Towards Unified Vision Language Models for Forest Ecological Analysis in Earth Observation

Xizhe Xue[1,2], **Xiao Xiang Zhu**[1,2*],

[1]Technical University of Munich
[2]Munich Center for Machine Learning
xizhe.xue@tum.de, xiaoxiang.zhu@tum.de

Abstract

Recent progress in vision language models (VLMs) has enabled remarkable perception and reasoning capabilities, yet their potential for scientific regression in Earth Observation (EO) remains largely unexplored. Existing EO datasets mainly emphasize semantic understanding tasks such as captioning or classification, lacking benchmarks that align multimodal perception with measurable biophysical variables. To fill this gap, we present REO-Instruct, the first unified benchmark designed for both descriptive and regression tasks in EO. REO-Instruct establishes a cognitively interpretable logic chain in forest ecological scenario: human activity → land-cover classification → ecological patch counting → above-ground biomass (AGB) regression, bridging qualitative understanding and quantitative prediction. The dataset integrates co-registered Sentinel-2 and ALOS-2 imagery with structured textual annotations generated and validated through a hybrid human–AI pipeline. Comprehensive evaluation protocols and baseline results across generic VLMs reveal that current models struggle with numeric reasoning, highlighting an essential challenge for scientific VLMs. REO-Instruct offers a standardized foundation for developing and assessing next-generation geospatial models capable of both description and scientific inference.

Code — https://github.com/zhu-xlab/REO-Instruct

Introduction

Earth Observation (EO) data has become a essential resource for diverse scientific research, crucial in areas such as ecological monitoring (Allen et al. 2025), weather forecasting (Nguyen et al. 2024), disaster response and population dynamics analysis (Batista e Silva et al. 2020). These researches require not only a thorough understanding of image content but also precise regression of the real-world scientific attributes they represent (Benson et al. 2024; Hamilton et al. 2024). Regression, in this context, involves modeling the relationship between environmental metrics and relevant features, enabling accurate predictions and deeper insights into complex geospatial or ecological processes.

Against this backdrop, the emergence of Vision Language Models (VLMs) (Ouyang et al. 2022; Radford et al. 2021; Liu et al. 2024b; Achiam et al. 2023) offers promising directions for advancing EO analysis. As illustrated in Figure 1(a), existing VLM applications in EO focus mainly on perception tasks (e.g., detection) and image content description tasks (e.g., VQA). However, the potential of VLMs for scientific regression tasks, such as predicting environmental attributes, has received limited attention, despite its essential role in EO applications.

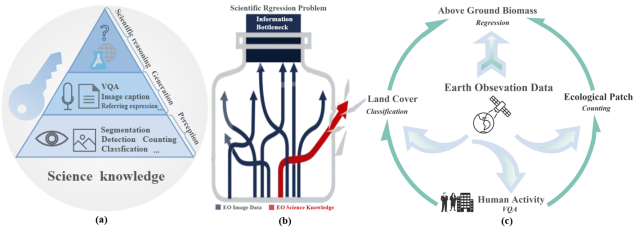

Figure 1: **(a). Hierarchical structure of VLM capabilities**: From basic perception tasks to higher-order reasoning tasks; **(b). Advantages of VLM for EO regression tasks**: By integrating scientific domain knowledge with EO image data, VLMs overcome the information bottleneck of traditional image-only regression models, enabling deeper insights and improved scientific reasoning; **(c). Interplay between regression and generation tasks**: Using AGB estimation as an example, the intrinsic link between regression and generation targets allows collaborative processing in a unified framework, enhancing prediction accuracy and reliability.

To bridge the gap, this work explores the potential of VLMs to jointly tackle both visual description and scientific regression tasks in EO. Compared with traditional methods based on visual contents, VLMs offer unique advantages for solving regression problem, as illustrated in Figure 1(b), By integrating multimodal EO data and embedding domain knowledge in language, the VLMs overcome information bottlenecks and conducts regression using more comprehensive information. Despite the potential of jointly modeling regression and generation tasks with VLMs, several key technical challenges are required to be considered:

- **Lack of instruction-tuning data and evaluation benchmarks:** Existing EO VLM dataset typically emphasizes visual content description, resulting in insufficient fine-

*Corresponding Author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1: Comparison of recent methods across different tasks. LM represents large language model, SM represents small model, such as U-Net, ViT. Text means using language input or not. M-task and Cls denote multi-task and classification, respectively.

Method	Venue	LM/SM	Text	M-task	Cls	VQA	Scientific Regression
Open-Canopy (Fogel et al. 2024)	CVPR/25	SM	✓	×	×	×	Canopy height
Contextformer (Benson et al. 2024)	CVPR/24	SM	×	×	×	×	Geospatial vegetation
TorchSpatial (Wu et al. 2024)	NeurIPS/24	SM	×	✓	✓	×	Landscape variables, Urbanization metrics, SDGs indicators
SRMS (Hamilton et al. 2024)	NeurIPS/24	LM	✓	×	×	×	Species range
LHRS-bot (Muhtar et al. 2024)	ECCV/24	LM	✓	✓	✓	✓	×
GeoChat (Kuckreja et al. 2024)	CVPR/24	LM	✓	✓	✓	✓	×

tuning data and standardized benchmarks for scientific regression tasks. This limits the models' ability to learn accurate numeric mappings between EO inputs and scientific variables.

- **Divergent latent feature requirements:** Generation tasks always focus on describing visual content using features that align with human perception. Scientific regression tasks predict physical properties from raw EO data. They depend on subtle spectral or spatial patterns that may not be visible to the human eye. This difference in feature needs makes it challenging for a single VLM to handle both tasks effectively.
- **Error accumulation in multi-step number generation:** Numerical values are split into multiple discrete tokens during tokenization. Under the autoregressive next-token prediction mechanism, an error in predicting any one token can cascade through subsequent tokens, degrading the accuracy of the final numeric output.
- **Conflicting optimization objectives:** VLMs are trained to predict the next token based on semantic coherence rather than minimizing numeric loss. This mismatch means that optimizing for fluent language generation may undermine the precision needed for accurate numeric regression, making joint fine-tuning challenging.

To address these challenges, this paper introduces REO-Instruct, a novel benchmark dataset in the EO domain that uniquely enables the exploration of VLMs' generation and scientific regression capabilities. Specifically, we selected Above Ground Biomass (AGB) estimation and ecological patch counting as representative scientific regression tasks, while VQA and classification as the representative generation task. Around these tasks, the dataset provides extensive text annotations and domain-specific knowledge, including descriptions of land cover types, human activities, and ecological contexts. As illustrated in Figure 1(c), REO-Instruct highlights the intrinsic relationships between regression targets (e.g., AGB values) and generation objectives (e.g., land cover classes).

In summary, our contributions include:

- **Conceptual framework for scientific VLMs in EO:** We introduce a new perspective that integrates content understanding and quantitative regression within Vision–Language Models, guided by a logical-chain design that links semantically and scientifically related tasks. This framework enables coherent multimodal learning and bridges descriptive and numeric reasoning in Earth Observation.
- **Creation of REO-Instruct:** We construct **REO-Instruct**, a large-scale multimodal benchmark that unifies generation and scientific regression tasks in EO. The dataset integrates co-registered Sentinel-2 and ALOS-2 imagery with structured text annotations, forming a cognitively interpretable chain from human activity and land cover to ecological patch counts and above-ground biomass (AGB).
- **Comprehensive evaluation and insights:** We benchmark representative VLMs on REO-Instruct with standardized protocols, revealing their current limitations in scientific numeric reasoning and establishing transparent baselines for future development of scientifically grounded VLMs in remote sensing.

Related Work

Vision Language Datasets for EO

Training VLMs for EO requires specialized image-text datasets, which are typically developed in two main ways: creating datasets from scratch and expanding existing EO datasets.

Creating datasets from scratch involves sourcing raw EO data and pairing it with annotations. For example, RS-GPT (Hu et al. 2023) created 2,500 high-quality RSI text pairs with expert manual captioning. GRAFT (Mall et al. 2023) linked ground-level image captions from social media with RSIs using geographical tags, obtaining a large dataset without manual captioning. SkyScript (Wang et al. 2024b) and LHRS-Align (Muhtar et al. 2024) utilized OpenStreetMap to generate captions, with LHRS-Align further leveraging the language-only LLM Vicuna-v1.5 for caption production based on geographic features.

Expanding existing EO datasets involves transforming existing annotations into textual descriptions. For instance, RemoteCLIP (Liu et al. 2024a) converted object detection annotations into image captions using textual templates,

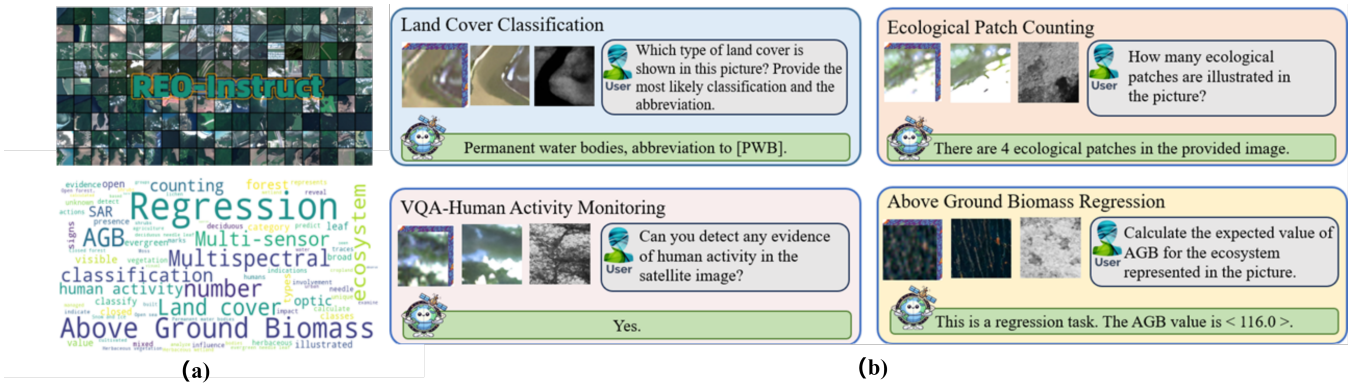

Figure 2: (a). RGB modality examples and word cloud of REO-Instruct benchmark; (b). Screenshots of some image-texts annotation pairs in REO-Instruct benchmark.

significantly increasing the available training data. Similarly, EarthGPT (Zhang et al. 2024a) adopted a similar template-based approach. The RS5M (Zhang et al. 2024b) dataset, currently the largest with 5 million RSIs, employed BLIP2 to generate captions, selecting the best variants using CLIP. GeoChat (Kuckreja et al. 2024) constructed question-answer pairs by describing target characteristics, which were subsequently processed by a LLM. Additionally, SkyEyeGPT (Zhan, Xiong, and Yuan 2025) combined object detection and VQA datasets to create a multitask dialogue instruction dataset.

These datasets set strong benchmarks for tasks like image captioning, VQA, and visual grounding, focusing mainly on interpreting image content. However, despite laying a solid foundation for future research, they still underutilize the potential of multimodal EO data collected from diverse sensors, particularly in addressing scientific regression challenges.

Scientific Regression Tasks in EO

Scientific regression tasks play a key role in EO, enabling essential scientific applications across climatology, ecology, and geophysics. Recent advances have driven significant progress in EO-related regression tasks (Benson et al. 2024; Hamilton et al. 2024; Wu et al. 2024). Common tasks include species range estimation, geospatial vegetation forecasting, population density regression, forest cover prediction, nightlights intensity estimation, elevation mapping, and notably, above ground biomass (AGB) estimation (Lang et al. 2023; Sialelli et al. 2024). Several advanced methods have emerged addressing these tasks with multimodal data, as summarized in Table 1. Classic regression models in EO typically rely on satellite imagery (Li et al. 2020; Laurin et al. 2018; Rodda et al. 2024), complemented by ground-truth data for model training and validation. However, relying solely on EO imagery for scientific regression faces inherent information bottlenecks, as some key environmental and anthropogenic factors may be underrepresented. Incorporating domain-specific knowledge or intermediate interpretation results as textual inputs or embeddings could unlock new possibilities for deeper understanding and reliable prediction.

REO-Instruct Benchmark

Developing a unified EO-VLM for both generation and scientific regression tasks is crucial yet currently hindered by lacking benchmarks. To bridge this gap, we propose REO-Instruct, a large-scale multimodal benchmark integrating EO imagery with domain-rich text annotations, supporting comprehensive model assessment.

Logical Chain Construction and Task Selection

Given the diverse range of EO-based generation and scientific regression tasks, constructing a logical chain within the benchmark requires careful consideration guided by the following principles: 1) **The output of the generation and regression tasks should be derivable from the input EO data**, improving the reliability and interpretability of the results; 2) **Generation and regression tasks must share tightly-coupled logical relationships**. It is unreasonable and unnecessary to use a unified framework to handle unrelated tasks. 3) **Text annotations must be clear, concise, and professionally structured**. Recent studies (Tang, Yang, and Song 2024; Lewkowycz et al. 2022; Song et al. 2024; Vacareanu et al. 2024) demonstrate that while perception tasks are relatively insensitive to minor semantic variations, slight textual perturbations can dramatically affect regression predictions in LLMs. Therefore, clear annotations are critical to mitigating unintended variability.

Based on these guidelines, we specifically focus on forest environmental monitoring, establishing a concise logical chain that incorporates critical and necessary factors, including human activities (anthropogenic impacts), ecological patch counts (reflecting biodiversity), land cover classification and biomass measurements.

EO Data Collection Principles and Overview

To ensure the dataset is comprehensive and representative, we follow these three main principles: 1) EO image modalities must be sufficient and necessary for precise forecasting into surface and vegetation characteristics; 2) The element of EO image data must be balanced, diverse, and representative. It should cover various land cover types, AGB values, human in-

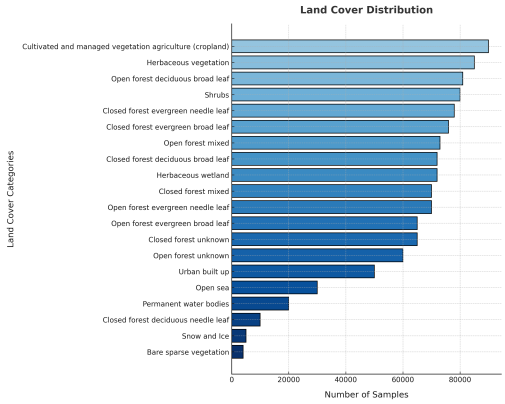

Figure 3: Land cover distribution based on the number of samples in each category.

fluences, and geographic distributions to build generalizable models.

REO-Instruct benchmark leverages the AGBD dataset (Sialelli et al. 2024), which encompasses imagery collected during the years 2019-2020. Inspired by prior work (Li et al. 2020; Laurin et al. 2018; Rodda et al. 2024), the proposed benchmark includes three types of EO data: multispectral (MS) images, RGB three-channel optical images, and Synthetic Aperture Radar (SAR) images. The multispectral data comes from Sentinel-2 L2A, covering 13 spectral bands at a 10-meter spatial resolution. We extracted bands [4,3,2] from each multispectral image to create corresponding RGB image. The 25-meter resolution SAR data originates from ALOS-2 PALSAR-2 products, featuring both HH and HV polarization back scatter. More detailed information can refer to (Shimada et al. 2014). Images from different modalities have been spatially aligned, resulting in 25×25-pixel patches corresponding to a 250m×250m observation area.

These EO images captured by different sensors are further enriched with domain-specific text annotations, leveraging land cover data and the GEDI AGB data to establish text annotation. REO-Instruct benchmark comprises a significant volume of image-text pairs, with 1.6 million pairs in training set, approximately 20k pairs in validation set and 36K pairs in testing set. Figure 2 shows image examples and focus areas of the proposed REO-Instruct benchmark.

Text Annotations in REO-Instruct

Text annotations must incorporate scientific and domain-specific knowledge relevant to generation and scientific regression tasks. To generate the textual component of our dataset, we utilize ChatGPT-4o[1], guided by a carefully designed prompt that ensures the inclusion of pertinent domain knowledge. Following the construction principle, the text annotations in the REO-Instruct benchmark cover several key aspects:

[1]This work invoked ChatGPT-4o, also known as GPT-4o, using OpenAI's official API.

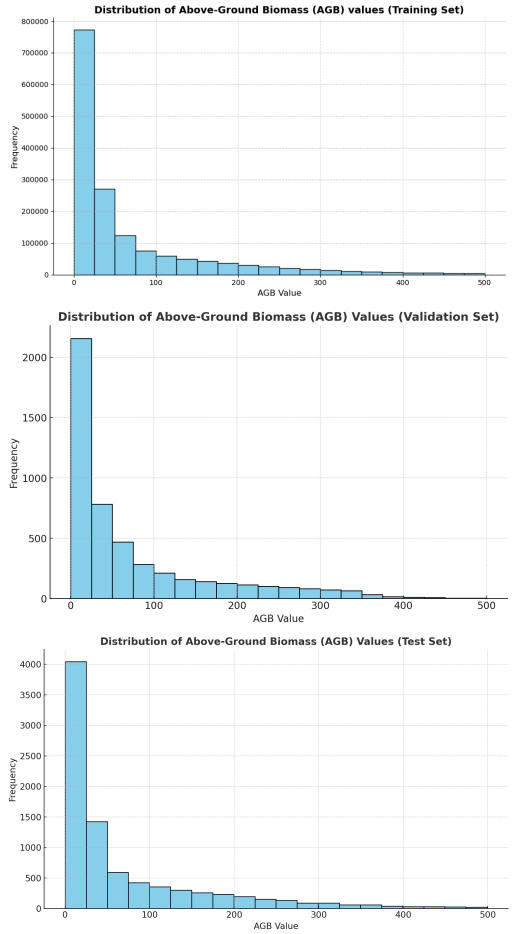

Figure 4: Distribution of Above-Ground Biomass (AGB) values in REO-Instruct. The histogram shows the frequency of AGB values.

- **Land Cover Classification:** Text annotations provide detailed descriptions of land cover types based on Copernicus Global Land Cover Layers. Each image is assigned to one of the professional land cover classes, such as *Closed forest, evergreen needleleaf forest*. These labels enable models to learn vegetation structures, land cover compositions, and spatial patterns critical for ecological and environmental analysis. A complete list of more than 20 land cover categories and their distributions in REO-Instruct can be found in Figure 3.
- **Ecological Patch Counting:** The annotations provide estimates of the number of ecological patches within each observed area. An ecological patch is defined as a continuous land cover unit with distinct ecological characteristics, such as vegetation type, land use, or habitat features. This annotation reflects vegetation richness and fragmentation, where a higher number of patches indicates greater biodiversity and land cover complexity, offering critical ecological insights.
- **VQA-Human Activity Monitoring:** This part of annotations include questions and answers about human-made

Table 2: Results (%) on land-cover classification.

Method	Modality	OA↑	MA-Pre↑	MA-Recl↑	MA-F1↑
Qwen2-VL*	RGB	3.77	0.38	1.14	0.57
ChatGPT-4o	RGB	3.97	12.27	3.63	5.60

*: metrics exclude unanswerable queries. Qwen2-VL deems 96.06% unanswerable.

Table 3: Results on ecological patch counting task.

Method	Modality	RMSE↓	MAE↓	R-squared↑	OA (%)↑
Qwen2-VL*	RGB	12.74	4.30	-121.79	9.75
ChatGPT-4o	RGB	5.13	4.79	-18.42	2.43
LLaVA*	RGB	1.31	1.06	-0.27	25.34

*: metrics exclude unanswerable queries. LLaVA answers 79.31% of RGB and 21.20% of MS queries, while Qwen2-VL only answers 20.31% questions.

features like urban structures, agricultural fields. These annotations discuss the potential impact of human activities on natural landscapes, facilitating the study of human-environment interactions such as urban expansion and deforestation-driven land-use change.

- **Above Ground Biomass Regression:** Text annotations further provide ground-truth quantitative estimates of Above Ground Biomass (AGB), reported in megagrams per hectare (Mg/ha). These values allow models to connect textual descriptions with visual inputs, enabling direct supervision for biomass regression tasks and supporting advanced multimodal learning frameworks. Details about AGB value distribution in REO-Instruct are shown in Figure 4.

Prompt Design and Manual Correction

To guide ChatGPT-4o in generating the most relevant answers while ensuring diversity in the responses, we have carefully designed a large number of prompts. For instance, to ensure the accuracy of land cover category descriptions, we restrict the responses to be selected from a predefined set of available categories. To ensure the diversity of questions and answers, we have set up more than 100 templates. Each conversation is generated by randomly selecting a template, while also introducing new variations to create unique responses. Additionally, we categorize the questions explicitly to ensure that the regression head is activated solely for regression-related tasks. Meanwhile, the generation head does not produce irrelevant content, which could interfere with the regression head's decision-making.

Because automatic generation can introduce errors or drift, we added a manual correction step to keep our annotation accurate and reliable. First, every model-assisted QA pair goes through an automated check. A script compares each annotation against trusted sources, like the *Copernicus Global Land Cover Map*, and fixes any mismatches. For example, if the land cover map marks an area as "urban" but ChatGPT-4o's label doesn't mention human activity, we correcte the model-generated labels to match. Any entries that remain unclear

Table 4: Results on Human Activity Monitoring task.

Model	**Modality**	**Accuracy (%)** ↑
Owen2-VL	RGB	21.52
ChatGPT-4o	RGB	33.79
LLaVA	RGB	44.92
GeoChat	RGB	45.03
LHRS-Bot	RGB	47.87
LLaVA	MS	48.08

Table 5: Results on AGB regression task.

Method	Modality	RMSE↓	MAE↓	R^2 ↑
LLaVA†	RGB	116.60	67.90	-0.51
U-Net	MS	81.27	49.19	0.32
LLaVA*†	MS	115.74	67.49	-0.45

: Unanswerable cases excluded; †: Models fine-tuned on REO-Instruct. LLaVA*† only counted 89.31% of questions with definite answers.

are removed. Next, senior experts review the cleaned test and validation sets by hand. They read through each annotation to confirm facts and ensure terminology is consistent with domain standards. This two-step process automatic filtering followed by expert review helps us deliver a benchmark that's both scientifically sound and easy to trust.

Experiments

We present results on the REO-Instruct benchmark to evaluate the performance of representative VLMs. Each downstream test subset contains approximately 8.6K unique samples, ensuring no image overlap.

For comparison, we include both domain-specific and general-purpose VLMs. Specifically, we evaluate two EO focused large models, GeoChat (Kuckreja et al. 2024) and LHRS-Bot (Muhtar et al. 2024), alongside general-purpose VLMs including LLaVA-1.5-7B (Liu et al. 2024b), Qwen2-VL-7B (Wang et al. 2024a), and ChatGPT4o (Achiam et al. 2023). In addition, we consider the results obtained when using different modalities of EO imagery as inputs to REO-VLM. This initial exploration aims to assess how varying input modalities influence inference accuracy. For GeoChat, we modified its *clip_interpolate_embeddings* component to adapt to the image resolution of REO-Instruct. Unless otherwise specified, for the comparison models, we use the official 7B versions. Following common evaluation practices in EO-VLM research (Li et al. 2024; Kuckreja et al. 2024; Muhtar et al. 2024), we directly evaluate these publicly released models without any additional training on our dataset. To ensure fairness during evaluation, we further provided guiding prompts that explained the questions and offered a range of possible answers to other compared methods.

Guiding Prompts

To define the scope of our questions and expected answers, we provide guiding prompts during testing for all comparison algorithms, except for our proposed method. These prompts

Prompt: This is a detailed satellite image showcasing a diverse land cover. Your task is to analyze the features visible in this image and classify the land cover based on the types of vegetation and landforms present. Consider the textures, colors, and arrangement of features such as trees, water bodies, and built-up areas. Choose the most probable one class from the following options: 'No input data available', 'Closed forest, evergreen needle leaf', 'Closed forest, deciduous needle leaf', 'Closed forest, evergreen broad leaf', 'Closed forest, deciduous broad leaf', 'Closed forest, mixed', 'Closed forest, unknown', 'Open forest, evergreen needle leaf', 'Open forest, deciduous needle leaf', 'Open forest, evergreen broad leaf', 'Open forest, deciduous broad leaf', 'Open forest, mixed', 'Open forest, unknown', 'Shrubs', 'Herbaceous vegetation', 'Herbaceous wetland', 'Moss and lichen', 'Bare / sparse vegetation', 'Cultivated and managed vegetation/agriculture (cropland)','Urban / built up', 'Snow and Ice', 'Permanent water bodies', 'Open sea', 'Missing value'. If you cannot predict, please answer <NA>. If you can predict, please give the predicted category and mark the category with <>.

Prompt: This is a satellite earth observation picture. Could you estimate the above-ground biomass for the ecosystem shown in the image? Please answer with a number.

Prompt: This is a satellite image. Please judge whether there are any signs of human activities in the image?

Prompt: This is a satellite image. How many ecological patches are contained in the ecosystem of the area shown in the image?

Figure 5: Test guiding prompts for compared methods.

shown in Figure 5 assist VLMs in better understanding and addressing multiple tasks.

Experimental results

We present experimental results across four representative tasks: land cover classification (Table 2), ecological patch counting (Table 3), VQA-based human activity monitoring (Table 4), and above-ground biomass (AGB) regression (Table 5) on the REO-Instruct benchmark. Building on these quantitative results, we summarize three key observations.

- In the human-activity monitoring task, MS inputs consistently outperform RGB imagery, and EO-specific VLMs (GeoChat, LHRS-Bot) achieve higher accuracy than general-purpose models, underscoring the benefit of spectral cues and domain-specialized pre-training.
- For forest ecology–oriented questions, general-purpose open-source and closed-source VLMs exhibit very low answer accuracy, revealing a substantial gap in forestry and ecological knowledge that cannot be bridged by generic web-scale corpora alone.
- In the regression tasks, all compared methods yield negative R^2 values under both RGB and MS settings, indicating that they fail to learn meaningful numerical relationships from EO inputs and remain far from reliable for EO-driven quantitative estimation.

Conclusion

This work benchmarks VLMs in EO for forest ecological analysis through the introduction of **REO-Instruct**, a multimodal benchmark designed to align descriptive and quantitative tasks within a unified framework. Experiments across representative tasks demonstrate that while current VLMs excel in content understanding, they still face significant challenges in numerical reasoning and scientific regression. These findings highlight the necessity of regression-aware modeling strategies and multimodal data integration for advancing scientific applications of VLMs. Overall, this work establishes a standardized foundation for benchmarking and developing next-generation **scientific VLMs in forest ecological analysis**, with future efforts directed toward extending this framework to broader geoscientific domains.

Acknowledgments

The work of Xizhe Xue is funded by the Technical University of Munich Global Postdoc Fellowship. This work is also supported by Munich Center for Machine Learning.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Allen, M. J.; Dorr, F.; Mejia, J. A. G.; Martínez-Ferrer, L.; Jungbluth, A.; Kalaitzis, F.; and Ramos-Pollán, R. 2025. M3LEO: A Multi-Modal, Multi-Label Earth Observation Dataset Integrating Interferometric SAR and Multispectral Data. *NeurIPS*, 37: 104694–104723.

Batista e Silva, F.; Freire, S.; Schiavina, M.; Rosina, K.; Marín-Herrera, M. A.; Ziemba, L.; Craglia, M.; Koomen, E.; and Lavalle, C. 2020. Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature communications*, 11(1): 4631.

Benson, V.; Robin, C.; Requena-Mesa, C.; Alonso, L.; Carvalhais, N.; Cortés, J.; Gao, Z.; Linscheid, N.; Weynants, M.; and Reichstein, M. 2024. Multi-modal learning for geospatial vegetation forecasting. *CVPR*, 27788–27799.

Fogel, F.; Perron, Y.; Besic, N.; Saint-André, L.; Pellissier-Tanon, A.; Schwartz, M.; Boudras, T.; Fayad, I.; d'Aspremont, A.; Landrieu, L.; et al. 2024. Open-Canopy: A country-scale benchmark for canopy height estimation at very high resolution. *arXiv preprint arXiv:2407.09392*.

Hamilton, M.; Lange, C.; Cole, E.; Shepard, A.; Heinrich, S.; Mac Aodha, O.; Van Horn, G.; and Maji, S. 2024. Combining Observational Data and Language for Species Range Estimation. *NeurIPS*.

Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; and Li, X. 2023. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*.

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. *CVPR*, 27831–27840.

Lang, N.; Jetz, W.; Schindler, K.; and Wegner, J. D. 2023. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, 7(11): 1778–1789.

Laurin, G. V.; Balling, J.; Corona, P.; Mattioli, W.; Papale, D.; Puletti, N.; Rizzo, M.; Truckenbrodt, J.; and Urban, M. 2018. Above-ground biomass prediction by Sentinel-1 multitemporal data in central Italy with integration of ALOS2 and Sentinel-2 data. *Journal of Applied Remote Sensing*, 12(1): 016008–016008.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving quantitative reasoning problems with language models. *NeurIPS*, 35: 3843–3857.

Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; and Zhu, X. X. 2024. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*.

Li, Y.; Li, M.; Li, C.; and Liu, Z. 2020. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific reports*, 10(1): 9952.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024a. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *NeurIPS*, 36.

Mall, U.; Phoo, C. P.; Liu, M. K.; Vondrick, C.; Hariharan, B.; and Bala, K. 2023. Remote sensing vision-language foundation models without annotations via ground remote alignment. *ICLR*.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *ECCV*, 440–457.

Nguyen, T.; Jewik, J.; Bansal, H.; Sharma, P.; and Grover, A. 2024. Climatelearn: Benchmarking machine learning for weather and climate modeling. *NeurIPS*, 36.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. 8748–8763. PMLR.

Rodda, S. R.; Fararoda, R.; Gopalakrishnan, R.; Jha, N.; Réjou-Méchain, M.; Couteron, P.; Barbier, N.; Alfonso, A.; Bako, O.; Bassama, P.; et al. 2024. LiDAR-based reference aboveground biomass maps for tropical forests of South Asia and Central Africa. *Scientific Data*, 11(1): 334.

Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; and Lucas, R. 2014. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sensing of environment*, 155: 13–31.

Sialelli, G.; Peters, T.; Wegner, J. D.; and Schindler, K. 2024. AGBD: A Global-scale Biomass Dataset. *arXiv preprint arXiv:2406.04928*.

Song, X.; Li, O.; Lee, C.; Yang, B.; Peng, D.; Perel, S.; and Chen, Y. 2024. Omnipred: Language models as universal regressors. *arXiv preprint arXiv:2402.14547*.

Tang, E.; Yang, B.; and Song, X. 2024. Understanding LLM Embeddings for Regression. *arXiv preprint arXiv:2411.14708*.

Vacareanu, R.; Negru, V.-A.; Suciu, V.; and Surdeanu, M. 2024. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. *arXiv preprint arXiv:2404.07544*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; and Rajagopal, R. 2024b. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *AAAI*, 38(6): 5805–5813.

Wu, N.; Cao, Q.; Wang, Z.; Liu, Z.; Qi, Y.; Zhang, J.; Ni, J.; Yao, X.; Ma, H.; Mu, L.; et al. 2024. Torchspatial: A location encoding framework and benchmark for spatial representation learning. *NeurIPS*.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2025. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 64–77.

Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; and Mao, X. 2024a. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024b. 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*.