Multimodal Learning for Earth Observation: Automating Satellite Image Captioning with Geo-FMs

Luca Chiarabini[1], Antony Zappacosta[1], Daniela Espinoza Molina[1], Rıdvan Salih Kuzu[1], Andrés Camero[1]

[1] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Remote Sensing Institute (IMF), Data Science Department (DAS) - Weßling

Abstract

The automatic generation of captions for satellite images can enhance the accessibility and interpretability of Earth Observation (EO) data. In this study, we compare two approaches to image captioning: TerraMind, a model developed within the FAST-EO project specifically for satellite imagery, and BLIP-2, a generic multimodal model trained on RGB images. The dataset used, SmallMinesDS, consists of annotated satellite images from five districts in Ghana, where unregulated small-scale gold mining threatens cocoa farmlands. Our evaluation focuses on caption accuracy, specificity, and adaptability to EO imagery, highlighting the strengths and limitations of each approach in the context of environmental monitoring.

Dataset

SmallMinesDS[1] consists of a collection of annotated satellite images covering five districts in Ghana. Each image is labeled to indicate the presence or absence of mining activity.

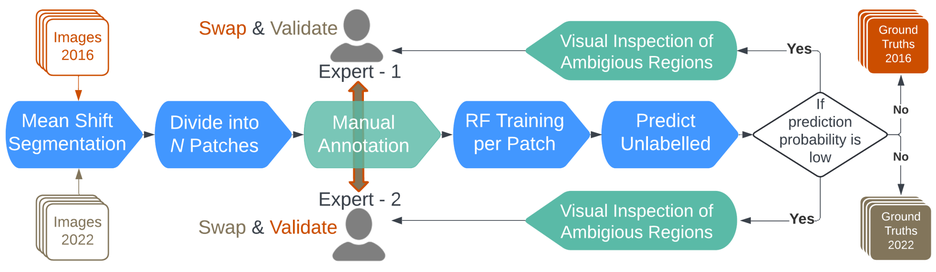

Figure 1. SmallMinesDS creation workflow.

We co-registered Sentinel-1 (S1), Sentinel-2 (S2), and Copernicus Digital Elevation Model (COPDEM) data, and annotated mining areas as illustrated in the **Figure 1**. A Random Forest (RF) classifier was applied to propagate the labels based on initial manual annotations, followed by a final round of manual review..

BLIP-2

BLIP-2[2] is a multimodal model that connects frozen image encoders and large language models via a lightweight Querying Transformer (Q-Former). Trained in two stages, it achieves state-of-the-art performance on tasks like image captioning and visual question answering, outperforming larger models in zero-shot settings with fewer trainable parameters.

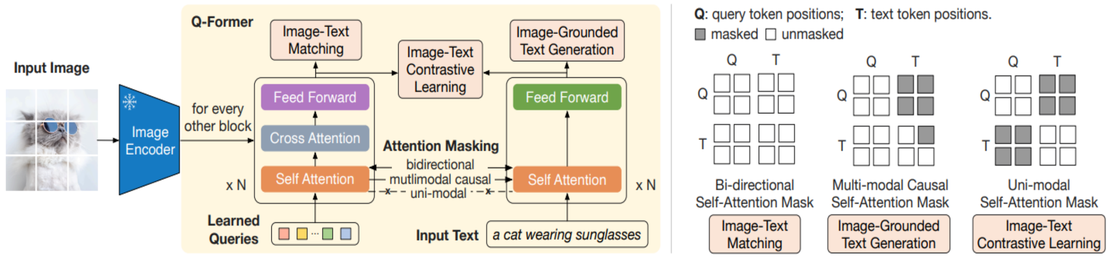

Figure 2. Blip2 network

Terramind: A Multi-modal Geo-Foundation Model

Terramind is a multi-modal geo-foundational model for Earth Observation, currently under development within the FAST-EO[3] project. It integrates multi-sensor, multi-temporal, and multi-modal inputs using efficient encoder-decoder architectures, ensuring interoperability across sources like Sentinel-1, Sentinel-2, and EnMAP. The model combines contrastive language-image pretraining (CLIP-based) with a LLaMA-based dual encoder to improve alignment between satellite imagery and textual descriptions.

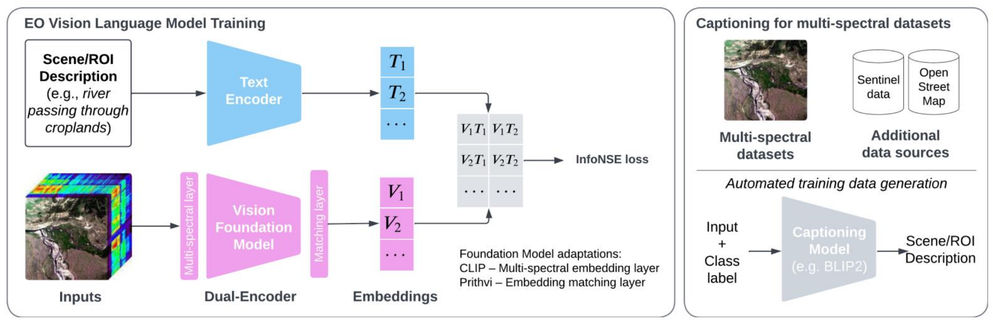

Figure 3. Terramind high level architecture.

Results

We generated captions for approximately 1,300 satellite images from the SmallMinesDS dataset. Each image contains 13 bands: 10 Sentinel-2 L2A bands (at 10m and 20m resolution), VH and VV polarizations from Sentinel-1, and a COPDEM elevation layer. The TerraMind beta model, specifically developed for EO applications, processes all 12 S2 bands, with the missing 60m bands mocked using interpolation. TerraMind is a transformer-based model with approximately 470 million parameters, trained on multimodal EO data. For comparison, we used BLIP-2, a general-purpose vision-language model with 2.7 billion parameters, which was fed only RGB bands. Ground truth captions were generated using the Llava-next[4] model, following the same protocol adopted in TerraMind's training. All input images were provided in GeoTIFF format and preprocessed to ensure consistent resolution and band alignment prior to inference.

Metric	BLIP-2	TerraMind
BLEU	0.0032	0.0127
METEOR	0.0536	0.1134
ROUGE-L	0.1093	0.1577
SBERT Similarity	0.2980	0.4568
CIDEr	0.0024	0.0095

While TerraMind, as a domain-specific model, outperforms BLIP-2 across all metrics, its overall performance remains suboptimal. These results suggest that, on the SmallMines dataset, further fine-tuning is required to enhance the model's captioning capabilities.

Conclusion and Future Work

Our quantitative evaluation shows that the TerraMind (beta) model outperforms BLIP-2 across all automatic captioning metrics, confirming the value of domain-specific approaches for Earth Observation. However, its performance remains insufficient for consistently high-quality captions. These findings highlight the need for fine-tuning on datasets like SmallMinesDS. Future work will focus on refining TerraMind to enhance caption accuracy and assess its role in supporting environmental monitoring.

References

[1] Oforo-Ampofo, Stella; Zappacosta, Antony, et al. (2025). SmallMinesDS: A Multi-Modal Dataset for Mapping Artisanal and Small-Scale Gold Mines, *IEEE Geoscience and Remote Sensing Letters*

[2] Li, J., Li, D., Savarese, S., & Hoi, S.C. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *International Conference on Machine Learning*.

[3] Kuzu, R.S., Brunschwiler, T., Cavallaro, G., et al. (2025). FAST-EO: Multi-Modal Foundation Models for Scalable Earch Observation and Earth Sciences. *ESA-NASA International Workshop on AI Foundation Models for EO, Frascati, Italy*

[4] Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., & Li, C. (2024). LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *ArXiv, abs/2407.07895*.

DLR