qwen	0.220	0.609	0.735	**0.743**⋆	0.720	0.818	0.912	0.762	**0.710**⋆
GPT-4o w/o ft	**0.262**	**0.612**	**0.805**⋆	0.676	0.715	**0.816**	**0.919**⋆	**0.776**	0.667
Qwen w/o ft	0.099	0.450	0.585	0.588	0.563	0.708	0.834	0.653	0.539

(a) Multi-label region classification metrics on the fine-tune set.

Model	BLEU-4↑	SPIDEr↑	BERT-F1↑	1-CHAIR-s↑	1-CHAIR-i↑	Avg. Cap. Len.
GPT-4.1 w/o ft (Initial)	0.160	0.440	0.902	0.438	0.843	149
GPT-4.1 w/o ft (Extra)	0.163	0.440	0.896	0.423	0.837	206
GPT-4.1 w/o ft (Reviewed)	0.152	0.438	0.901	**0.522**⋆	**0.864**⋆	140
GPT-4.1 (Initial)	**0.188**⋆	**0.510**	**0.905**⋆	0.428	0.841	161
GPT-4.1 (Extra)	0.173	**0.510**	0.897	0.358	0.828	217
GPT-4.1 (Reviewed)	**0.184**	**0.517**⋆	**0.903**	**0.473**	**0.853**	161

(b) Captioning metrics on the fine-tune set.

Table 1: Evaluation of model performance on the fine-tuning set, comparing models before (w/o ft) and after fine-tuning. The best score in each metric is marked with a star (⋆) and the top two scores are in **bold**.

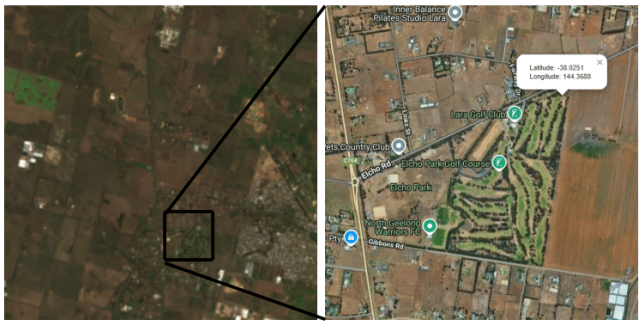
Reviewed caption. To evaluate the impact of each stage, we score all three versions (*Initial*, *Extra*, and *Reviewed*) on a held-out reference set using BLEU-4 (Papineni et al. 2002), SPIDEr (Liu et al. 2017), and BERTScore-F1 (Zhang et al. 2020) for semantic quality, and CHAIR-s/i (Rohrbach et al. 2019) for hallucination. As shown in Table 1b, the *Reviewed* captions provide the best overall balance (SPIDEr 0.517; 1-CHAIR-i 0.853), and are used throughout the dataset. This process yields 196,262 high-quality captions, which make up the LANDSAT30-AU-CAP dataset.

VQA generation. To construct LANDSAT30-AU-VQA, we prompt GPT-4.1 to generate multiple-choice questions (MCQs) from 9,735 captioned images. Each MCQ is designed to assess one of eight Landsat-specific reasoning tasks (see Table 2 and Fig. 3). Human reviewers then refine the questions by correcting ambiguous phrasing, replacing weak distractors, and discarding low-quality items. As shown in Fig. 2b, original VQA from GPT-4.1 confuses the width of a linear feature and incorrectly classified it as a highway. We intentionally included such errors as incorrect options to force finer distinctions. Example question-answer pairs are shown in Table 2 and Fig. 3 This results in 17,725 validated question-answer pairs.

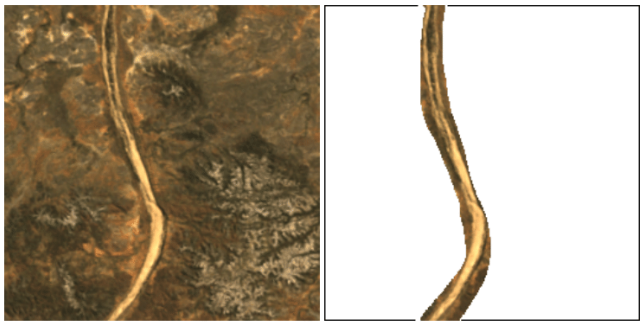
Together, these two processes complete the LANDSAT30-AU corpus, providing resolution-aware, multi-sensor, and temporally grounded textual supervision for training and evaluating VLMs on real-world satellite imagery.

4 Landsat30-AU Dataset

This section details the two LANDSAT30-AU sub-datasets, providing key statistics and a comparison with existing remote sensing vision-language corpora.



(a) A golf course appears in the image. Decision: Keep.

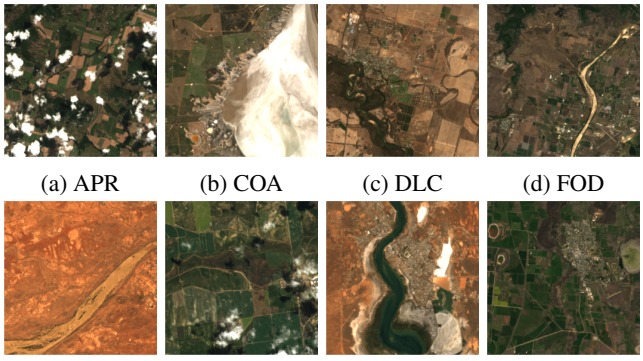


(b) Which objects in the image? Fix: from highway to river.

Figure 2: Examples of the human verification process. (a) A correct caption is kept. (b) An incorrect answer is fixed.

Type	# VQA	Task Focus	Example (Fig.3)
APR	2,102	Crop-season inference from field texture	*Fig. 3a*: "Wet or dry season?" Options: wet_season, **dry_season**
COA	2,129	Cloud/haze usability assessment	*Fig. 3b*: "Scene usable despite cloud?" Options: **Fully usable**, Not usable
DLC	2,479	Dominant land-cover type	*Fig. 3c*: "Main cover type?" Options: Urban, Forest, **Field**
FOD	2,000	Detectability of thin man-made objects	*Fig. 3d*: "Prominent thin structure?" Options: Railway, Pipeline, **None**
MOP	2,418	Presence of macro-objects	*Fig. 3e*: "Which object is visible?" Options: Railway, Large building, **River**
NUM	2,244	Numerosity estimation	*Fig. 3f*: "Water-body count?" Options: Four, Two, Three, **Zero**
SRI	2,419	Spatial-relation inference	*Fig. 3g*: "River vs. urban fabric?" Options: Only south, **Both sides**, Only north
USR	1,934	Urban-scale recognition	*Fig. 3h*: "Settlement type?" Options: Major city, **Small town**, Rural

Table 2: The LANDSAT30-AU-VQA taxonomy. This table outlines the eight question categories, their respective task focus, and an example for each. The correct answer in the examples is shown in **bold**.



(a) APR (b) COA (c) DLC (d) FOD

(e) MOP (f) NUM (g) SRI (h) USR

Figure 3: Landsat30-AU-VQA categories.

Landsat30-AU-Cap. LANDSAT30-AU-CAP consists of **196,262** image-caption pairs aligned with **low-resolution** Landsat imagery from **four satellites** spanning 36 years (**1988-2024**). Each caption is visually grounded and tailored to Landsat's spatial resolution, offering detailed semantic content that reflects the constraints and opportunities of low-resolution Earth observation. This dataset supports training and evaluation of captioning models on real-world, multi-sensor, multi-temporal satellite imagery.

Landsat30-AU-VQA. LANDSAT30-AU-VQA contains **17,725** multiple-choice question-answer pairs covering **eight** remote sensing tasks designed to capture common reasoning challenges in low-resolution imagery. These include: (1) inferring cropping season from field texture (Agro-Phenology Reasoning, APR), (2) evaluating cloud and haze interference (Cloud-Occlusion Assessment, COA), (3) identifying dominant land-cover types (Dominant Land Cover, DLC), (4) detecting thin or sub-pixel structures (Fine-Object Detectability, FOD), (5) identifying large visible features (Macro-Object Presence, MOP), (6) estimating object counts (Numerosity, NUM), (7) reasoning about spatial layout (Spatial-Relation Inference, SRI), and (8) classifying settlement scale (Urban-Scale Recognition, USR). Examples are in Table 2 and Fig. 3.

Comparison with Remote-Sensing VLM Datasets

We compare LANDSAT30-AU to six key remote sensing vision-language datasets: RSICD (Rosario and Noever 2023), SKYSCRIPT (Wang et al. 2023), CHATEARTHNET (Yuan et al. 2024), GIT-10M (Liu et al. 2025), GAIA (Zavras et al. 2025), and EARTHDIAL (Soni et al. 2025).

Scope and diversity. Table 3 compares datasets across five key metrics: total images (# img), the number of Landsat images (# LS image) and number of source Landsat satellites (# LS Sats), whether the imagery is georeferenced (Geo-loc.), and the Landsat imagery temporal span (Span), highlighting differences in scale, Landsat imagery diversity, and spatio-temporal coverage.

While EarthDial offers a larger number of Landsat images (1.6 million), it is restricted to a single satellite (Landsat 8) and lacks geolocation metadata. In contrast, LANDSAT30-

Dataset	# img/LS img	# LS Sats.	Geo-loc.	Span
RSICD	10k/0	0	No	-
SkyScript	5M/15k	2	Yes	2013-2023
ChatEarthNet	173k/0	0	No	-
Git-10M	16M/0	0	Yes	-
GAIA	41k/2k	2	Yes	2013-2024
EarthDial	11M/1.6M	1	No	2013-2024
Landsat30-AU	**196k/196k**	**4**	**Yes**	**1988-2024**

Table 3: LANDSAT30-AU vs. other remote sensing VLM datasets. *Span* is blank when no Landsat imagery is present.

AU spans four Landsat satellites (Landsat 5, 7, 8, and 9) over a 36-year period (1988-2024), with each image accompanied by precise geographic coordinates and acquisition dates. This rich spatiotemporal coverage enables models to learn from diverse sensor characteristics and location-aware patterns, making LANDSAT30-AU uniquely suited for multi-sensor, long-term Earth observation tasks.

Linguistic and semantic richness. Table 4 presents a comparison of caption length and lexical diversity across Landsat-related datasets. EarthDial does not include captions, and SkyScript provides only very short ones, averaging 9.3 words. GAIA offers high-quality captions, with an average length of 183.3 words and strong lexical diversity as measured by the Mean Segmental Type-Token Ratio (MSTTR) at 0.84. However, it includes only around 2,000 image-caption pairs. In contrast, LANDSAT30-AU provides **196,262** captions with both scale and linguistic richness, featuring an average length of 165.4 words and 0.82 MSTTR.

Dataset	LS Pairs	Vocab	Avg. Cap. Len.	MSTTR ↑
SkyScript	15k	1,049	9.3	-
GAIA	2k	2,325	**183.3** ⋆	**0.84** ⋆
EarthDial	**1.6M** ⋆	**9,251** ⋆	-	-
Landsat30-AU	**196k**	**4,405**	**165.4**	**0.82**

Table 4: Linguistic properties of Landsat-related datasets. The best score in each metric is marked with a star (⋆), and the top two are in **bold**.

5 Benchmark Evaluation

Task Settings. LANDSAT30-AU includes two distinct tasks for evaluating Landsat imagery understanding:

- **Image-Captioning:** This is a generative captioning task requiring VLMs to produce detailed descriptions of Landsat images. We use a test set of $1,005$ human-verified image-caption pairs from Stage 2 image captioning and compare the VLM-generated captions against reference captions using BLEU-4, SPIDEr, BERT-F1, 1-CHAIR-s, 1-CHAIR-i, and Average Caption Length.
- **VQA:** A multiple-choice VQA task that evaluates a model's ability to understand Landsat imagery content, to infer information beyond the visual data, and address challenges specific to 30-meter GSD. We report **per-category accuracy** across eight VQA categories. We use a 15% split of LANDSAT30-AU-VQA as the test set.

Implementation Details. To structure our evaluation, we group the models based on their training domain. The Specialized category comprises remote sensing VLMs, including EarthDial (Soni et al. 2025) and RS-LLaVA (Bazi et al. 2024), and reasoning VLMs, such as GLM-4.1V (GLM-V) (Team et al. 2025c) and MiMo-VL (MiMo) (Team et al. 2025a). The General category consists of foundational models like Qwen2.5-VL (Qwen), Gemma 3 (Gemma3) (Team et al. 2025b), Llama-3.2 (Llama) (Grattafiori et al. 2024), and LLaVA-OneVision (LLaVA) (Li et al. 2024). We ran the two reasoning models in a zero-shot setting, enforcing a maximum output of 8,192 tokens, while the remaining models were evaluated in a one-shot setting.

Furthermore, we fine-tuned two of the general models, Qwen and Llama (creating Qwen-ft and Llama-ft), using LoRA (Hu et al. 2021) on 15% of the respective training data for each task (LANDSAT30-AU-CAP for captioning and LANDSAT30-AU-VQA for VQA).

RQ1: How do Specialized VLMs perform compared to General models?

Settings. We analyze the performance of Specialized VLMs including remote sensing VLMs (EarthDial, RS-LLaVA) and reasoning VLMs (GLM-V, MiMo) against general models (without fine-tune).

Results. The specialized models exhibit distinct trade-offs. RS-LLaVA proves to be a competent semantic captioner, while EarthDial lags significantly (Table 5a). Both models show strong sentence-level hallucination control, suggesting a shared cautiousness in their design. However, their VQA performance reveals critical flaws: EarthDial fails on tasks like APR, COA, NUM, SRI, and USR, while RS-LLaVA surprisingly struggles with fundamental SRI and USR, achieving the lowest score of all models.We hypothesize this stems from a domain mismatch between their training corpora and our Landsat imagery. The reasoning VLM MiMo achieves the second-highest overall VQA score (0.7555), notably excelling in NUM and MOP, showcasing the value of its chain-of-thought capabilities. However, its verbose captions lead to the worst sentence-level hallucination rate of 0.3831 on the 1-CHAIR-s metric. GLM-V is the most factually grounded captioner with the best hallucination score, but its VQA performance is unremarkable. Ultimately, neither remote sensing nor reasoning VLMs demonstrate the consistent, all-around competence required for robust Landsat imagery analysis, as the stark performance divergence even within the same category reveals strong, conflicting biases inherited from their unique training domains.

RQ2: Can fine-tuning improve VLM performance in Landsat imagery understanding?

Settings. We compare the performance of the base Qwen and Llama models against their fine-tuned ones (Qwen-ft, Llama-ft) on both the captioning and VQA tasks.

Type	Model	Size	BLEU-4↑	SPIDEr↑	BERTScore-F1↑	1-CHAIR-s↑	1-CHAIR-i↑	Avg. Cap. Len.
Specialized	EarthDial	4B	0.0210	0.0726	0.8379	**0.5920**	0.8197	140
	RS-LLaVA	7B	0.0975	0.2095	0.8874	**0.5920**	0.8119	139
	MiMo	7B	0.0338	0.0958	0.8601	0.3831	0.7805	168
	GLM-V	9B	0.0420	0.1177	0.8668	**0.6259***	**0.8496**	155
General	Qwen	7B	0.0350	0.1114	0.8693	0.4697	0.7959	124
	LLaVA	8B	0.0258	0.1286	0.8643	0.5483	0.8437	103
	Llama	11B	0.0726	0.1695	0.8800	0.5483	0.8296	147
	Gemma 3	12B	0.0542	0.1246	0.8751	0.3572	0.8019	149
General with ft	Qwen-ft	7B	**0.1395***	**0.3054***	**0.8935***	0.4657	**0.8549***	157
	Llama-ft	11B	**0.1129**	**0.2767**	**0.8914**	0.5224	0.8016	124

(a) Performance on the image captioning task.

Type	Model	Size	APR↑	COA↑	DLC↑	FOD↑	MOP↑	NUM↑	SRI↑	USR↑	Overall
Specialized	EarthDial	4B	0.2349	0.1034	0.7527	**0.9900**	0.6116	0.4362	0.5124	0.1552	0.4829
	RS-LLaVA	7B	**0.6857**	0.8088	0.7124	0.8700	0.6309	0.4985	0.2617	0.1034	0.5724
	MiMo	7B	0.4000	0.4577	0.9247	0.9333	**0.8430**	**0.6142**	**0.9421***	0.8897	**0.7555**
	GLM-V	9B	0.4571	0.3636	0.7285	0.6267	0.6749	0.5863	0.6997	0.8828	0.6287
General	Qwen	7B	0.2984	**0.8966**	**0.9409**	0.7167	0.7603	0.5312	0.9284	0.8207	0.7428
	LLaVA	8B	0.3937	0.7900	0.8306	0.5900	0.7245	0.4659	0.8512	0.1034	0.6096
	Llama	11B	0.3111	0.8558	0.6022	0.6633	0.7135	0.5757	0.8953	0.1034	0.6025
	Gemma 3	12B	0.6730	0.8150	0.9220	0.4533	0.7934	0.3234	0.9311	**0.9310***	0.7356
General with ft	Qwen-ft	7B	**0.7016***	**0.9530***	**0.9651***	**1.0***	**0.8678***	**0.6588***	0.9229	**0.8966**	**0.8710***
	Llama-ft	11B	0.5238	0.8558	0.8682	**1.0***	0.8402	0.6024	**0.9339**	0.1276	0.7315

(b) Performance on the VQA task, reported as accuracy per category

Table 5: Evaluation of VLMs on Landsat30-AU. Bold indicates a top-2 score. * indicates the best score.

Results. As shown in Table 5, fine-tuning provides a decisive performance boost on both models. On the captioning task, Qwen-ft achieves state-of-the-art results, leading in BLEU-4 (0.1395), SPIDEr (0.3054), and BERTScore-F1 (0.8935), while simultaneously demonstrating strong hallucination control, with a 1-CHAIR-i score of 0.8549. While Llama-ft also saw a substantial 63% gain in its SPIDEr score, it revealed a nuanced trade-off, with a slight increase in object hallucination. The most compelling evidence lies in the VQA tasks, where fine-tuning specifically improved performance on domain-specific challenges. For instance, accuracy on APR more than doubled for Qwen-ft, while both fine-tuned models achieved perfect scores on FOD, effectively learning the resolution limits of the imagery. Qwen-ft achieves the highest overall accuracy (0.8710) and secures top scores in six of the eight reasoning categories. These results unequivocally demonstrate that even limited, efficient fine-tuning is critical for adapting VLMs to the specific visual and logical challenges of Landsat imagery analysis.

RQ3: What are the strengths and weaknesses of VLMs on Landsat imagery?

Settings. We analyze the per-category VQA accuracies across all VLMs in Table 5b.

Results. Models consistently excel at direct perceptual tasks, such as identifying dominant land cover (DLC), confirming the presence of macro-objects (MOP), or correctly assessing the absence of sub-pixel features (FOD). This indicates a strong baseline for grounded visual recognition.

However, performance degrades significantly as tasks demand more abstract or contextual reasoning. Numerosity (NUM) emerges as a universal bottleneck across all models. Similarly, tasks requiring contextual assessment of the entire scene, such as judging urban scale (USR) or cloud usability (COA), produce highly polarized results, suggesting that only some models have learned the necessary holistic interpretation skills. The most abstract reasoning tasks, like inferring seasonality from texture (APR) or deducing complex spatial relationships (SRI), remain the most challenging and are often the primary beneficiaries of targeted fine-tuning. This pattern suggests that while current VLMs have mastered direct perception for Landsat imagery.

6 Conclusion

We introduce a new dataset derived from optical imagery across the Landsat 5, 7, 8, and 9 missions. This dataset is organized into two subsets: LANDSAT30-AU-CAP, containing 196,262 image-captioning pairs, and LANDSAT30-AU-VQA, comprising 17,725 human-verified VQA samples. Both components are built from 30-meter resolution Landsat imagery and are curated to facilitate VLM training and validation within this specific domain. Our benchmark evaluation reveals that while fine-tuning is critical for adapting models, significant challenges remain in complex tasks, thereby highlighting key areas for future VLM development.

References

Bai, S.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bazi, Y.; Bashmal, L.; Al Rahhal, M. M.; Ricci, R.; and Melgani, F. 2024. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery. *Remote Sensing*, 16(9).

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. arXiv:2102.08981.

Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; and Wang, Z. 2022. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–19.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Elisseeff, A.; and Weston, J. 2001. A Kernel Method for Multi-Labelled Classification. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'01, 681–687. Cambridge, MA, USA: MIT Press.

Ge, J.; Zhang, X.; Zheng, Y.; Guo, K.; and Liang, J. 2025. RSTeller: Scaling up Visual Language Modeling in Remote Sensing with Rich Linguistic Semantics from Openly Available Data and Large Language Models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226: 146–163.

Geoscience Australia. 2024. What is Analysis Ready Data? https://www.ga.gov.au/scientific-topics/dea/about/what-is-analysis-ready-data. Page last updated 18 June 2024. Accessed: 2025-07-28.

Geoscience Australia. 2025. DEA Land Cover (Landsat). Derivative raster dataset, version 2.0.0. Covers 1988–2024; Creative Commons Attribution 4.0 Licence; DOI 10.26186/149976; accessed 23 June 2025.

Godbole, S.; and Sarawagi, S. 2004. Discriminative Methods for Multi-Labeled Classification. In Dai, H.; Srikant, R.; and Zhang, C., eds., *Advances in Knowledge Discovery and Data Mining*, 22–30. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-24775-3.

Grattafiori, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918.

Johnson, N.; Treible, W.; and Crispell, D. 2022. OpenSentinelMap: A Large-Scale Land Use Dataset using OpenStreetMap and Sentinel-2 Imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1332–1340.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4): 422–446.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27831–27840.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. arXiv:2406.12384.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, C.; Chen, K.; Zhao, R.; Zou, Z.; and Shi, Z. 2025. Text2Earth: Unlocking Text-Driven Remote Sensing Image Generation with a Global-Scale Dataset and a Foundation Model. *IEEE Geoscience and Remote Sensing Magazine*, 2–23.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved Image Captioning via Policy Gradient Optimization of SPIDEr. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 873–881. IEEE.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4): 2183–2195.

Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; and Li, Y. 2024. SkySenseGPT: A Fine-Grained Instruction Tuning Dataset and Model for Remote Sensing Vision-Language Understanding. arXiv:2406.10100.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. arXiv:2402.02544.

NASA Landsat Science. 2024. Landsat Next. https://landsat.gsfc.nasa.gov/satellites/landsat-next/mission-details/. Accessed: 2025-05-17.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. GPT-4.1 [Large Language Model]. https://platform.openai.com/docs/models/gpt-4.1.

OpenStreetMap Contributors. 2025. OpenStreetMap: Planet Dump. Data file, available from https://planet.openstreetmap.org/. Frequently updated global dataset under ODbL; accessed 23 June 2025.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. arXiv:1505.04870.

Qu, B.; Li, X.; Tao, D.; and Lu, X. 2016. Deep Semantic Understanding of High Resolution Remote Sensing Image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–5.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. arXiv:1809.02156.

Rosario, G.; and Noever, D. 2023. Satellite Captioning: Large Language Models to Augment Labeling. arXiv:2312.10905.

Schuhmann, C.; et al. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. arXiv:2210.08402.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1896–1906. Melbourne, Australia: Association for Computational Linguistics.

Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M. A.; Danish, M. S.; Fraccaro, P.; Watson, C. D.; Klein, L. J.; Khan, F. S.; and Khan, S. 2025. EarthDial: Turning Multi-Sensory Earth Observations to Interactive Dialogues. arXiv:2412.15190.

Stewart, A. J.; Lehmann, N.; Corley, I. A.; Wang, Y.; Chang, Y.-C.; Braham, N. A. A.; Sehgal, S.; Robinson, C.; and Banerjee, A. 2023. SSL4EO-L: Datasets and Foundation Models for Landsat Imagery. arXiv:2306.09424.

Team, C.; et al. 2025a. MiMo-VL Technical Report. arXiv:2506.03569.

Team, G.; et al. 2025b. Gemma 3 Technical Report. arXiv:2503.19786.

Team, V.; et al. 2025c. GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv:2507.01006.

U.S. Geological Survey. 2024. Landsat Next. https://www.usgs.gov/landsat-missions/landsat-next. Accessed: 2025-05-17.

U.S. Geological Survey. 2025. What Are the Band Designations for the Landsat Satellites? https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites. Accessed: 2025-07-20.

Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; and Rajagopal, R. 2023. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. arXiv:2312.12856.

Wulder, M. A.; et al. 2022. Fifty Years of Landsat Science and Impacts. *Remote Sensing of Environment*, 280: 113195.

Yuan, Z.; Xiong, Z.; Mou, L.; and Zhu, X. X. 2024. ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models. arXiv:2402.11325.

Zavras, A.; Michail, D.; Zhu, X. X.; Demir, B.; and Papoutsis, I. 2025. GAIA: A Global, Multi-Modal, Multi-Scale Vision-Language Dataset for Remote Sensing Image Analysis. arXiv:2502.09598.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024. RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.