Beyond the Visible: Multispectral Vision-Language Learning for Earth Observation

Clive Tinashe Marimo [⋆1], Benedikt Blumenstiel [⋆2] (✉), Maximilian Nitsche[1], Johannes Jakubik[2], and Thomas Brunschwiler[2]

[1] IBM Germany
[2] IBM Research Europe
bendikt.blumenstiel@ibm.com

Abstract. Vision-language models for Earth observation (EO) typically rely on the visual spectrum of data as the only model input, thus failing to leverage the rich spectral information available in the multispectral channels recorded by satellites. Therefore, we introduce Llama3-MS-CLIP—the first vision-language model pre-trained with contrastive learning on a large-scale multispectral dataset and report on the performance gains due to the extended spectral range. Furthermore, we present the largest-to-date image-caption dataset for multispectral data, consisting of one million Sentinel-2 samples and corresponding textual descriptions generated using Llama3-LLaVA-Next and Overture Maps data. We develop a scalable captioning pipeline, which is validated by domain experts. We evaluate Llama3-MS-CLIP on multispectral zero-shot image classification and retrieval using three datasets of varying complexity. Our results demonstrate that Llama3-MS-CLIP significantly outperforms other RGB-based approaches, improving classification accuracy by +6.77% on average and retrieval performance by +4.63% mAP compared to the second-best model. Our results emphasize the relevance of multispectral vision-language learning. The image-caption dataset, code, and model weights are available at https://github.com/IBM/MS-CLIP.

Keywords: Multispectral Data · Vision-Language Model · Earth Observation

1 Introduction

Vision-language models (VLM) have transformed computer vision, enabling powerful zero-shot learning and cross-modal retrieval capabilities [20,9]. By learning joint representations of images and text, these models generalize across tasks without requiring task-specific training data.However, existing VLMs, such as CLIP [20], are predominantly trained on natural RGB images, limiting their applicability to specialized domains such as Earth observation(EO) [28,25,16]. Conversely, effectively utilizing multispectral input data in VLMs represents an interesting and underexplored research topic in the machine learning community.

⋆ Equal contribution.

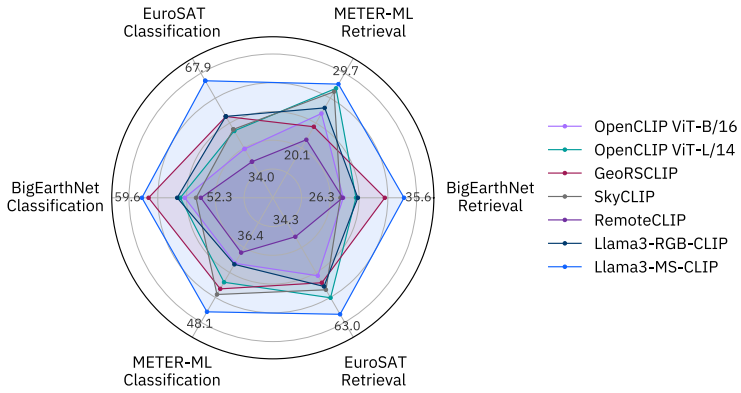


Fig. 1: Zero-shot classification and text-to-image retrieval results, measured in accuracy (%) ↑ and mAP@100 (%) ↑, respectively. We applied a smoothed min-max scaling and annotated the lowest and highest scores. The multispectral CLIP is outperforming other RGB-based models on most benchmarks.

EO relies on satellite imagery to monitor environmental changes, urban expansion, and agriculture [30]. To use VLMs for such EO applications, researchers proposed a range of domain-specific EO adaptations of CLIP [28,25,16,15]. Those models have been trained on up to five million remote sensing images with aligned image captions. Despite impressive performance, these models rely only on RGB input channels instead of leveraging the full spectral range available in multispectral (MS) satellite data for maximum effectiveness. Satellites like Sentinel-2 provide up to 13 spectral bands, capturing rich information far beyond visible wavelengths. Yet, until today, no large-scale multispectral dataset with image captions is publicly available for the research community.

To address this gap, we generate approximately one million captions for a popular, multispectral EO dataset derived from the Sentinel satellites called SSL4EO-S12 [24]. We develop an automated captioning approach using metadata tags from Overture maps and the multimodal large language model (MLLM) Llama3-LLaVA-Next-8B [12]. These captions provide semantic grounding for contrastive learning, allowing the model to align multispectral image representations with natural language. We then perform continual pre-training on OpenCLIP [9], adapting the model to the EO domain. Our model, Llama3-MS-CLIP[3], outperforms other VLMs on a range of downstream applications as depicted in Figure 1. The results demonstrate that multispectral data significantly enhances vision-language learning in EO, unlocking capabilities that RGB-based models fail to capture.

[3] Built with Meta Llama 3. While the model itself is not based on Llama 3 but OpenCLIP B/16, it is trained on captions generated by a Llama 3-derivative model. Therefore, the model name starts with Llama 3 following its license (`https://github.com/meta-llama/llama3/blob/main/LICENSE`).

Our contributions are threefold: (1) We create the largest multispectral image-caption dataset for EO, (2) we present the first multispectral EO VLM, surpassing current state-of-the-art performance, and (3) we propose novel best practices for model development and image captioning with multispectral data. The dataset, code, and model weights are available at `https://github.com/IBM/MS-CLIP` under a permissive license.

2 Related Work

Vision-language models have successfully enabled zero-shot learning and cross-modal retrieval by aligning image and text embeddings through contrastive learning. CLIP [20], OpenCLIP [9], and ALIGN [11] are among the most prominent models in this space, trained on vast datasets of internet-scale image-text pairs. These models excel at general vision tasks but are inherently biased toward natural RGB images. Their ability to generalize to remote sensing imagery is limited due to the domain gap between natural images and satellite imagery [20,28]. Spectral information beyond the visible range is key for applications like vegetation and disaster monitoring, as well as urban planning, which benefit from near-infrared and short-wave infrared reflectance measurements [23].

One of the major challenges in EO vision-language learning is the lack of large-scale image-text datasets with multispectral data [26]. Unlike natural images, satellite images do not inherently come with descriptive text. Most EO datasets provide only categorical labels or metadata, making it challenging to train models that require diverse textual supervision. While some approaches have attempted to generate captions for EO images, they often rely on metadata-based descriptions [25] or manually curated annotations [16]. These approaches do not scale to the large volumes of data required to train robust vision-language models. UCMC [19], RSICD [17], and RSITMD [27] are some famous examples of human-curated EO datasets that are often limited by their small size. RS5M [28] collected five million images from eleven source datasets and used BLIP-2 [13] to generate captions for the RGB images. RSCLIP [15] introduced a pseudo-labeling technique that automatically generates pseudo-labels from unlabeled data. ChatEarthNet [26] is the first multispectral dataset with over 100k samples, using ChatGPT-3.5 to generate captions. However, the captions are solely based on a small set of land-cover classes without visual input for the LLM.

Recent approaches have adapted VLMs to EO by continual pre-training on domain-specific RGB datasets. For instance, SkyCLIP [25], RSCLIP [15], and RemoteCLIP [16] built large-scale image-text datasets and adapted CLIP-based backbones. GeoRSCLIP [28] was pre-trained on RS5M, the largest known image-test dataset in the domain with five million images. GRAFT [18] utilized co-located street-view images to correlate satellite imagery with language. Despite these advancements, all these methods rely solely on RGB data, ignoring the rich spectral information available in multispectral EO imagery.We further note the emergence of autoregressive approaches that generate language output based on optical or radar images, like in TerraMind [10].

In contrast to prior work, we introduce a self-supervised approach based on multimodal large language models (MLLM) and Overture annotations to automatically generate captions for multispectral EO imagery (i.e., Sentinel-2 images). By fine-tuning OpenCLIP on this dataset, we enable multispectral vision-language learning, allowing the model to leverage spectral information beyond the visible spectrum.

3 Automated Captioning

The effectiveness of vision-language models relies heavily on the availability of high-quality image-text datasets. Thus, we introduce Llama3-SSL4EO-S12 captions, a novel dataset of text data aligned with SSL4EO-S12 v1.1 [2]. It provides detailed natural language descriptions required for contrastive learning of multispectral vision-language models.

SSL4EO-S12 v1.1 [2] consists of 975k co-registered images of optical data from Sentinel-2 L1C (top-of-atmosphere) and Sentinel-2 L2A (bottom-of-atmosphere) as well as synthetic aperture radar (SAR) data from Sentinel-1 GRD. The dataset covers 244k global locations centered around urban areas, with a 264×264 pixel size at 10 m resolution, each including samples from four seasons.

We generate captions by employing a multimodal large language model, specifically Llama-LLaVA-Next-8B. The model was selected based on a qualitative comparison and a quantitative evaluation of three MLLM models using METEOR (Metric for Evaluation of Translation with Explicit Ordering) [7], comparing ground truth captions with generations. We tested BLIP2 [13], used for the captioning in the RS5M [28] dataset, Llama3-LLaVA-Next-8B [12], and RS-LLaVA [1], a domain-specific adaptation of LLaVA 1.5. We assessed these models using UCM Captions [19], RSICD [17], and RSITMD [27] that provide human-annotated captions. Llama3-LLaVA-Next-8B reaches an average METEOR score of 0.20 compared to 0.16 for RS-LLaVA and 0.10 for BLIP2. All scores and some examples are provided in the supplementary material.

The captioning process consists of the following steps: First, we extract the RGB channels from S-2 L2A data and scale it to a uint8 value range of 0—255 as no publicly available MLLM supports multispectral inputs. The images are resized to 224×224 pixels as input for the captioning model. We then extract geographical tags from the Overture Maps base layer[4] that provides additional contextual information about land cover, infrastructure, and other features in the satellite image. The geographical instances are then sorted and filtered by size, i.e., all features smaller than 2500 square meters (5×5 pixels) are omitted. We use all tags of each instance as they include additional information, like intermittent rivers. We further add the names of places to avoid hallucinations. Otherwise, we observe that the model often refers to popular places incorrectly, like labeling most universities as *Harvard* or *Berkeley*. Finally, we prompt the MLLM to generate captions in a structured manner by following a

[4] Overture Maps: `https://docs.overturemaps.org` (Version: 2024-03-12-alpha.0)

chain-of-thought approach. First, the model is prompted to generate three relevant question-answer pairs, guiding the model in producing the final caption. The prompt includes further instructions to avoid hallucinations and increase the caption quality. We repeat the generation if any Q&A pair or the caption is missing in the output. We include the prompt and other details on the captioning process in the supplementary material.

Figure 2 shows example images and their corresponding generated captions. While we do observe several hallucinations in the generated captions, they are also more diverse and include more details than other image-caption datasets such as SkyScript or RS5M, which are based on heuristics [25] or the much simpler MLLM BLIP2 [28,14] (see supplementary material for examples). Figure 2 includes three examples with hallucinations to showcase their different forms. For example, the model sometimes imagines landmasses in ocean patches, man-made or water features, and provides wrong counts or length estimations.

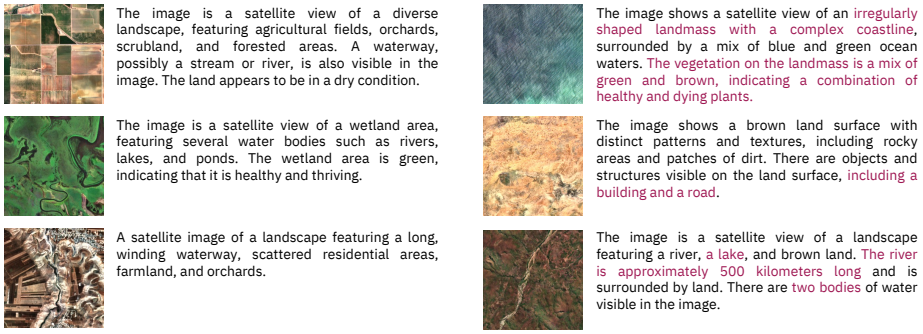


Fig. 2: Examples of image-caption pairs of high quality (left) and hallucination examples (right) of our generated pre-training dataset. We highlight hallucinations in red.

We evaluate the captions quantitatively by comparing our validation set with manually labeled EO datasets: RSITMD [27], RSICD [17], and UCM Captions [19]. The generated captions exhibit a much higher average n-gram diversity of 0.75 compared to only 0.48 to 0.49 in the three human-annotated datasets. The similarity between captions is also lower, showing a higher lexical variety in the SSL4EO-S12-captions.

To assess the quality of our dataset, we asked domain experts[5] to conduct a manual evaluation using a random subset with more than one thousand captions from the validation split. Domain experts rated the captions based on completeness and presence of hallucinations. The caption completeness represents whether all relevant features in the image are covered in the caption and is measured on a scale from 0 (Terrible) to 5 (Excellent). Additionally, the ex-

[5] 14 researchers from the FAST-EO project working in the Earth observation domain.

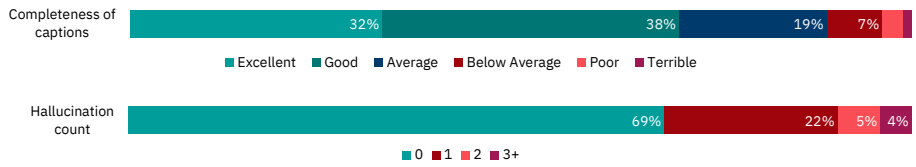


Fig. 3: Evaluation results of the caption quality in the Llama3-SSL4EOS12 dataset based on 1.3k captions reviewed by domain experts. Completeness evaluates if all relevant features of an image are mentioned in the caption, while hallucinations are the number of incorrect features.

perts counted hallucinations. Figure 3 presents the distribution of expert ratings, indicating that over 85% of the captions are considered to include most of the relevant features in the image. The human assessment further demonstrates that two-thirds of the evaluated data is free from hallucinations. If hallucinations are present, we typically observe only one hallucinated feature within an image. We provide details of the quantitative comparison and the human evaluation in the supplementary material.

The manual evaluation of the generated captions demonstrates that automated captioning with a general-purpose MLLM and additionally provided tags is feasible and leads to mostly correct captions. Furthermore, the quantitative assessment of the full validation set indicates that the captioning model uses a more diverse vocabulary than existing human-annotated datasets. Different from datasets like ChatEarthNet [26] or SkyScript [25] that do not use multimodal LMMs, our pipeline can capture scene-specific features like snow, clouds, or colors. While we do want to highlight the challenge of hallucinations in the dataset, our experiments show that VLMs can learn semantic concepts from the correct annotations. Furthermore, the alignment with S-1 GRD data in SSL4EO-S12 and the question-answer pairs provides additional potential for the EO community.

4 Llama3-MS-CLIP

Llama3-MS-CLIP is trained with self-supervised contrastive language-image pre-training (CLIP) [20], visualized in Figure 4. We modified the input layer to handle Sentinel-2's spectral bands beyond RGB by extending the patch embedding for the additional channels. We initialize the corresponding weights with zero tensors so that during continual pre-training, the model starts from RGB input and can iteratively include additional channels based on optimizing the loss landscape. Hence, the model can slowly learn to leverage the additional information. Our initial experiments suggest that this initialization strategy outperforms random initialization, where the continual pre-training would be disrupted due to the noise that originates from the random weights for the additional channels.

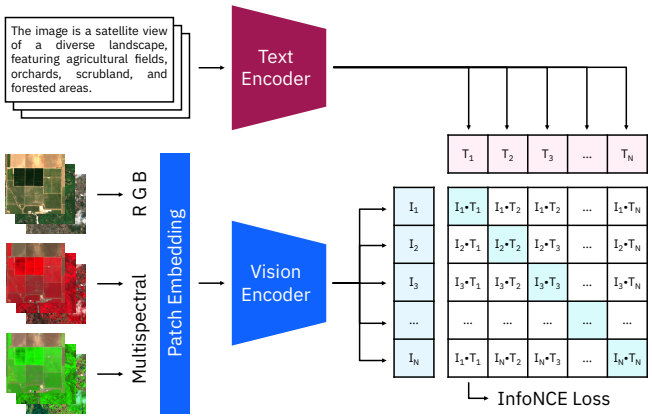


Fig. 4: The CLIP model consists of two encoders for text and images [20]. We extended the RGB patch embeddings to multispectral input and initialized the weights of the additional input channels with zeros. During the continual pre-training, the images and texts of each batch are encoded and combined. The loss increases the similarity of matching pairs while decreasing other combinations.

Following CLIP, we utilize the InfoNCE loss, a contrastive loss function, to align embeddings of semantically similar samples while separating semantically dissimilar ones through cross-modal supervision. The InfoNCE loss encourages the embeddings of matching (positive) pairs (x_i, y_i) to be similar, while pushing apart non-matching (negative) pairs (x_i, y_j) with $j \neq i$. Here, $\text{sim}(\cdot, \cdot)$ is a pairwise similarity measure, and τ is a temperature parameter that scales the logits. Minimizing the loss thus maximizes the similarity of each positive pair relative to all negative pairs. We summarize the training objective in equation 1.

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{\text{sim}(x_i,y_i)/\tau}}{\sum_{j=1}^{N}e^{\text{sim}(x_i,y_j)/\tau}} + \log\frac{e^{\text{sim}(y_i,x_i)/\tau}}{\sum_{j=1}^{N}e^{\text{sim}(y_i,x_j)/\tau}}\right] \quad (1)$$

5 Experimental Setup

In the following, we outline our pre-training and evaluation setting, including the downstream datasets and benchmark models.

Pre-training. We use the implementation and model weights provided by OpenCLIP [9] and perform continual pre-training of the ViT-B/16 using the SSL4EO-S12-captions. The model has 150 million parameters, split between the image and text encoders, and was initially pre-trained on LAION-2B, an English subset of the LAION-5B [21] dataset. While many EO models use the ViT-B/32 version with a patch size of 32 [28,25,16], we find a patch size of

16 more appropriate for low-resolution images with many details. We used an AdamW optimizer with a learning rate of 4e-5, 50 warm-up steps, and a cosine decay scheduler that updated after each training step. The model was trained for up to five epochs on NVIDIA-A100 GPUs with a global batch size of 1200. The final model was selected based on the lowest validation loss reached after one epoch for the multispectral version and two epochs for the RGB version. Based on prior experiments, all layers are unfrozen during the pre-training of Llama3-MS-CLIP, but only the projection layers are trained for the RGB data.

Benchmark Models. We evaluate OpenCLIP [5] ViT-B/16 and ViT-L/14 as well as three RGB-based EO-specific models based on ViT-B/32 backbones. SkyCLIP [25] used remote sensing images with rich semantics covered in Open Street Map to construct a dataset comprising 2.6 million images and generated captions with a simple heuristic by just listing all Open Street Map tags. They performed fully unfrozen continual pre-training using the ViT B/32 backbone initialized from the LAION 2B weights by OpenCLIP [9]. RemoteCLIP [16] proposed a data scaling approach to existing datasets via annotation unification. For images with bounding box annotations, a box-to-caption generation approach was applied. The mask-to-box conversion method was used to generate captions for datasets with available semantic segmentation. The resulting high-resolution dataset consisted of 165k images, each accompanied by five captions. RemoteCLIP is based on the OpenAI CLIP weights [20] and was adapted with fully unfrozen weights. The authors of GeoRSCLIP [28] used the images from BigEarthNet [22] and ten other datasets. They generated captions based on the annotations and metadata using BLIP-2 [13]. Subsequently, they fine-tuned the OpenAI ViT B/32 and ViT H/14 models applying parameter-efficient fine-tuning techniques.

Downstream Datasets. Our zero-shot evaluation focuses on low-resolution Sentinel-2 imagery from EuroSAT [8], BigEarthNet [22], and METER-ML [29]. EuroSAT includes 64×64 patches from ten land-use/land-cover (LULC) classes. BigEarthNet consists of S-2 L2A patches with 19 multi-labels covering a more diverse set of LULC classes and has an input size of 120×120 pixels. Finally, METER-ML covers images with seven classes of different methane sources like *landfills*, *coal mines*, or *natural gas processing plants*. METER-ML includes S-2 images of size 72×72 and high-resolution RGB images from NAIP with size 720×720. Methane is visible in the SWIR S-2 bands (bands 11 and 12) and, therefore, is an especially interesting downstream task. Since EuroSAT and METER-ML only include S-2 L1C data, we downloaded L2A data for their test sets to better align the inputs with the pre-training data. We observed improvements for METER-ML and therefore evaluated on L2A data for this task. We perform additional experiments with the METER-ML-NAIP data and the RESISC45 [4] dataset. The latter includes RGB images of size 256×256 with a spatial resolution ranging from 30 m to 0.2 m. The 45 scene classes range from

landscapes like *wetland* to large objects like *airplane*.

Evaluation. We assess our model's zero-shot capabilities on previously unseen EO datasets. Specifically, we evaluate two tasks: zero-shot classification and text-to-image retrieval. We adopt a template-based approach that leverages multiple prompts of the form *a satellite photo of {class name}* and averages these for the class embedding. For zero-shot classification, we compute the similarity between each image and all possible class labels, assigning the class with the highest similarity score. The zero-shot classification performance is measured using macro top-1 accuracy. We use the test set defined by CLIP for EuroSAT [20] and the official test split for all other datasets.

For the multi-label dataset BigEarthNet, we transform each class into a binary classification task. For each class, we calculate the similarity between the image embedding and the respective text embedding of that class and compare it to the mean similarity between the image and all other classes, as formulated in Equation 2. Here, $\hat{y}_i$ is the predicted label for class i, x is the image embedding, c_i is the class embedding, $\mathrm{sim}(\cdot,\cdot)$ is the dot-product similarity, and K is the total number of classes. We also compare this method to a negative-class approach (e.g., *"other features"*), which boosts accuracy but substantially lowers recall and the F1 score (results in the supplementary material).

$$\hat{y}_i = \begin{cases} 1, & \text{if } \mathrm{sim}(x, c_i) > \frac{1}{K-1} \sum_{j \neq i} \mathrm{sim}(x, c_j), \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

For text-to-image retrieval, we calculate the similarity between a given class label and all test images, rank these scores in descending order, and then compute the mean average precision over the top 100 results (mAP@100). We report the average mAP@100 across all classes. As this retrieval procedure is class-based, it naturally extends to both single-label and multi-label datasets.

6 Results

We first analyze Llama3-MS-CLIP's performance and compare it against RGB-based EO models. Next, we evaluate our RGB-only model on two high-resolution tasks. Finally, we present ablation studies to investigate the effects of multispectral continual pre-training.

Table 1 presents the zero-shot classification and retrieval results for Llama3-MS-CLIP, OpenCLIP baselines, and other RGB-based EO VLMs. Llama3-RGB-CLIP is an ablation trained using only the RGB channels of SSL4EO-S12 v1.1. Llama3-MS-CLIP achieves an average top-1 accuracy of 58.54% for classification, surpassing the untuned baseline by +14.48 percentage points (pp), followed by GeoRSCLIP [28] with 51.77%. In text-to-image retrieval, Llama3-MS-CLIP outperforms all other models as well, exhibiting a 9.43pp improvement over its base model, OpenCLIP ViT-B/16. Domain-specific approaches generally outperform general-purpose baselines in zero-shot tasks, except for RemoteCLIP.

Table 1: Evaluation results on EuroSAT, BigEarthNet, METER-ML, and the overall average. We report zero-shot classification results in accuracy (%) ↑ and text-to-image retrieval results in mAP@100 (%) ↑. The best-performing model is highlighted in bold, and the second-best model is underlined.

	Zero-shot classification				Text-to-image retrieval			
Model	ESAT	BEN	M-ML	Avg	ESAT	BEN	M-ML	Avg
OpenCLIP B/16 [20]	39.36	54.28	38.54	44.06	48.77	26.70	24.62	33.36
OpenCLIP L/14 [20]	46.90	54.85	42.28	48.01	<u>56.92</u>	28.57	<u>28.99</u>	<u>38.16</u>
GeoRSCLIP [28]	52.92	<u>58.80</u>	43.59	<u>51.77</u>	51.36	<u>32.80</u>	22.33	35.50
SkyCLIP [25]	47.54	52.88	<u>44.70</u>	48.37	53.96	26.29	28.41	36.22
RemoteCLIP [16]	34.02	52.28	36.42	40.91	34.34	26.62	20.08	27.01
Llama3-RGB-CLIP	<u>52.96</u>	55.23	38.74	48.98	52.72	28.84	25.60	35.72
Llama3-MS-CLIP	**67.86**	**59.63**	**48.13**	**58.54**	**63.03**	**35.62**	**29.72**	**42.79**

These findings underscore the effectiveness of our curated dataset and the importance of multispectral pre-training. While our RGB-based variant yields only a minor improvement over the baseline, incorporating multispectral channels leads to a substantial performance gain. Notably, GeoRSCLIP [28], despite being adapted with five times more training samples, still falls short of bridging the gap created by the missing multispectral information.

We provide example predictions from Llama3-MS-CLIP in Figure 5 to illustrate common behavior. On EuroSAT, the model is nearly always correct for general classes like *residential*, *sea/lake*, or *industrial*, but it confuses certain pairs such as *permanent crop* and *annual crop*. On METER-ML, samples are often mistakenly classified as *wastewater treatment plants*. In contrast, *concentrated animal feedings operations* (COFAs) like farms and *other features* (negative class) are mostly correctly identified.

In BigEarthNet's multi-label scenario, our approach tends to produce numerous false positives. However, these misclassifications are usually semantically correlated (e.g., *inland waters* or *beaches* predictions for *marine waters*). Introducing an extra negative class (*other features*) mitigates false positives but leads to many more false negatives. This results in higher accuracy overall, but the F1 score drops for six of the seven models. Since F1 reflects both precision and recall, we report the original method here and the alternative strategy in the supplementary material. In both cases, MS-CLIP achieves the highest F1 and significantly outperforms five other models in accuracy.

We further analyse the embedding space of the single-label datasets EuroSAT and METER-ML in Figure 6 and compare Llama3-MS-CLIP with the base model. While many EuroSAT classes overlap in the OpenCLIP UMAP visualization, the clusters are more distinct for Llama3-MS-CLIP. Differentiating between classes in METER-ML is much more difficult, which is reflected in the lower accuracy and the embedding space visualization with no large clusters for

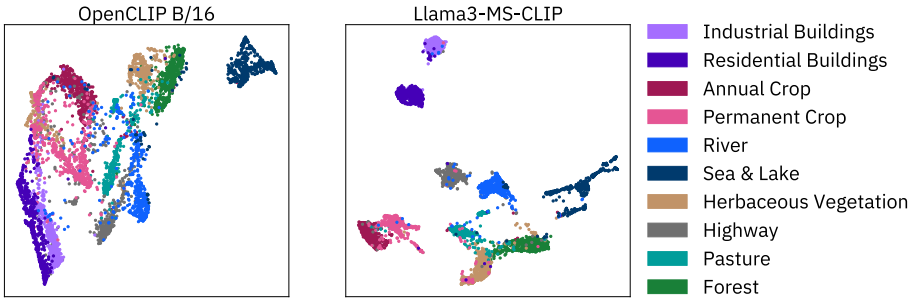
Fig. 5: Prediction examples from Llama3-MS-CLIP.

any model. However, some classes like *wastewater treatment* or *refineries and terminals* form small clusters in the Llama3-MS-CLIP plot, while being more distributed for OpenCLIP.

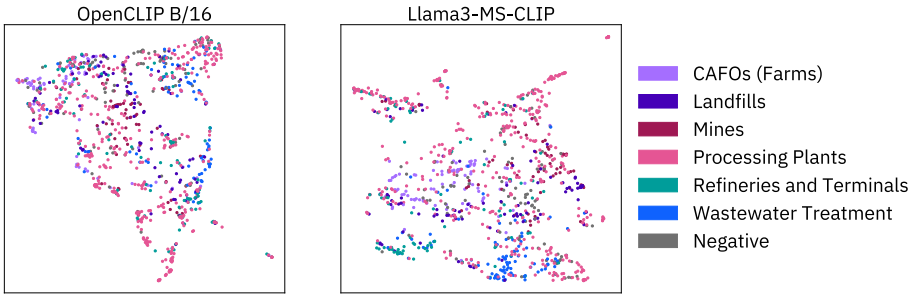
6.1 RGB Experiments

We conduct additional experiments on two RGB datasets that feature high-resolution imagery, aiming to assess the generalization capabilities of our models trained solely on low-resolution Sentinel-2 data. Table 2 shows that Llama3-RGB-CLIP achieves results on par with the untuned base model. While it scores 1.71pp lower for METER-ML-NAIP classification, it outperforms the baseline in all other tasks. GeoRSCLIP achieves the best zero-shot classification but underperforms in retrieval tasks, and RemoteCLIP again delivers the lowest results across metrics.

Comparing performance on METER-ML for both Sentinel-2 and NAIP data reveals that all models benefit from the higher resolution, with improvements ranging from 3pp to 17pp in both classification and retrieval. Our RGB variant is 5.25pp more accurate than its multispectral version in classification using S-2 imagery. Although high-resolution imagery clearly enhances VLM performance, its public availability is limited, whereas Sentinel-2 data is openly accessible every five days. Notably, relying on low-resolution data for domain adaptation does not reduce performance on high-resolution tasks, and other domain-specific approaches, even those trained on high-resolution data, only show marginal improvements over the baseline.



(a) EuroSAT



(b) METER-ML

Fig. 6: UMAP plot of the embedding spaces from OpenCLIP B/16 and Llama3-MS-CLIP for EuroSAT and METER-ML using the S-2 test images. The UMAP settings are 10 nearest neighbors and min distance 0.0.

6.2 Ablation Studies

We conduct several ablation experiments to study various design choices systematically. Specifically, we investigate prompt templates, weight initialization for the patch embedding, different input bands, and strategies for freezing/unfreezing model layers.

Our initial prompt templates follow RS5M, which are already adapted for EO and outperform the original CLIP templates used for EuroSAT (see Table 3). Extending and refining these prompts leads to a gain of on average 3.47pp in classification and 1.16pp in retrieval, relative to the baseline.

Figure 7(a) compares two approaches to initialize the multispectral patch embeddings. While a naive solution is to set each new channel's weights to the mean of the RGB ones, we find that zero-initialization produces superior performance in all six tasks. We hypothesize that starting from zero with a short warm-up phase lets the model adjust gradually to the additional input channels instead of the more sudden interruption with mean initialization.

We analyse the weights of the patch embedding before and after continual pre-training. The absolute values of the patch embedding for the multispectral

Table 2: Zero-shot evaluation results for the high-resolution RGB datasets METER-ML NAIP and RESISC45. We report zero-shot classification results in accuracy (%) ↑ and text-to-image retrieval results in mAP@100 (%) ↑. The two best-performing models are highlighted in bold and underlined.

	Zero-shot classification			Text-to-image retrieval		
Model	M-ML-N	RESISC45	Avg	M-ML-N	RESISC45	Avg
OpenCLIP B/16 [20]	<u>55.09</u>	67.45	61.27	37.43	64.30	50.87
OpenCLIP L/14 [20]	53.48	**72.82**	<u>63.15</u>	**42.55**	**70.88**	**56.72**
GeoRSCLIP [28]	**59.03**	<u>68.28</u>	**63.66**	34.93	60.04	47.49
SkyCLIP [25]	53.88	67.68	60.78	<u>39.02</u>	<u>66.79</u>	<u>52.91</u>
RemoteCLIP [16]	39.35	66.90	53.13	23.80	56.88	40.34
Llama3-RGB-CLIP	53.38	68.26	60.82	38.50	66.49	52.50

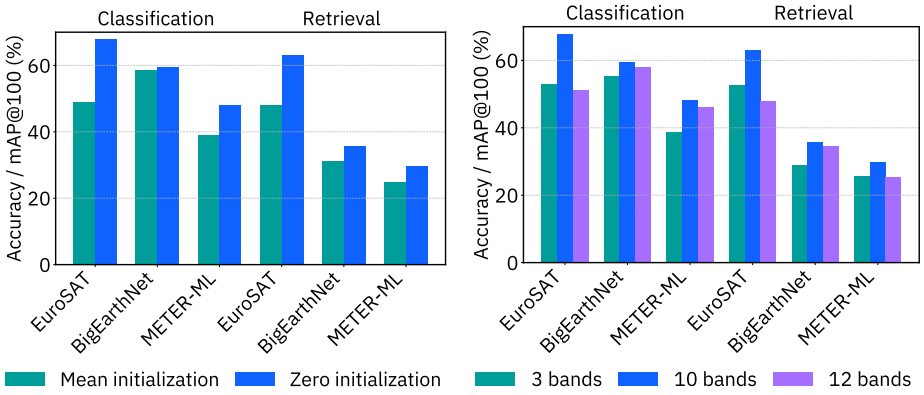
Table 3: Zero-shot evaluation results for different text templates using the Llama3-MS-CLIP model. We report zero-shot classification results in accuracy (%) ↑ and text-to-image retrieval results in mAP@100 (%) ↑. The best-performing method is highlighted in bold.

	Zero-shot classification				Text-to-image retrieval			
Templates	ESAT	BEN	M-ML	Avg	ESAT	BEN	M-ML	Avg
CLIP templates [20]	67.64	60.25	37.33	55.07	61.14	34.31	29.45	41.63
RS5M templates [28]	67.28	**60.17**	47.02	58.15	60.19	34.69	**30.99**	41.95
MS-CLIP templates	**67.86**	59.63	**48.13**	**58.54**	**63.03**	**35.62**	29.72	**42.79**

channels are much lower than for the RGB channels, showing that the model did not fully adapt to the additional input due to the limited number of weight updates with only one million samples. At the same time, changes in the multispectral channels are ∼2800 times higher than the changes in the RGB weights, which are adjusted by only 0.03% compared to the OpenCLIP weights. This shows the additional information Llama3-MS-CLIP leverages from the multispectral channels and highlights the need for even larger multispectral EO datasets for longer pre-training without overfitting.

In Figure 7(b), we examine the effect of using three, ten, or all twelve Sentinel-2 bands. Although using all bands might seem beneficial, certain tasks decline in performance compared to using only the RGB bands. Omitting bands 1 and 10 and training on ten bands leads to the overall best results. The dropped bands both have a 60 m spatial resolution, and their information might not align well with the other bands. Dropping these bands is common among EO models (e.g., in [6]), suggesting low meaningful information for many use cases.

We further compare various strategies for freezing and unfreezing model layers in Table 4. The patch embedding was unfrozen in every setting to adapt



(a) Patch embedding weights (b) Input bands

Fig. 7: Ablation experiments for the initialization of additional patch embedding channels (a) and the selected input bands (b). The final setting is in blue.

Table 4: Ablation experiment with different unfrozen layers. We either unfreeze the attention layers, projection layers, or all layers in the image and text encoders. We report zero-shot classification results in accuracy (%) ↑ and text-to-image retrieval results in mAP@100 (%) ↑. The best-performing method is highlighted in bold.

		Zero-shot classification				Text-to-image retrieval			
Image enc.	Text enc.	ESAT	BEN	M-ML	Avg	ESAT	BEN	M-ML	Avg
Attention l.	Attention l.	56.12	58.05	45.30	53.15	48.76	33.80	24.96	35.84
Projection l.	Projection l.	32.00	56.25	43.49	43.91	34.70	34.47	21.44	30.20
All layers	None	65.82	58.87	**48.63**	57.77	55.46	34.19	29.08	39.57
All layers	All layers	**67.86**	**59.63**	48.13	**58.54**	**63.03**	**35.62**	**29.72**	**42.79**

to the multispectral input. Freezing layers can, in principle, reduce the risk of catastrophic forgetting. Yet, our experiments show that fully unfreezing the image and text encoders leads to the best results when adapting to multispectral data. By contrast, selectively fine-tuning only the projection layer yields lower performance than the baseline. We performed a similar ablation study for Llama3-RGB-CLIP with contrary results. Fine-tuning only the projection layer resulted in the best performance. Keeping the earlier layers frozen avoids forgetting pre-trained features with RGB input, but it cannot capture the multispectral information when including the additional multispectral channels.

7 Discussion and Limitations

In this work, we demonstrate the benefit of leveraging MLLMs for accelerating image captioning in order to curate datasets for subsequent vision-language model training. While MLLMs are already able to digest and caption RGB representations of remote sensing images themselves, they cannot leverage multispectral data—leaving room for specialized models for Earth observation. We address this gap by coupling the automated caption generation on RGB imagery with multi-spectral data for the Llama3-MS-CLIP. However, we acknowledge limitations when using MLLMs to generate synthetic image captions. First, we understand that there exists a risk of propagating errors or biases (e.g., in the form of hallucinations) of the MLLM further into the self-supervised models that are trained on top of the synthetic data. It will be relevant to identify such ripple effects that result from training self-supervised models on synthetically generated data of MLLMs in order to understand how errors and biases are propagated, reduced, or reinforced by downstream models. Second, we note that the existing dataset likely benefits from increased diversity, as the word count graph in the supplementary section highlights a trend of similar topics in many of the captions. Third, we advocate for considering human-in-the-loop systems during the caption generation process in settings where errors by the MLLMs are not acceptable.

Based on the synthetically generated captions, Llama3-MS-CLIP demonstrates the benefit of using multi-spectral data during pretraining. Even though we observe significant performance improvements when leveraging multi-spectral data instead of RGB data, our experiments also show that the model is potentially not yet saturated. This is indicated by comparably low weights in the patch embedding of Llama3-MS-CLIP for the non-visible channels compared to the visible RGB spectrum. Overall, this experiment reinforces our expectation that the performance of the model will further improve with longer continuous pretraining, leveraging a larger and more diverse training corpus.

Finally, we see a possibility for future research to work on integrating and merging pixel-level training strategies with the image-level training we employ in this work. This merge could improve the model's capability to capture detailed image nuances and help differentiate between closely related classes. Pixel-level understanding might also unlock additional progress on other tasks that we did not explore in this work, including semantic segmentation and object detection.

8 Conclusion

We introduce a multispectral vision-language dataset of low-resolution, multispectral Sentinel-2 data with corresponding captions. Our automated captioning strategy scales easily, reducing the need for costly human annotations. On top of this dataset, we build Llama3-MS-CLIP, the first CLIP-like multispectral VLM. Our experiments show that our model significantly improves zero-shot classification and retrieval compared to other methods, even domain-specific adaptations

trained on larger datasets. We see significant potential in leveraging the open-sourced Llama3-MS-CLIP in downstream applications and as a vision encoder for building multispectral MLLMs.

Acknowledgments. We thank the remote sensing experts for reviewing the generated captions and Niklas Kopp for providing the embedding space analysis.

Disclosure of Interests. This work is part of the FAST-EO project funded by the European Space Agency (ESA), contract number 4000143501/23/I-DT.

References

1. Bazi, Y., Bashmal, L., Al Rahhal, M.M., Ricci, R., Melgani, F.: Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. Remote Sensing **16**(9) (2024). https://doi.org/10.3390/rs16091477
2. Blumenstiel, B., Braham, N.A.A., Albrecht, C.M., Maurogiovanni, S., Fraccaro, P.: Ssl4eo-s12 v1. 1: A multimodal, multiseasonal dataset for pretraining, updated. arXiv preprint arXiv:2503.00168 (2025)
3. Blumenstiel, B., Jakubik, J., Kühne, H., Vössing, M.: What a mess: Multi-domain evaluation of zero-shot semantic segmentation. Advances in Neural Information Processing Systems **36**, 73299–73311 (2023)
4. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
5. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
6. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. Advances in Neural Information Processing Systems **35**, 197–211 (2022)
7. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
8. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
9. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (2021). https://doi.org/10.5281/zenodo.5143773
10. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al.: Terramind: Large-scale generative multimodality for earth observation. arXiv preprint arXiv:2504.11171 (2025)
11. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)

12. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: Llava-next: Stronger llms supercharge multimodal capabilities in the wild (2024), https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
15. Li, X., Wen, C., Hu, Y., Zhou, N.: Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. International Journal of Applied Earth Observation and Geoinformation **124**, 103497 (2023)
16. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing (2024)
17. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
18. Mall, U., Phoo, C.P., Liu, M.K., Vondrick, C., Hariharan, B., Bala, K.: Remote sensing vision-language foundation models without annotations via ground remote alignment. arXiv preprint arXiv:2312.06960 (2023)
19. Qu, B., Li, X., Tao, D., Lu, X.: Deep semantic understanding of high resolution remote sensing image. In: 2016 International conference on computer, information and telecommunication systems (Cits). pp. 1–5. IEEE (2016)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
22. Sumbul, G., Charfuelan, M., Demir, B., Markl, V.: Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In: IGARSS IEEE international geoscience and remote sensing symposium. pp. 5901–5904. IEEE (2019)
23. Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, Þ.E., Blumenstiel, B., Ghosal, R., de Oliveira, P.H., Almeida, J.L.d.S., Sedona, R., Kang, Y., et al.: Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. arXiv preprint arXiv:2412.02732 (2024)
24. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. IEEE Geoscience and Remote Sensing Magazine **11**(3), 98–106 (2023)
25. Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R.: Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5805–5813 (2024)
26. Yuan, Z., Xiong, Z., Mou, L., Zhu, X.X.: Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. Earth System Science Data Discussions **2024**, 1–24 (2024)

27. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. arXiv preprint arXiv:2204.09868 (2022)
28. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: 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 (2024)
29. Zhu, B., Lui, N., Irvin, J., Le, J., Tadwalkar, S., Wang, C., Ouyang, Z., Liu, F.Y., Ng, A.Y., Jackson, R.B.: Meter-ml: a multi-sensor earth observation benchmark for automated methane source mapping. arXiv preprint arXiv:2207.11166 (2022)
30. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geoscience and Remote Sensing Magazine **5**(4) (2017)