ACADEMIC POSITION · POST-DOCTORAL POSITION

Multi-Modal Vision–Language Model Architecture for Earth Observation

ANR GEO-ReSeT Project (geo-reset.sylvainlobry.com)

Vision-Language Models | Multimodal Learning | Remote Sensing | Foundation Models | Computer Vision

POSITION DETAILS

Duration	12 months (Start as soon as possible)
Indicative salary	2,585–3,200 € gross/month (depending on experience)
Institute	Université Paris Cité — LIPADE, Team SIP (*Systèmes Intelligents de Perception*)
Location	45 rue des Saints-Pères, 75006 Paris, France
Supervision	Sylvain Lobry
Status	Open until filled

APPLICATION PROCEDURE

Send an email to **sylvain.lobry@u-paris.fr** with subject:

`[Postdoc GEO-ReSeT] FirstName LastName`

Required documents:

- Updated CV
- Cover letter
- Contact information of a referee / supervisor

01 Introduction

Earth Observation (EO) produces massive amount of data from diverse sensors, including Very High Resolution (VHR) optical imagery, multi-spectral time series (e.g., Sentinel-2), Synthetic Aperture Radar (SAR, e.g., Sentinel-1). Combining these modalities is key to monitoring environmental evolution, urban growth, and natural disasters [1]. To make such information accessible beyond remote sensing specialists, there is a growing need for Vision Language Models (VLMs) capable of interpreting complex scenes and answering natural-language queries [2, 3].

However, most existing EO VLMs [4, 5] pair standard RGB vision backbones (often pretrained on natural images) with a Large Language Model (LLM). Because of this design, they cannot process arbitrary spectral bands, non-optical modalities such as SAR, varying spatial resolutions, and multi-temporal dynamics, while remaining dependent on limited manually annotated datasets.

Within the ANR-funded **GEO-ReSeT** project (geo-reset.sylvainlobry.com), our group has developed building blocks that address these limitations individually: **Atomizer** [6] and **RAMEN** [7] enable sensor-agnostic, resolution-flexible encoding of heterogeneous EO observations, while **GroundSet** [8] generates spatially grounded supervision from cadastral vector records without manual labeling. The objective of this postdoctoral position is to bring these efforts together to propose, train, and validate an architecture for a multi-modal EO VLM (Figure 1).

The work plan is flexible and will be defined jointly with the candidate based on their background and interests. The position is funded for 12 months in the **SIP team** at **LIPADE** (Université Paris Cité), with computing resources available locally and via the Jean Zay supercomputer, targeting publications at leading computer vision conferences (e.g., CVPR, ICCV) or remote sensing journals (e.g., IEEE TGRS, ISPRS PHOTO). **The position is open immediately and expected to start as soon as possible.**

02 Background

Existing EO VLMs such as **GeoChat** [4] and **EarthGPT** [5] have demonstrated strong progress on remote sensing visual question answering, captioning, and region-level grounding. However, they typically rely on fixed-resolution RGB vision backbones or modality-specific adapters, limiting their ability to handle arbitrary spectral bands, non-optical sensors, or multi-temporal series natively.

Recent works in our group address two core prerequisites for a general-purpose multi-modal EO VLM: flexible multi-sensor representation learning and scalable spatial grounding. First, **Atomizer** [6] breaks satellite images down into sets of metadata-conditioned scalar tokens (encoding band wavelength, spatial resolution, and acquisition time), allowing a single model to process arbitrary sensor configurations without resampling. Second, **RAMEN** [7] introduces a resolution-adjustable multi-modal encoder in a unified latent space, supporting optical, multi-spectral, SAR, and elevation inputs while letting users control output spatial precision at inference time. Third, **GroundSet** [8] leverages cadastral vector databases to automatically create large-scale spatially grounded supervision for object localization, counting, and spatial reasoning.

03 Research Directions (Flexible Work Plan)

- **Sensor-agnostic visual tokenization for LLMs:** Designing an effective interface (e.g., cross-attention resamplers, resolution-conditioned projectors, or hybrid tokenizers) connecting encoders like **RAMEN** [7] and **Atomizer** [6] to an LLM without losing

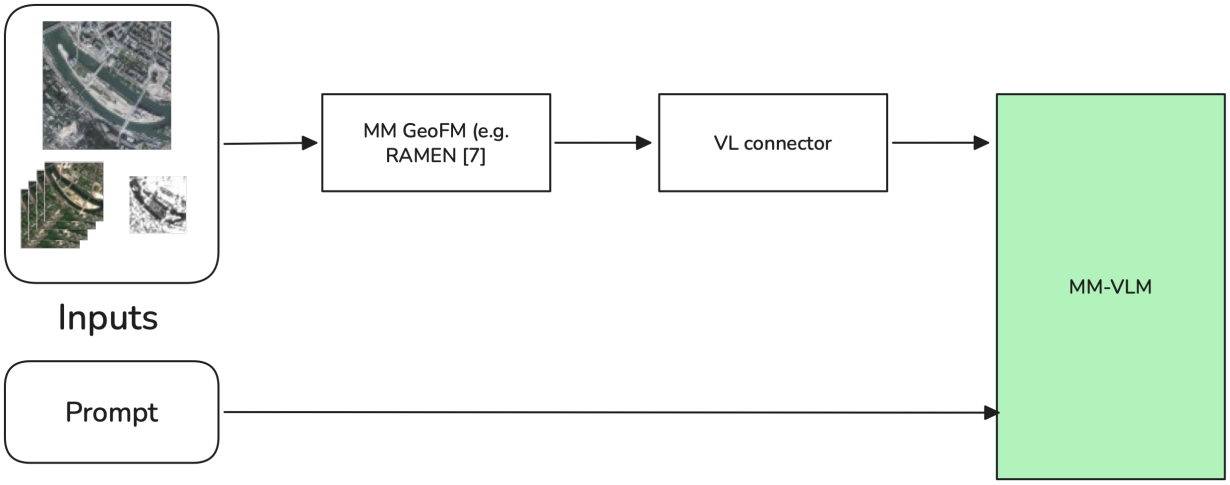

Figure 1: Conceptual direction for the GEO-ReSeT multi-modal EO VLM.

fine spatial, spectral, or temporal cues. This includes investigating how to compress variable-length multi-sensor token sets into LLM-compatible representations and aligning them with language using scalable supervision such as **GroundSet** [8] and existing EO VQA corpora [2].

- **Evaluation of the model:** Designing and running comprehensive evaluation protocols for the resulting multi-modal EO VLM across diverse sensors and resolutions, including spatial grounding, temporal reasoning, and robustness to missing modalities, alongside open-source model and code release within GEO-ReSeT.

04 Desired Background for the Candidate

We are looking for a recently graduated PhD in Computer Science, Machine Learning, or a related field with a strong background in computer vision and deep learning, and proficiency in Python (PyTorch). Prior experience with Vision–Language Models, LLMs, multimodal representation learning, or remote sensing data is a strong advantage.

About LIPADE, the SIP Team, and the GEO-ReSeT Project

The postdoctoral researcher will join the **SIP (*Systèmes Intelligents de Perception*)** team at **LIPADE** (*Laboratoire d'Informatique de Paris Descartes*), Université Paris Cité (45 rue des Saints-Pères, 75006 Paris), and work closely with Sylvain Lobry, Camille Kurtz, Laurent Wendling (UPCité) and Diego Marcos (INRIA). He will also collaborate with the PhD students and collaborators behind Atomizer, RAMEN within the ANR **GEO-ReSeT**.

References

[1] M. Dalla Mura et al. "Challenges and opportunities of multimodality and data fusion in remote sensing". In: *Proceedings of the IEEE* 103.9 (2015), pp. 1585–1601.

[2] S. Lobry, D. Marcos, J. Murray, and D. Tuia. "RSVQA: Visual question answering for remote sensing data". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.12 (2020), pp. 8555–8566.

[3] D. Tuia et al. "Toward a collective agenda on AI for Earth science data analysis". In: *IEEE Geoscience and Remote Sensing Magazine* 9.2 (2021), pp. 88–104.

[4] K. Kuckreja et al. "GeoChat: Grounded Large Vision-Language Model for Remote Sensing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 27831–27840.

[5] W. Zhang et al. "EarthGPT: A Universal Multi-modal Large Language Model for Multi-sensor Image Comprehension in Remote Sensing Domain". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–20.

[6] H. Riffaud de Turckheim, S. Lobry, R. Interdonato, and D. Marcos. "Atomizer: Generalizing to new modalities by breaking satellite images down to a set of scalars". In: *British Machine Vision Conference (BMVC)*. 2025.

[7] N. Houdré et al. "RAMEN: Resolution-Adjustable Multimodal Encoder for Earth Observation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2026.

[8] R. Ferrod et al. "GroundSet: A Cadastral-Grounded Dataset for Spatial Understanding with Vector Data". In: *arXiv preprint arXiv:2603.14609* (2026).