

GeoLangBind: Unifying Earth Observation with Agglomerative Vision–Language Foundation Models

Zhitong Xiong¹
Nils Lehmann¹

Yi Wang¹ Weikang Yu^{1,2} Adam J Stewart¹ Jie Zhao¹
Thomas Dujardin¹ Zhenghang Yuan¹ Pedram Ghamisi²
Xiao Xiang Zhu^{1*}

¹ Technical University of Munich ² Helmholtz-Zentrum Dresden-Rossendorf

{zhitong.xiong, xiaoxiang.zhu}@tum.de

Abstract

Earth observation (EO) data, collected from diverse sensors with varying imaging principles, present significant challenges in creating unified analytical frameworks. We present GeoLangBind, a novel agglomerative vision–language foundation model that bridges the gap between heterogeneous EO data modalities using language as a unifying medium. Our approach aligns different EO data types into a shared language embedding space, enabling seamless integration and complementary feature learning from diverse sensor data. To achieve this, we construct a large-scale multimodal image–text dataset, GeoLangBind-2M, encompassing six data modalities. GeoLangBind leverages this dataset to develop a zero-shot foundation model capable of processing arbitrary numbers of EO data channels as input. Through our designed Modality-aware Knowledge Agglomeration (MaKA) module and progressive multimodal weight merging strategy, we create a powerful agglomerative foundation model that excels in both zero-shot vision–language comprehension and fine-grained visual understanding. Extensive evaluation across 23 datasets covering multiple tasks demonstrates GeoLangBind’s superior performance and versatility in EO applications, offering a robust framework for various environmental monitoring and analysis tasks. The dataset and pretrained models will be publicly available.

1. Introduction

Earth observation (EO) technologies have undergone remarkable advancement in recent years, generating vast amounts of multimodal geospatial data spanning different sensors and scales [54, 55, 59, 70]. In response to the exponential growth of remote sensing data, researchers have proposed foundation models pretrained on large-scale EO data, including SatMAE [12], Scale-MAE [38], HyperSIGMA [46], SSL4EO [39, 47], and SpectralGPT [20]. However, these

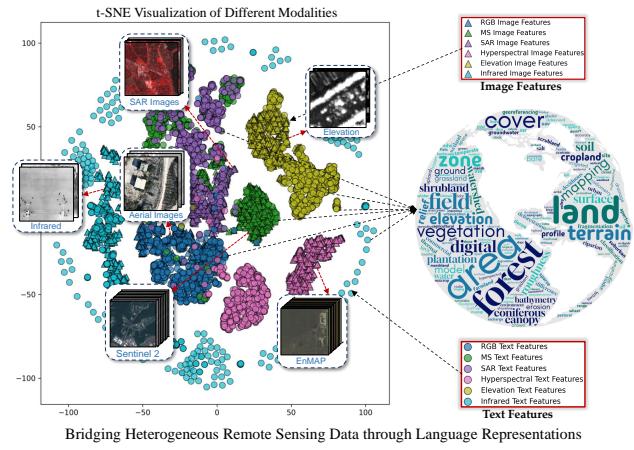


Figure 1. GeoLangBind leverages language as a bridge to unify heterogeneous remote sensing data, enabling a common representation space across diverse modalities for improved multimodal learning.

models typically focus on a single EO data modality, limiting their ability to leverage complementary characteristics across different sensor types.

To address this limitation, recent works such as OFA-Net [58], DOFA [57], AnySat [5], OmniSat [6], and FoMo-Net [8] have introduced foundation models capable of learning from multiple EO modalities within a unified framework. These approaches have demonstrated promising results in multimodal EO representation learning. However, their focus remains on unsupervised visual pretraining, lacking a structured way to align EO data with language for better interpretability and retrieval. Meanwhile, multimodal large language models like GPT [33], LLaMA [44], LLaVA [29], and DeepSeek [27] have shown remarkable capabilities in handling diverse and complex tasks. Vision-language foundation models, such as CLIP [36] and SigLIP [64], have played a crucial role as vision encoders of such multimodal large language models. Inspired by these advances, EO-

specific adaptations like RemoteCLIP [28], RS5M [67], and SkyScript [50] have attempted to bring contrastive learning to remote sensing. However, these models are primarily limited to high-resolution RGB optical imagery, and can not handle other diverse EO modalities (e.g., multispectral data, Synthetic Aperture Radar data) that are critical for many applications.

Training a single vision–language foundation model capable of supporting multiple heterogeneous EO data modalities presents several challenges: 1) the lack of comprehensive multimodal EO datasets with aligned textual descriptions, 2) the design of a flexible architecture to handle variable-length input channels across different modalities, and 3) addressing the imbalance among different modalities in the training data. To address such challenges, we propose GeoLangBind, a unified multimodal vision–language foundation model that handles diverse data modalities and aligns them to the shared language feature space, as illustrated in Fig. 1. In addition, GeoLangBind enhances fine-grained semantic understanding via knowledge agglomeration [37] from multiple teacher models beyond contrastive learning.

To achieve this, we combine the following key designs: 1) a wavelength-aware dynamic encoder to handle multiple data modalities, 2) a modality-aware knowledge agglomeration module, and 3) progressive weight-space merging [53] to scale the model efficiently across multiple modalities. Our contributions are as follows:

1. We construct GeoLangBind-2M, a comprehensive dataset of two million image–text pairs spanning six distinct EO data modalities to facilitate the training of multimodal geospatial foundation models.
2. We propose an effective Modality-aware Knowledge Agglomeration (MaKA) module to enhance continuous pre-training and improve fine-grained image understanding.
3. We design a progressive weight merging strategy that enables scalable model adaptation across multiple modalities, outperforming traditional data mixing approaches.
4. We conduct extensive experiments on zero-shot classification, semantic segmentation, and cross-modal image retrieval, demonstrating the superiority of GeoLangBind across diverse tasks.

Benchmarking results indicate that GeoLangBind achieves state-of-the-art performance while being flexible in handling diverse data modalities with any number of spectral bands.

2. Related work

Vision foundation models in Earth observation Existing pretrained models predominantly focus on RGB data, such as GFM [31], Scale-MAE [38], and Cross-Scale MAE [43]. Others specialize in multispectral imagery, with FG-MAE [48] and SatMAE [12] tailored for Sentinel-2 data, while SSL4EO-L [39] is trained on Landsat imagery. SpectralGPT [20] leverages a 3D generative transformer for

spectral data analysis. Beyond optical data, CROMA [14] introduces dual unimodal encoders for multispectral and Synthetic Aperture Radar (SAR) data, using a cross-modal radar–optical transformer to learn unified deep representations. DeCUR [49] enhances bimodal learning by decoupling unique and shared features across modalities. Meanwhile, Satlas [7] assembles a large-scale multi-sensor dataset, pre-training separate models for each sensor. DOFA [57] introduces a neural plasticity-inspired hypernetwork that adapts to different sensor wavelengths, enabling joint training on five distinct sensors. OmniSat [6] proposes to fuse features across multiple EO modalities to learn robust multimodal representations without labeled data. SkySense [18] utilizes separate visual encoders for different data modalities and designs a dedicated module for multimodal fusion.

Zero-shot foundation models Zero-shot foundation models like CLIP [36] have significantly advanced computer vision by enabling models to generalize across tasks without explicit task-specific training. Building upon CLIP, SigLIP [64] introduces a pairwise sigmoid loss function to enlarge the training batch and enhance the model performance. Zero-shot foundation models have also been investigated in EO applications. SkyScript [50] constructs a large-scale image–text dataset by linking remote sensing images with OpenStreetMap [34] semantics. RemoteCLIP [28] introduces an image–text dataset by generating captions from existing annotated datasets. GeoRSCLIP [67] introduces RS5M, an image–text dataset with 5M pairs, created by filtering existing datasets and generating captions for label-only remote sensing data. Despite these advances, existing vision–language models mainly focus on RGB data and have limitations in learning representations from multiple EO data modalities, which are crucial for holistic environmental analysis. To unify multiple data modalities, ImageBind [16] aligns six modalities within a shared embedding space. LanguageBind [69] extends this by using language as a universal intermediary for modality alignment. While these approaches demonstrate the potential of language or vision as unifying mediums, applying them to EO data is still challenging due to the scarcity of multimodal image–text datasets and the diverse scales of Earth systems.

3. Dataset construction: GeoLangBind-2M

GeoLangBind-2M contains a broad range of data modalities including RGB, SAR, multispectral, hyperspectral, infrared, and elevation data, as listed in Tab. 1. Each image is paired with a high-quality text annotation, as illustrated in Fig. 2. All datasets are implemented using the TorchGeo library [40]. In the following, we introduce the composition of the dataset.

RGB image–text datasets High-resolution RGB images with textual descriptions form a significant portion of GeoLangBind-2M. We incorporate parts of the RemoteCLIP [28] dataset, i.e. Seg4 and Det10, which are

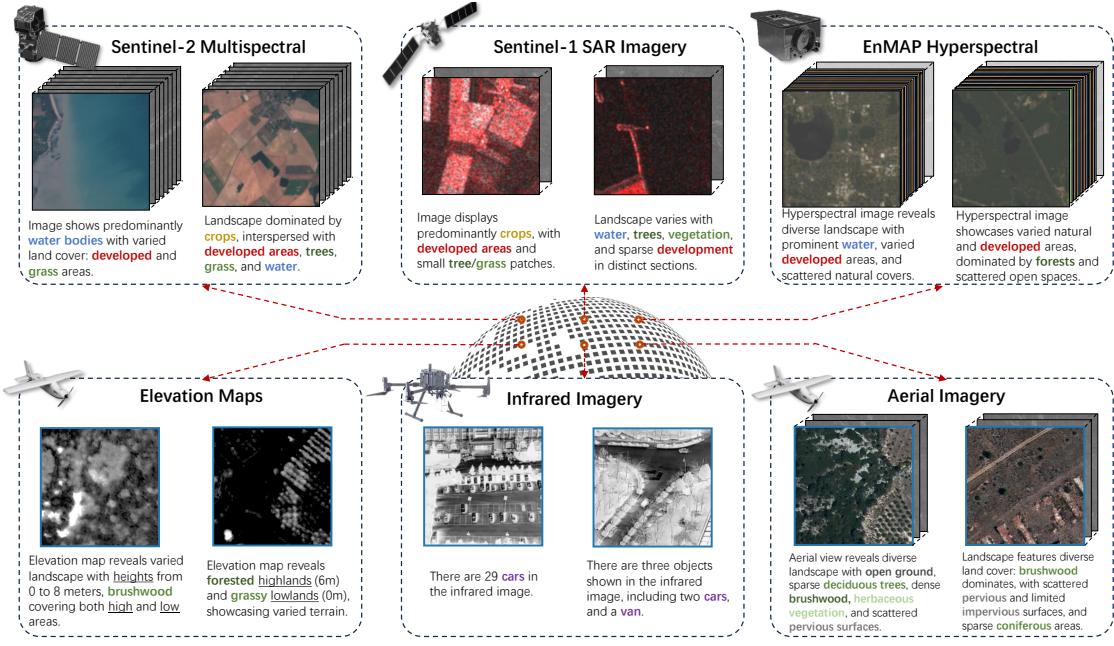


Figure 2. Visualization of data samples from our GeoLangBind-2M. The dataset includes imagery from six different sensors and modalities: Sentinel-2 multispectral, Sentinel-1 SAR, EnMAP hyperspectral, elevation maps, infrared imagery, and aerial imagery. Each sample is paired with textual descriptions capturing key land cover types, objects, and geographic features.

Table 1. Summary of dataset sources of GeoLangBind-2M. * indicates image–text datasets that are created in this work. SAR = synthetic aperture radar, MSI = multispectral imagery, HSI = hyperspectral imagery, and IR = infrared.

Image–Text Datasets	Knowledge Domain	# Samples
Seg4	General segmentation	41,172
Det10	General detection	110,800
Flair2-RGB-caption*	Land cover analysis	61,711
VRS-train	General detection	20,264
SkyScript dataset	OpenStreetMap tags	1,518,888
SAR Datasets		
MMflood-caption*	Flood mapping	6,181
SAR-ship-caption*	Ship detection	5,984
ChatEarthNet-SAR-caption*	Land cover analysis	95,620
MSI/HSI Datasets		
ChatEarthNet-S2-caption*	Land cover analysis	95,620
NLCD-hyper-caption*	Land cover analysis	15,000
Elevation Dataset		
Flair2-Elevation-caption*	Digital Elevation Model	61,711
Infrared Dataset		
IR-ship-caption*	Ships in infrared imagery	18,032
Total	—	2,050,983

constructed by generating captions from existing segmentation and detection datasets. We also include VRS-Bench [26], which aggregates multiple object detection datasets. SkyScript [50] is a large-scale image–text dataset for EO, from which we include the top 30% highest quality

samples after filtering. Beyond existing datasets, we introduce Flair2-RGB-caption [15], a newly created image–text dataset. Using semantic information as prompts, we employ the Pixtral-12B [2] model to generate captions for detailed land cover features.

SAR image–text datasets To enhance the dataset with SAR–text pairs, we create MMflood-SAR-caption [32], which contains captions emphasizing water extent and flooded regions. Additionally, the SAR-ship-caption dataset includes captions describing ships in SAR images. ChatEarthNet-SAR-caption further extends the collection by covering diverse SAR scenes from Sentinel-1, focusing on land cover types.

Multispectral, hyperspectral, and elevation image–text datasets We construct ChatEarthNet-S2-caption by summarizing long, detailed descriptions from the ChatEarthNet dataset [63]. This dataset contains Sentinel-2 multispectral imagery covering more than 10 land cover types. Furthermore, we introduce NLCD-hyper-caption, a hyperspectral dataset with 200+ EnMAP bands [17]. Like Flair2-RGB-caption, we use Pixtral-12B to generate rich captions, leveraging annotated land cover maps for enhanced descriptiveness. Similarly, we also generate captions using Pixtral-12B to construct the Flair2-Elevation-caption dataset. Further details on the construction of GeoLangBind-2M are provided in the **Supplementary materials**.

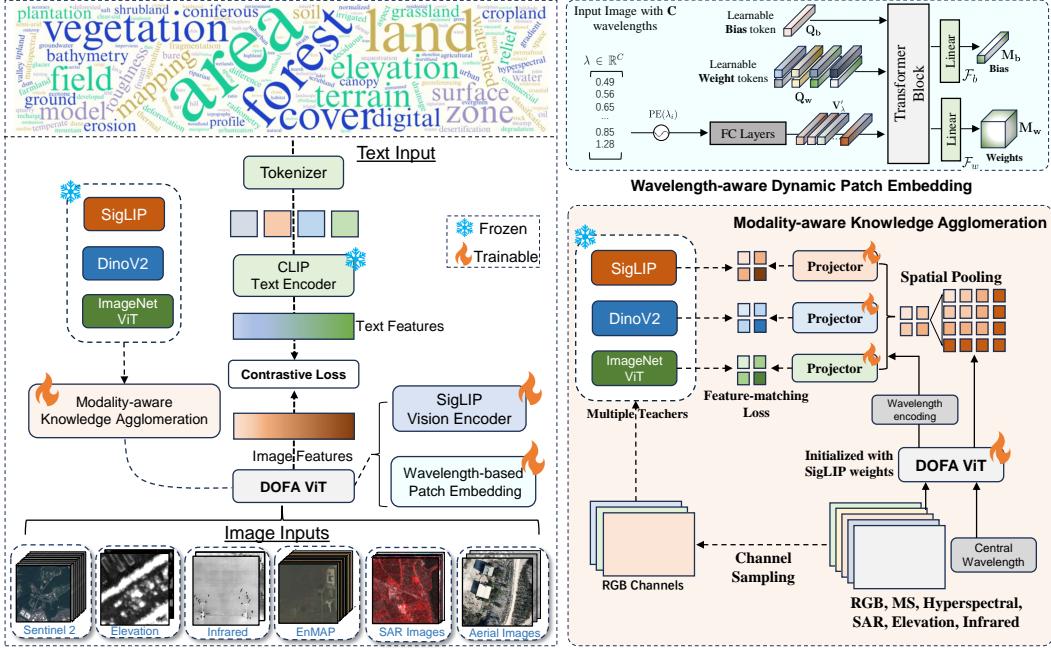


Figure 3. GeoLangBind aims to unify and align diverse Earth observation data modalities into the shared language feature space while enhancing fine-grained image understanding. The model uses modality-aware knowledge agglomeration (MaKA) to transfer knowledge from multiple teachers. The wavelength-aware dynamic patch embedding module processes multimodal images dynamically.

4. Methodology

4.1. GeoLangBind model design

GeoLangBind aims to unify diverse data modalities and embed them into a shared language feature space while enhancing fine-grained image understanding ability. To this end, we design a wavelength-aware dynamic encoder, a modality-aware knowledge agglomeration module, and a progressive weight-space merging approach.

4.2. Wavelength-aware dynamic encoder

To handle the diversity of spectral bands across modalities, we adopt the dynamic encoder architecture from DOFA [57], as illustrated in Fig. 3. Given an input image $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ with C spectral channels, we define its corresponding central wavelengths as $\lambda \in \mathbb{R}^C$. For modalities without well-defined wavelengths, such as elevation, we assign default wavelengths equivalent to those of the RGB channels. First, a 1D sine-cosine positional encoding is applied to embed these wavelengths into a higher-dimensional space: $\mathbf{V}_\lambda = PE(\lambda) \in \mathbb{R}^{C \times d_\lambda}$, where d_λ is the embedding dimension. The encoded wavelengths \mathbf{V}_λ are then transformed using two fully-connected layers with residual connections.

Next, a lightweight Transformer [45] encoder processes the transformed wavelength embeddings. Specifically, we concatenate the transformed embeddings \mathbf{V}'_λ with the learnable weight query tokens \mathbf{Q}_w and a learnable bias query

token \mathbf{Q}_b as input to the Transformer encoder. After the Transformer layer, the output features are passed through two fully-connected layers to generate the dynamic convolutional weights \mathbf{M}_w and biases \mathbf{M}_b . Finally, these dynamically generated weights and biases are used for patch embedding (implemented as convolution). This process enables the model to handle different data modalities with varying spectral channels using a unified deep encoder.

4.3. Modality-aware knowledge agglomeration

We introduce the modality-aware knowledge agglomeration (MaKA) module, which leverages wavelength as a modality-specific condition when distilling features from different data modalities. MaKA comprises three key components: a wavelength-aware prompt generator, which encodes wavelength values into a learnable embedding vector; a modality-aware conditional layer normalization, which adapts normalization parameters based on modality-specific characteristics; and a feature projector, which transforms input features to facilitate effective feature matching and knowledge distillation. Below, we introduce each of them in detail.

Wavelength-aware prompt generator Each modality is associated with a set of wavelengths, which we encode using a sine-cosine positional encoding function. Given input wavelengths $\lambda \in \mathbb{R}^C$, we first obtain their positional embeddings $\mathbf{V}_\lambda \in \mathbb{R}^{C \times d_\lambda}$. To generate the modality-aware prompt vector, we apply a linear projection $\mathbf{W}_{proj} \in \mathbb{R}^{d_\lambda \times d_\lambda}$ to each

row of \mathbf{V}_λ , then compute the mean over the C wavelengths:

$$\mathbf{V}_p = \frac{1}{C} \sum_{i=1}^C (\mathbf{W}_{\text{proj}} \mathbf{V}_{\lambda_i}), \quad (1)$$

where $\mathbf{V}_p \in \mathbb{R}^d$ is the fused modality embedding. This final embedding vector \mathbf{V}_p captures modality-specific information derived from the provided wavelengths.

Modality-aware layer normalization Given the feature map from the last layer of the GeoLangBind model, we first interpolate it to match the teacher’s feature resolution, denoted as $\mathbf{F} \in \mathbb{R}^{d \times H' \times W'}$, omitting the batch dimension for simplicity. With \mathbf{F} and a modality prompt $\mathbf{V}_p \in \mathbb{R}^d$, our goal is to perform layer normalization that dynamically adapts to the modality. To achieve this, we introduce modality-aware conditional layer normalization, formulated as follows.

First, we derive two learnable modulation vectors, $\gamma, \beta \in \mathbb{R}^d$, from the modality prompt via a linear transformation:

$$\begin{bmatrix} \gamma \\ \beta \end{bmatrix} = \mathbf{W}_{\text{prompt}} \mathbf{V}_p \in \mathbb{R}^{2d}, \quad (2)$$

where $\mathbf{W}_{\text{prompt}} \in \mathbb{R}^{2d \times d}$. We then split the output into $\gamma, \beta \in \mathbb{R}^d$ and reshape them to broadcast over the spatial dimensions:

$$\gamma \rightarrow (1, d, 1, 1), \quad \beta \rightarrow (1, d, 1, 1). \quad (3)$$

Let $\text{LN}(\mathbf{F})$ denote the standard channel-wise layer normalization of \mathbf{F} . The modality-aware layer normalization is then defined as:

$$\mathbf{F}' = \text{LN}(\mathbf{F}) + \gamma \odot \text{LN}(\mathbf{F}) + \beta, \quad (4)$$

where \odot denotes the Hadamard product. Note that we use the modality-aware layer normalization to learn the residual information specific to each modality.

Projector module A convolutional layer is applied to adjust the feature representation \mathbf{F}' to channel dimensions d_t of the corresponding teachers. The final representation is then used to compute the feature-matching loss for model distillation. Our design incorporates modality awareness into the feature translation process by encoding wavelength information and adapting the layer normalization parameters accordingly. This ensures that the learned features are both *aligned* and *modality-specific*, enhancing robust adaptation across different input modalities.

Model training To train the model, we use a feature-matching loss to align the translated spatial features with the teacher’s features. As illustrated in Fig. 3, we extract three RGB channels from multispectral or hyperspectral images as input to the teacher model. The student features $\mathbf{F}_s \in \mathbb{R}^{d_t \times H' \times W'}$ are then matched to the teacher’s features $\mathbf{F}_t \in \mathbb{R}^{d_t \times H' \times W'}$ using a combination of L_1 loss, cosine

embedding loss to maximize the cosine similarity, and mean squared error (MSE) loss:

$$L_{\text{match}} = L_{L_1}(\mathbf{F}_s, \mathbf{F}_t) + L_{\text{mse}}(\mathbf{F}_s, \mathbf{F}_t) + L_{\cos}(\mathbf{F}_s, \mathbf{F}_t). \quad (5)$$

We utilize three pre-trained models as teachers: SigLIP [64], DINOv2 [35], and ViT [13]. During training, the GeoLangBind model is initialized with SigLIP pre-trained weights and trained on GeoLangBind-2M using both the pairwise sigmoid loss L_{siglip} for contrastive learning and distillation loss for feature matching. The total loss function is:

$$L = L_{\text{siglip}} + \alpha_s L_{\text{match}}^{\text{siglip}} + \alpha_d L_{\text{match}}^{\text{dino2}} + \alpha_v L_{\text{match}}^{\text{vit}}, \quad (6)$$

where α_s, α_d , and α_v are balancing factors for feature-matching losses of SigLIP, DINOv2, and ViT, respectively.

4.4. Progressive Multimodal Weight Merging

To address the data imbalance issue, we propose a post-training weight-space merging strategy to scale the model to diverse data modalities effectively. Since RGB data is the most dominant modality, we split the dataset into two parts: data with RGB channels and data with other modalities. During training, we first train two separate GeoLangBind models on these subsets. After training, we merge their weights to obtain a unified model capable of handling both RGB and other modalities. Since the original SigLIP model has been trained on massive data, we also merge the weights of GeoLangBind with the original SigLIP weights to enhance the model. Thus, we perform two rounds of weight merging to merge these three model weights: 1) merge the original SigLIP weights with the model trained on the RGB subset of GeoLangBind-2M (GeoLangBind-RGB); 2) merge the intermediate model with the model trained on the remaining five data modalities (GeoLangBind-Others).

For clarity, let θ_{siglip} be the original SigLIP model weights, θ_{rgb} be the GeoLangBind model weights trained on the RGB subset, θ_{others} be the GeoLangBind model weights trained on the other five modalities. In the first step, we merge θ_{siglip} and θ_{rgb} using a simple linear weight merging strategy:

$$\theta^* = (1 - m_1) \theta_{\text{siglip}} + m_1 \theta_{\text{rgb}}, \quad (7)$$

where m_1 controls the weighting ratio between the two models. Next, we merge the intermediate weights θ^* with θ_{others} to obtain the final GeoLangBind model:

$$\theta = (1 - m_2) \theta^* + m_2 \theta_{\text{others}}, \quad (8)$$

where m_2 determines the contribution of the non-RGB modalities. We conduct ablation studies in Sec. 5.3 by varying m_1 and m_2 to search for optimal weighting ratios. Although more sophisticated weight merging methods exist [3], we adopt a simple linear approach to evaluate its effectiveness for heterogeneous EO data. Investigating advanced merging strategies remains a topic for future work.

Table 2. Zero-shot comparison of various models on scene and fine-grained classification tasks. Bold values indicate the best performance.

ViT	Model	Scene classification								Fine-grained classification			
		SkyScript	AID	EuroSAT	fMoW	Million-AID	PatternNet	RESISC	RSI-CB	Avg.	Roof shape	Smoothness	Surface
Base	CLIP-original	40.16	69.55	32.11	17.62	57.27	64.09	65.71	41.26	49.66	31.50	26.80	61.36
	Human-curated captions	40.03	71.05	33.85	18.02	57.48	66.56	66.04	42.73	50.82	28.50	27.80	60.91
	RemoteCLIP	27.06	87.05	30.74	11.13	46.26	56.05	67.88	44.55	49.09	30.50	21.00	43.86
	CLIP-laion-RS	40.77	69.55	37.63	19.16	56.59	64.79	64.63	41.79	50.59	28.83	27.60	62.27
	SkyCLIP-50	52.98	70.90	33.30	19.24	62.69	72.18	66.67	46.20	53.02	26.00	38.00	67.73
Large	GeoLangBind-B-224	70.39	77.60	52.30	20.17	64.91	76.68	67.21	49.53	59.85	44.33	20.00	72.73
	CLIP-original	55.06	69.25	41.89	26.19	57.88	71.39	66.70	43.02	53.76	37.50	25.40	42.73
	Human-curated captions	56.09	72.95	41.96	26.33	58.47	74.86	68.70	44.60	55.41	37.00	26.60	40.00
	RemoteCLIP	34.40	70.85	27.81	16.77	47.20	61.91	74.31	50.79	49.99	34.33	34.20	55.45
	CLIP-laion-RS	58.81	71.70	54.30	27.21	60.77	72.68	71.21	48.21	57.82	40.50	37.60	53.41
	SkyCLIP-20	67.94	71.95	53.63	28.04	65.68	78.62	70.70	50.03	59.98	44.83	26.80	61.36
	SkyCLIP-30	69.08	72.15	52.44	27.77	66.40	79.67	70.77	50.19	59.99	46.17	30.80	64.32
	SkyCLIP-50	70.89	71.70	51.33	27.12	67.45	80.88	70.94	50.09	59.93	46.83	35.80	67.50
GeoLangBind-L-384	GeoLangBind-L-384	76.83	75.50	59.04	29.10	70.16	80.17	73.15	51.62	64.45	61.83	26.00	81.36

Table 3. Zero-shot comparison of various models on GeoBench datasets. Bold values indicate the best performance.

Model (Base version)	m-bigearthnet			m-so2sat		m-forestnet	
	Precision	Recall	F1	RGB	Sentinel-2 (5 bands)	RGB	Landsat-8 (5 bands)
SigLIP	45.34	12.36	16.82	12.88	-	8.16	-
RemoteCLIP	33.82	20.35	18.84	10.75	-	8.46	-
SkyCLIP-50	39.91	19.58	20.32	11.97	-	10.78	-
GeoLangBind-B-224	47.70	20.37	23.69	17.95	14.60	13.60	17.02

5. Experiments

5.1. Implementation details

GeoLangBind is trained using 8x NVIDIA A100 40GB GPUs for the base model and 8x NVIDIA A100 80GB GPUs for the large version. Our code is implemented based on OpenCLIP [21]. For the base version, we use a ViT-Base architecture (patch size 16), initialized with WebLI-pretrained weights [9]. The teacher models are ViT-Base-SigLIP (224), DINOv2-ViT-L/14, and ViT-Huge (ImageNet). The large version is based on SoViT-400m[4] (patch size 14, input size 384), with teacher models ViT-So400m-SigLIP (384), DINOv2-ViT-L/14, and ViT-Large (ImageNet). Training is conducted for 5 epochs on RGB datasets and 20 epochs on other modalities using AdamW (learning rate 5e-4, weight decay 1e-7). The batch size is 140 for the base model and 35 for the large model.

5.2. Performance comparison

We follow the evaluation protocols of SkyScript [50] and evaluate the models via zero-shot classification tasks. Specifically, the following eleven image scene classification datasets are used: AID [54], EuroSAT [19], fMoW [11], Million-AID [30], PatternNet [68], NWPU-RESISC45 [10], RSI-CB256 [25], as well as the fine-grained attribute classification datasets. In addition, we further compare these models on some datasets from the GEO-Bench suite [24]: m-bigearthnet [41], m-so2sat [71], and m-forestnet [22].

Overall zero-shot performance As shown in Tab. 2, our GeoLangBind models establish new state-of-the-art results

in zero-shot remote-sensing image classification. Compared with existing CLIP-based baselines such as CLIP-original, RemoteCLIP, CLIP-laion-RS, and the various SkyCLIP configurations, GeoLangBind-B-224 (Base) consistently outperforms all counterparts on scene-level tasks. Its average accuracy across the eight scene datasets is significantly higher than that of the other Base models. The Larger variant of GeoLangBind with higher resolution yields further accuracy gains. This highlights that both capacity and resolution play vital roles in capturing the spatial and spectral details crucial for complex remote-sensing tasks. On three fine-grained roof-attribute classification tasks, both GeoLangBind variants achieve strong accuracy on roof shape and surface, significantly outperforming other methods. Although the results on smoothness remain more modest, our approach still provides a clear improvement over existing approaches.

Zero-shot classification on GEO-Bench The zero-shot classification results on three GEO-Bench datasets are presented in Tab. 3. GeoLangBind demonstrates competitive or superior zero-shot multi-label classification performance compared to existing models on the m-bigearthnet dataset (43 classes). For the m-so2sat dataset (17 classes), when utilizing five spectral bands, the performance decreases slightly due to the significantly less multispectral training data compared to RGB. Note that competitive models cannot process 5-band images as input for zero-shot classification. On the m-forestnet dataset (12 classes), GeoLangBind substantially outperforms other models and shows remarkable accuracy improvements when using five spectral bands as input. Notably, the m-forestnet dataset uses Landsat-8 satellite data,

Table 4. Recall@1, recall@5, and recall@10 for cross-modal retrieval on three benchmark datasets. Bold values indicate the best performance.

Model	RSICD						RSITMD						UCM-caption					
	Image to Text			Text to Image			Image to Text			Text to Image			Image to Text			Text to Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
<i>RSICD, RSITMD, and/or UCM-caption seen in training</i>																		
AMFMN [61]	5.39	15.08	23.40	4.90	18.28	31.44	10.63	24.78	41.81	11.51	34.69	54.87	16.67	45.71	68.57	12.86	53.24	79.43
LW-MCR-u [60]	4.39	13.35	20.29	4.30	18.85	32.34	9.73	26.77	37.61	9.25	34.07	54.03	18.10	47.14	63.81	13.14	50.38	79.52
GaLR [62]	6.59	19.85	31.04	4.69	19.48	32.13	14.82	31.64	42.48	11.15	36.68	51.68	—	—	—	—	—	—
<i>RSICD, RSITMD, and/or UCM-caption not seen in training</i>																		
CLIP-original	6.59	20.68	31.75	3.62	14.28	23.63	10.18	30.31	42.04	8.31	24.96	39.03	37.62	78.10	89.52	28.12	64.99	77.19
CLIP-laion-RS	8.42	23.70	35.86	5.81	19.49	30.25	13.72	32.08	44.91	10.57	31.48	46.96	39.52	79.52	90.00	29.71	62.60	80.37
SkyCLIP-30	8.97	24.15	37.97	5.85	20.53	33.53	11.73	33.19	47.35	10.19	32.47	49.08	38.57	84.29	93.81	31.83	64.19	81.96
GeoLangBind-L	8.42	25.16	37.05	8.15	24.90	37.73	13.94	30.31	44.47	13.45	38.70	55.78	43.33	81.43	93.81	35.28	71.62	87.80

Table 5. Partial fine-tuning results on six segmentation tasks. All models are trained with a frozen backbone for 20 epochs. Reported numbers are mean intersection over union (mIoU). Missing values are due to the inability of the model to adapt to this domain.

Method	Models	Backbone	m-pv4ger-seg	m-nz-cattle	m-NeonTree	m-cashew-plant	m-SA-crop	m-chesapeake
Supervised	DeepLabv3 U-Net	ResNet101 ResNet101	93.4 94.1	67.6 80.5	53.9 56.6	48.6 46.6	30.4 29.9	62.1 70.8
	rand. init. Scale-MAE [38]	ViT-B ViT-L	81.7 83.5	74.1 76.5	51.7 51.0	32.4 —	29.0 —	47.1 61.0
	GFM [31]	Swin-B	92.0	75.0	51.1	—	—	63.8
Self-supervised Vision FMs	Cross-Scale MAE [43]	ViT-B	83.2	77.9	52.1	—	—	52.3
	CROMA [14]	ViT-B	—	—	—	30.1	31.4	—
	DOFA [57]	ViT-B ViT-L	94.5 95.0	81.4 81.8	58.8 59.4	51.5 56.9	33.0 32.1	65.3 66.3
Zero-shot FMs	RemoteCLIP [28]	ViT-L	94.1	79.8	56.2	37.9	31.3	59.5
	Skyscript-20 [50]	ViT-L	93.9	79.2	57.2	38.0	30.6	61.5
	GeoLangBind-B-224	ViT-B	94.3	80.7	56.1	53.6	31.8	62.7
	GeoLangBind-L-384	Large	94.7	82.2	59.0	45.8	32.6	63.0

which is not represented in our training dataset.

Cross-modal retrieval performance Tab. 4 compares the GeoLangBind model against recent methods on three benchmark caption datasets: RSICD, RSITMD, and UCM-caption. We report Recall@1, Recall@5, and Recall@10 for both image-to-text and text-to-image retrieval. Despite the inherent difficulty of remote-sensing cross-modal tasks, GeoLangBind-L consistently achieves competitive or higher recall scores than most competitive baselines, including CLIP-original, CLIP-laion-RS, and SkyCLIP-30.

Semantic segmentation performance Tab. 5 shows that GeoLangBind models work well across six segmentation tasks despite being zero-shot vision–language models. For all models, the backbone remains frozen while UPerNet [56] is trained as the segmentation head. GeoLangBind-B-224 achieves strong mIoU scores on m-pv4ger-seg and m-nz-cattle datasets. GeoLangBind-L-384 further improves results across all datasets, achieving competitive or better performance than vision foundation models. These results demonstrate GeoLangBind’s versatility and ability on fine-grained pixel-level understanding tasks.

5.3. Ablation studies

Weight merging Tab. 6 presents the results of our ablation experiments, evaluating the impact of weight merging across

Table 6. Ablation experiments of models on the AID zero-shot classification dataset.

Model	Distill.	Top-1 Accuracy	Top-5 Accuracy
SigLIP	—	65.95	96.20
GeoLangBind-B-RGB	✓	76.90	94.55
SigLIP+GeoLangBind-B-RGB	✓	77.70	95.50
SigLIP+GeoLangBind-B-RGB	✗	72.00	95.80
GeoLangBind-B-Others	✓	64.70	91.95
SigLIP+GeoLangBind-B-Others	✓	69.50	96.20
SigLIP+GeoLangBind-B-Others	✗	66.30	91.20
GeoLangBind-B-All	✓	71.60	93.65
GeoLangBind-B-All-Merging	✓	77.60	96.65

Table 7. Ablation experiments of different distillation methods on the segmentation datasets.

Model	MADOS	m-nz-cattle	m-NeonTree
SigLIP	57.2	77.5	52.3
GeoLangBind-B w/o MaKA	61.7	81.5	58.1
GeoLangBind-B w/ MaKA	62.3	82.2	59.0

different models on the AID zero-shot classification dataset. The baseline SigLIP model can be significantly improved by incorporating the GeoLangBind-B-RGB model, trained on the RGB part of GeoLangBind-2M. Merging SigLIP with GeoLangBind-B-RGB further enhances performance.

Table 8. Ratio for merging SigLIP and GeoLangBind-B-RGB.

m_1	AID (Top-1)	RSI-CB (Top-1)	Roof shape (Top-1)
0	65.95	38.01	41.17
0.1	44.25	27.33	37.67
0.3	56.75	31.87	38.00
0.5	74.65	41.80	44.33
0.7	77.40	45.36	53.67
0.9	77.70	47.03	57.83
1.0	76.90	46.72	58.50

The model trained with our MaKA module (for distillation) can largely outperform the non-distilled version. This highlights the role of distillation in improving weight fusion. GeoLangBind-B-Others, trained on non-RGB modalities performs lower than its RGB counterpart. When merged with SigLIP, the performance can be significantly improved.

GeoLangBind-B-All is the model trained with all the data modalities mixing together, i.e. using data mix instead of weight merging. When compared with the GeoLangBind-B-All-Merging, the performance is significantly lower. This comparison demonstrates that weight merging is more effective and flexible than direct mixing data modalities.

Knowledge distillation As presented in Tab. 7, we conducted ablation studies on the MADOS [23], m-nz-cattle [1], and m-NeonTree [52] datasets. GeoLangBind-B without MaKA represents a model trained using a baseline distillation method that does not incorporate modality-specific wavelengths as conditioning. The results demonstrate the significant effectiveness of the MaKA module, showing consistent performance improvements across all datasets.

Weight merging ratios In Tab. 8 and Tab. 9, a linear search is conducted to determine the optimal ratios for weight merging. Based on the results, we select 0.9 for m_1 to merge the weights of SigLIP and GeoLangBind-RGB. We choose 0.5 for merging weights of the intermediate weights and GeoLangBind-Others. Thus we set $m_1 = 0.9$ and $m_2 = 0.5$ to derive our final model weights. The findings demonstrate that merged weights consistently outperform individual models (when the ratio equals 0 or 1), validating the effectiveness of our weight-merging approach.

Table 9. Ratio for merging weights of GeoLangBind-B-RGB and GeoLangBind-B-Others.

m_2	AID (Top-1)	RSI-CB (Top-1)	Roof shape (Top-1)
0	77.70	47.03	57.83
0.1	78.10	47.31	57.50
0.3	78.55	47.91	55.83
0.5	78.60	49.17	52.67
0.7	76.45	49.92	52.00
0.9	76.35	43.98	63.33
1.0	69.20	34.80	59.67

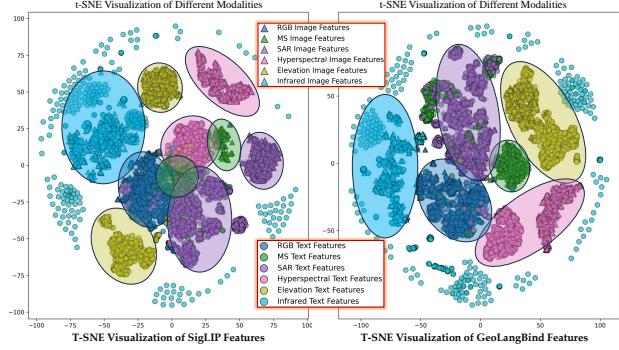


Figure 4. t-SNE visualization of feature distributions for the original SigLIP features (left) and GeoLangBind-L-384 model (right). Circle markers represent language features, while triangle markers represent image features.

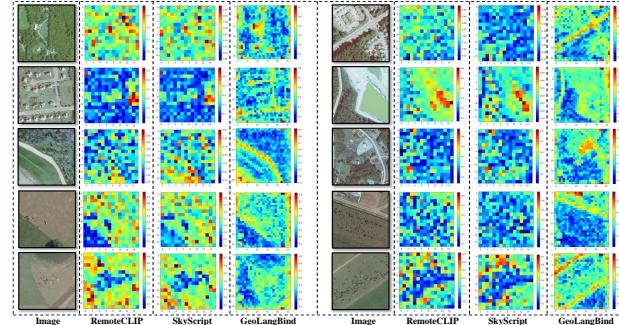


Figure 5. Visual comparison of deep features from RemoteCLIP, SkyScript, and GeoLangBind model.

5.4. Feature visualization and analysis

In the left side of Fig. 4, we present the t-SNE plot of SigLIP features. There is a significant separation between the language and image features for hyperspectral, SAR, elevation, and multispectral data modalities. This means that SigLIP struggles to align the representations of these data modalities. In contrast, features of GeoLangBind-L-384 demonstrate a much stronger feature alignment across all modalities, making it a powerful foundation for Earth observation tasks. In Fig. 5, we visually compare the feature heatmaps of RemoteCLIP, SkyScript, and GeoLangBind. We can see that GeoLangBind features are more spatially structured and preserve more object details.

6. Conclusion

We introduce GeoLangBind, a unified vision–language foundation model that bridges heterogeneous EO modalities through a shared language space. To achieve this, we construct GeoLangBind-2M, a large-scale image–text dataset containing six EO data modalities. GeoLangBind consists of a wavelength-aware encoder and a modality-aware knowl-

edge agglomeration module to enhance fine-grained image understanding. Additionally, progressive weight merging is proposed to scale the model training to multiple data modalities. Extensive experiments demonstrate the state-of-the-art performance of GeoLangBind across zero-shot classification, semantic segmentation, and cross-modal retrieval tasks.

References

- [1] Diab Abuaiadah and Alexander Switzer. Remote sensing dataset for detecting cows from high resolution aerial images, 2022. 8
- [2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Moncault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024. 3, 12, 16
- [3] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, pages 1–10, 2025. 5
- [4] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023. 6
- [5] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: An earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024. 1
- [6] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024. 1, 2
- [7] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 2
- [8] Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*, 2023. 1
- [9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 6
- [10] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 6, 12
- [11] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 6, 12
- [12] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1, 2
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [14] Anthony Fuller, Koreen Millard, and James R Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *arXiv preprint arXiv:2311.00566*, 2023. 2, 7
- [15] Anatol Garioud, Apolline De Wit, Marc Poupée, Marion Valette, Sébastien Giordano, and Boris Wattrelot. Flair# 2: textural and temporal information for semantic segmentation from multi-source optical imagery. *arXiv preprint arXiv:2305.14467*, 2023. 3, 12
- [16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 2
- [17] Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, et al. The EnMAP spaceborne imaging spectroscopy mission for Earth observation. *Remote Sensing*, 7(7):8830–8857, 2015. 3
- [18] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024. 2
- [19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. 6, 12
- [20] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2
- [21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 6
- [22] Jeremy Irvin, Hao Sheng, Neel Ramachandran, Sonja Johnson-Yu, Sharon Zhou, Kyle Story, Rose Rustowicz, Cooper Elsworth, Kemen Austin, and Andrew Y Ng. Forestnet: Classifying drivers of deforestation in indonesia using deep learning on satellite imagery. *arXiv preprint arXiv:2011.05479*, 2020. 6
- [23] Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and

- sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210: 39–54, 2024. 8, 12
- [24] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 12
- [25] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Hou, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large scale remote sensing image classification benchmark via crowdsource data. *arXiv preprint arXiv:1705.10450*, 2017. 6, 12
- [26] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024. 3
- [27] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1
- [28] Fan Liu, Delong Chen, Zhangqinyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 7
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [30] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021. 6, 12
- [31] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 2, 7
- [32] Fabio Montello, Edoardo Arnaudo, and Claudio Rossi. Mmflood: A multimodal dataset for flood delineation from satellite imagery. *IEEE Access*, 10:96774–96787, 2022. 3
- [33] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. 1
- [34] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org>. <https://www.openstreetmap.org>, 2017. 2
- [35] Maxime Oquab, Timothée Darzet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinnov2: Learning robust visual features without supervision, 2023. 5
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2
- [37] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12490–12500, 2024. 2
- [38] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 1, 2, 7
- [39] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. SSL4EO-L: Datasets and foundation models for Landsat imagery. *Advances in Neural Information Processing Systems*, 36:59787–59807, 2023. 1, 2
- [40] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. TorchGeo: Deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–12, Seattle, Washington, 2022. Association for Computing Machinery. 2
- [41] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019. 6
- [42] X. Sun, Y. Sun, and Z. Wang. High-resolution sar ship object detection dataset - 2.0. Online, 2020. 16
- [43] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-Scale MAE: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 7
- [44] Hugo Touvron, Thibaut Lavrill, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [46] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, et al. Hypersigma: Hyperspectral intelligence comprehension foundation model. *arXiv preprint arXiv:2406.11519*, 2024. 1
- [47] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. 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. 1

- [48] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023. 2
- [49] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenyang Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer, 2024. 2
- [50] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 2, 3, 6, 7, 12
- [51] Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation. *Ieee Access*, 8: 120234–120254, 2020. 16
- [52] Ben G Weinstein, Sarah J Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A Bohlman, and Ethan P White. A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network. *PLoS computational biology*, 17(7):e1009180, 2021. 8
- [53] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gonçalo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 2
- [54] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 1, 6, 12
- [55] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1
- [56] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 7
- [57] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *arXiv e-prints*, pages arXiv–2403, 2024. 1, 2, 4, 7
- [58] Zhitong Xiong, Yi Wang, Fahong Zhang, and Xiao Xiang Zhu. One for all: Toward unified foundation models for Earth vision. *arXiv preprint arXiv:2401.07527*, 2024. 1
- [59] Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. Earthnets: Empowering artificial intelligence for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 2024. 1
- [60] Zhiqiang Yuan, Wenkai Zhang, Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang, Kun Fu, and Xian Sun. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2021. 7
- [61] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*, 2022. 7
- [62] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 7
- [63] Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth System Science Data Discussions*, 2024:1–24, 2024. 3, 12
- [64] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 2, 5
- [65] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021. 16
- [66] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 16
- [67] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. 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. 2
- [68] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 6, 12
- [69] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 2
- [70] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 1
- [71] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuan-sheng Hua, Rong Huang, et al. So2sat lcz42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019. 6

A. Supplementary material

The supplementary material contains the following sections:

1. Details of the evaluation datasets;
 2. Details on the construction of GeoLangBind-2M;
 3. More visualization examples of GeoLangBind-2M;
 4. Ablation studies on the distillation loss terms;
 5. Confusion matrices of the zero-shot classification tasks;
 6. More visualization of the features;

A.1. Details of the evaluation datasets

We evaluate GeoLangBind on both datasets that have been established in previous works for direct comparison, as well as additional datasets that highlight the wide-ranging capabilities of the proposed model.

The following datasets are part of the SkyScript [50] evaluation framework, which we follow for direct comparison.

- SkyScript classification dataset [50]: contains 7,000 aerial RGB images with 70 classes and different objects than the SkyScript pretraining dataset.
 - AID [54]: contains 2,000 aerial RGB images and 30 image scene classes.
 - EuroSAT [19]: contains 27,000 Sentinel-2 images with 13 spectral bands and 10 image scene classes. We use the dedicated test split that contains 2,700 images.
 - fMoW [11]: contains 106,081 RGB aerial images with 62 image scene classes.
 - Million-AID [30]: contains 10,000 RGB aerial images of varying GSD with 51 image scene classes.
 - PatternNet [68]: contains 30,400 high-resolution RGB aerial images with 6–50 cm GSD and 38 image scene classes.
 - NWPU-RESISC45 [10]: contains 31,500 aerial RGB images with 0.2–30 m GSD and 45 image scene classes.
 - RSI-CB [25]: contains 24,747 aerial RGB images and 35 image scene classes.

Additionally, we choose the GEO-Bench [24] suite of datasets that cover a range of relevant remote sensing tasks across different domains, sensors, and geospatial locations. Tab. 10 contains a detailed overview of the number of samples, sensors, and target classes in the GEO-Bench datasets.

In addition, we conduct ablation experiments on the Marine Debris and Oil Spill (MADOS) [23] dataset. The MADOS dataset is a high-resolution multispectral Sentinel-2 dataset for marine pollution detection. Spanning 174 globally distributed scenes (2015–2022) with 1.5M annotated pixels, it covers diverse pollutants, sea surface features, and water-related classes. Unlike existing datasets, MADOS enables scalable, generalizable deep learning models for holistic marine pollution monitoring.

FLAIR2 (RGB)



Figure 6. Word cloud for the Flair2-RGB-caption dataset.

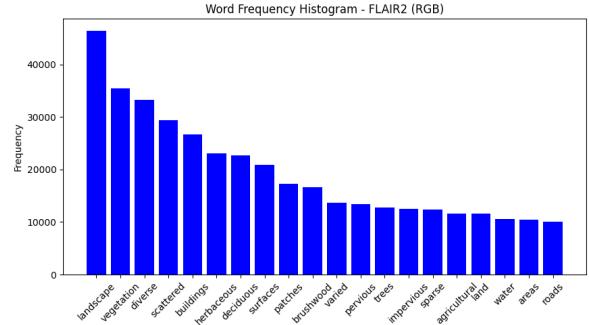


Figure 7. Word histogram for the Flair2-RGB-caption dataset.

A.2. Details on constructing GeoLangBind-2M

We construct our multimodal image–text dataset GeoLangBind-2M by assembling existing datasets and curating 8 new datasets. Details on the newly curated sub-datasets are as follows.

A.2.1. Flair2-RGB-caption

For the Flair2-RGB-caption dataset, originally designed for semantic segmentation land cover classification, we employed an approach similar to the ChatEarthNet [63] dataset construction. We select the training set of 61,711 images from FLAIR #2 [15] and derive contextual information from their corresponding pixel-level segmentation masks. This extracted semantic content, paired with the segmentation maps as visual context, serves as the input for generating comprehensive descriptions using the Pixtral 12B [2] language model.

Table 10. GEO-Bench dataset description.

Tasks	Name	Image Size	# Classes	Train	Val	Test	# Bands	Sensors
Classification	m-bigearthnet	120 x 120	43	20000	1000	1000	12	Sentinel-2
	m-so2sat	32 x 32	17	19992	986	986	18	Sentinel-2 + Sentinel-1
	m-forestnet	332 x 332	12	6464	989	993	6	Landsat-8
Segmentation	m-pv4ger-seg	320 x 320	2	3000	403	403	3	RGB
	m-chesapeake-landcover	256 x 256	7	3000	1000	1000	4	RGBN
		256 x 256	7	1350	400	50	13	Sentinel-2
	m-SA-crop-type	256 x 256	10	3000	1000	1000	13	Sentinel-2
	m-nz-cattle	500 x 500	2	524	66	65	3	RGB
	m-NeonTree	400 x 400	2	270	94	93	5	RGB + Hyperspectral + Elevation

Our semantic extraction process involves calculating the distribution percentages of various terrain categories within each image and formulating these statistics into structured contextual cues. We identify all land cover categories in each segmentation map, assigning distinctive color codes to represent different semantic classes. The designed prompt explicitly defines the correspondence between color codes and land cover categories, while also incorporating the calculated percentage coverage of each land/object type within the segmentation maps. This quantitative spatial information is then processed by the Pixtral 12B model to analyze distribution patterns and generate linguistically accurate descriptions.

With this semantic framework established, we craft a prompt for generating detailed image descriptions. The prompt template variables indicated in brackets represent dynamic content populated during the semantic processing phase. The exact prompt template used for Flair2-RGB-caption creation is shown below:

You are an AI visual assistant capable of describing a scene based on a segmentation map. The map represents different land cover types using specific colors. The legend is as follows:

- $[color_i]$ corresponds to $[label_i]$, occupying $[percent_i]\%$ of the image.

Do not mention colors, color coding, or technical details. Use the given land cover class names exclusively. Generate a brief yet natural description of the scene by extending the sentence: “The aerial image contains [presented_labels] land types.” Provide a concise summary of their spatial distribution.

In Fig. 6, we illustrate the word cloud for this dataset.

A.2.2. Flair2-Elevation-caption

The **Flair2-Elevation-Caption** dataset is constructed to generate detailed textual descriptions of terrain characteristics by leveraging both semantic segmentation maps and elevation maps. The dataset creation follows a structured approach, similar to the Flair2-RGB-caption dataset, ensuring that the generated captions effectively describe the topographical and land cover features of each image.

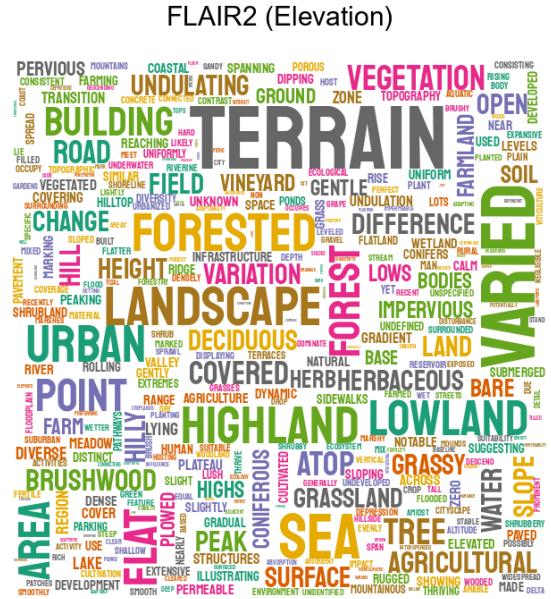


Figure 8. Word cloud for the Flair2-Elevation-caption dataset.

To begin, we select 61,711 images from the FLAIR #2 dataset, each accompanied by a segmentation map and an elevation map. The segmentation map provides land cover information, where each pixel corresponds to a specific land class, while the elevation map encodes height values representing terrain variation. Additionally, a class label mapping is used to associate numerical class IDs with their corresponding semantic labels, such as “forest,” “urban area,” or “water body.” This combination of datasets enables a comprehensive understanding of both land cover distribution and elevation patterns.

For each selected image, we analyze the elevation distribution by identifying the highest and lowest elevation values within the image patch. The corresponding land cover types at these extreme points are then determined using the segmentation map. This extraction process helps characterize the relationship between land cover and elevation, capturing how different landscapes correspond to varying altitude levels.

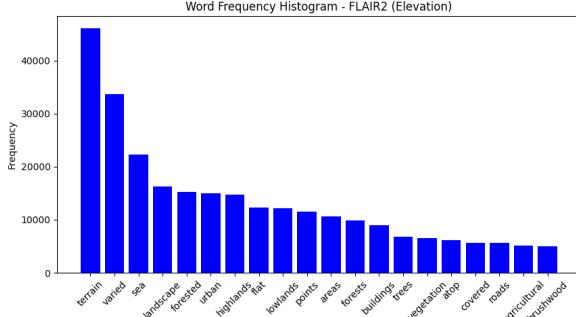


Figure 9. Histogram for the Flair2-elevation-caption dataset. Similarly to the other datasets of this appendix, FLAIR2-Elevation-caption has a focused vocabulary with many rare domain-specific terms. Its distribution deviates from a true power law due to frequent word repetition and an unusually long tail that is populated with remote sensing imagery terms. This mix of common and rare words could help models generalize by balancing domain-specific terms with diverse scene descriptors.

Using the extracted elevation and land cover information, we design a structured prompt to generate descriptive and natural textual captions. The prompt explicitly states the highest and lowest elevation values and their associated land cover types, followed by a request to summarize the overall terrain characteristics. The model is guided to describe whether the landscape is relatively flat or exhibits significant elevation changes, ensuring that the generated captions provide useful insights for downstream geospatial tasks.

To generate the final dataset, the structured prompts are input into a language model Pixtral 12B, which produces detailed, human-like descriptions of the elevation characteristics. These generated captions are then stored alongside their corresponding elevation and segmentation maps, forming the Flair2-Elevation-Caption dataset. The dataset offers rich, multimodal learning signals, allowing models to develop a deeper understanding of both land cover semantics and terrain variations.

This is an elevation map that indicates the height of each pixel. The highest areas, at an elevation of approximately [highest_height] meters, are [highest_class]. The lowest areas, at an elevation of approximately [lowest_height] meters, are [lowest_class].

Based on the provided context and elevation values, generate a concise and accurate description of the elevation map. Describe the image by briefly addressing:

- 1) The highest and lowest land cover types.
- 2) Whether the terrain is relatively flat or has significant elevation differences.

In Fig. 8, we demonstrate the word cloud to provide a better overview of the dataset. The histogram of the Flair2-Elevation-caption dataset is presented in Fig. 9.

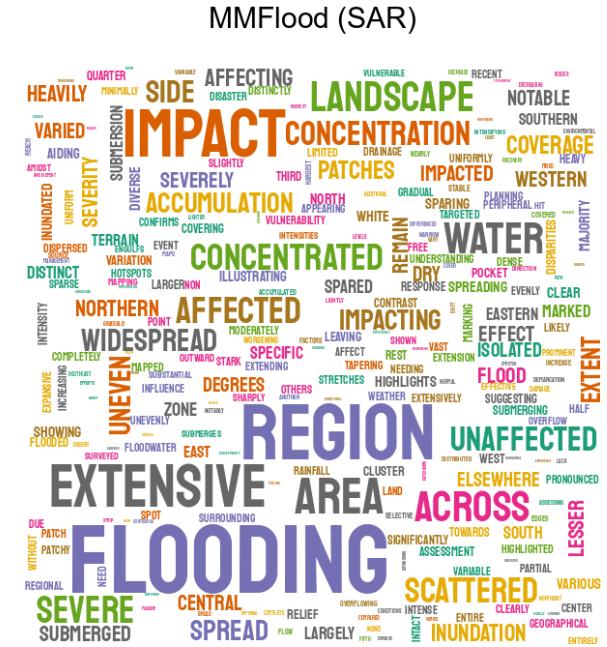


Figure 10. Word cloud for the MMflood-SAR-caption dataset.

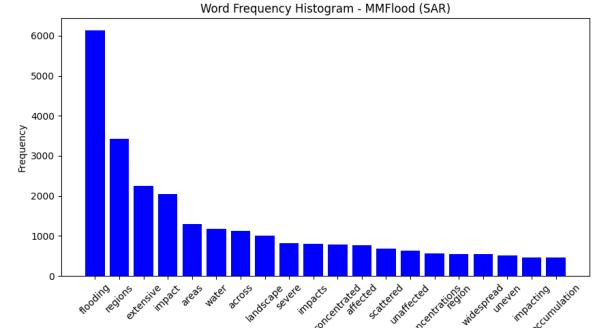


Figure 11. Word histogram for the MMflood-SAR-caption dataset.

A.2.3. MMflood-SAR-caption

The MMflood-SAR-caption dataset contains textual descriptions of flood maps generated using Pixtral 12B. In the following, we will describe the generation process in detail. Each flood map is represented as a binary mask, where white pixels indicate flooded areas, and black pixels represent non-flooded regions. To create meaningful textual descriptions, we analyze the extent of flooding and its spatial distribution before constructing structured prompts to guide the language model.

To quantify the severity of flooding, we calculate the total number of flooded pixels in each binary mask and express it as a percentage of the total image area. This value provides a direct measure of the flood extent, allowing the generated captions to capture variations in flood severity across differ-

ent samples. In addition to flood coverage, we assess spatial patterns by dividing the image into four quadrants: top, bottom, left, and right. By checking the presence of flooded pixels in each region, we determine whether the flooding is concentrated in specific areas or dispersed across the map.

Once these characteristics are extracted, we construct a structured prompt to guide Pixtral 12B in generating concise and informative descriptions. The prompt instructs the model to summarize the proportion of flooded areas and specify their locations within the image. To ensure clarity and brevity, the generated description is limited to fewer than 70 words, making it suitable for vision–language pretraining.

The prompts are then fed into Pixtral 12B, which generates natural language captions describing each flood map. These captions, stored alongside their corresponding binary masks, form the MMFlood dataset, enabling automated flood analysis, disaster monitoring, and geospatial AI research. By combining semantic flood mapping with natural language descriptions, this dataset enhances the ability of AI models to interpret and describe flood-affected regions in a human-readable format. We also provide the word cloud in Fig. 10 to show the detailed content of this dataset. In Fig. 11, we present the histogram of the words in the MMflood-SAR-caption dataset.

This is a flood map where areas marked with white pixels indicate flooded regions. The flooded area occupies approximately [flood_percentage]% of the entire map. Please analyze the flood map and provide insights into the affected areas.

You must generate a short description (less than 70 words) of the elevation image. First, describe the portion of floods; then introduce the location of the flooded areas.

After generating these detailed textual descriptions, we further refine the dataset by prompting Pixtral 12B to summarize each description into a concise 30-word caption. The prompts used in this summarizing process are as follows:

You are an AI assistant tasked with creating a concise 30-word caption that effectively summarizes the key points of the following content: [caption]

A.2.4. NLCD-hyper-caption

The caption generation process for hyperspectral images follows a methodology similar to the FLAIR #2 dataset, ensuring consistency in semantic information extraction and description generation. Hyperspectral images provide rich spectral information across multiple wavelengths, allowing for a more detailed classification of land cover types. However, the caption generation process remains aligned with the Flair2-RGB-Caption dataset, differing only in the specific prompt format used to guide the Pixtral 12B model. To cre-

NLCD (Hyperspectral)



Figure 12. Word cloud for the NLCD-hyper-caption dataset.

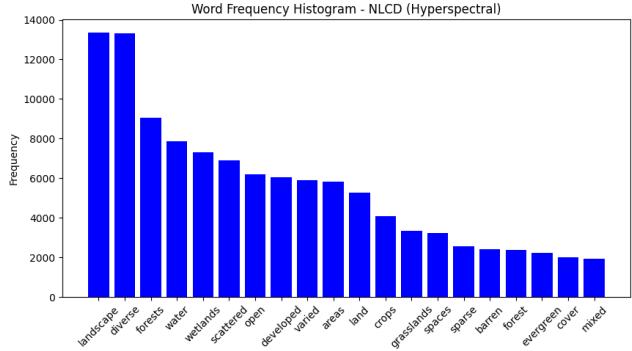


Figure 13. Word histogram for the NLCD-hyper-caption dataset.

ate captions, we first extract semantic information from the hyperspectral images using their corresponding segmentation maps. Each land cover type is identified and labeled, and its spatial distribution within the image is analyzed. Unlike conventional RGB imagery, where color-based segmentation might play a role, hyperspectral images rely on spectral signatures to differentiate land cover types, making this process crucial for accurate caption generation.

Once the land cover labels and their distributions are determined, we construct a structured prompt to generate natural language descriptions. The Pixtral 12B model is tasked with producing concise and human-readable descriptions of the scene. The prompt ensures that the generated captions focus solely on land cover types and their spatial distributions, avoiding any mention of technical details, color coding, or

hyperspectral data intricacies.

Following the generation of detailed textual descriptions, we further prompt Pixtral 12B to refine the descriptions into brief and natural captions. This step ensures that the dataset provides both long-form and short-form textual annotations, improving its usability for vision–language tasks, remote sensing analysis, and multimodal learning.

The prompt template used for hyperspectral image captioning is as follows:

You are an AI visual assistant capable of describing a scene based on a segmentation map. The map represents different land cover types using specific colors. The legend is as follows:

- $[color_i]$ corresponds to $[label_i]$, occupying $[percent_i]\%$ of the image.

Do not mention any colors, color coding, or technical details. Use the given class names. Only mention land cover types in the color legend.

Generate a brief and natural description of the scene by refining: "The hyperspectral image contains [presented_labels] land types." Provide a concise description of their spatial distributions (e.g., left, right, top, bottom).

To provide a clearer overview of this dataset, we showcase the word cloud in Fig. 12. We also showcase the histogram of the words in the NLCD-hyper-caption dataset in Fig. 13.

A.2.5. SAR-ship-caption and IR-ship-caption

For the SAR-ship-caption and IR-ship-caption subsets, we derive short captions from existing datasets to ensure high-quality textual descriptions. For IR-ship-caption, we utilize samples from the HIT-UAV dataset, specifically the data provided in EarthGPT [66], which is originally designed for visual question answering tasks. To generate descriptions for infrared images, we concatenate the corresponding question and answer pairs, transforming them into concise yet informative captions.

Similarly, for the SAR-ship-caption dataset, we generate captions for SAR images using three key datasets: HRSID [51], SSDD [65], and AIR-SARShip-2.0 [42]. These datasets provide extensive annotated SAR imagery, enabling the creation of meaningful captions. For further details on the captioning methodology, please refer to [66].

A.2.6. ChatEarthNet captions

Regarding the ChatEarthNet-S2-caption and ChatEarthNet-SAR-caption datasets, we reuse the detailed descriptions provided in the original ChatEarthNet dataset. Specifically, we use Pixtral 12B [2] to summarize these detailed descriptions into short image captions and generate question-answer pairs for visual question answering tasks. The word cloud for ChatEarthNet-caption datasets is presented in Fig. 14.

The word cloud is centered around the word 'GRASS' in large yellow font. Other prominent words include 'CROP', 'LAND', 'WATER', 'TREE', 'DEVELOPED', 'ACROSS', 'COVER', 'AREA', 'LANDSCAPE', and 'BARE'. The words are colored in shades of yellow, green, blue, and red, representing different semantic clusters or categories.

Figure 14. Word cloud for the ChatEarthNet-caption datasets.

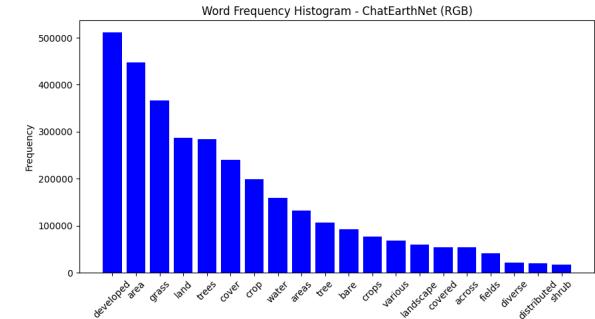


Figure 15. Word histogram for the ChatEarthNet-caption datasets.

To provide more insights into the dataset, we also show the histogram in Fig. 15. The prompt is as follows:

You are an AI assistant tasked with creating a concise 30-word caption that effectively summarizes the key points of the following content: [caption]

A.3. More dataset examples

These paired samples not only capture scene-level descriptions but also highlight fine-grained properties, such as object positions and fine-grained land cover types. Thereby the dataset enables the development of *foundation models* proficient in downstream performance for land cover mapping, flood monitoring, and object detection under challenging conditions (e.g., night or cloudy weather).

In Fig. 16 and Fig. 17, we provide more examples to

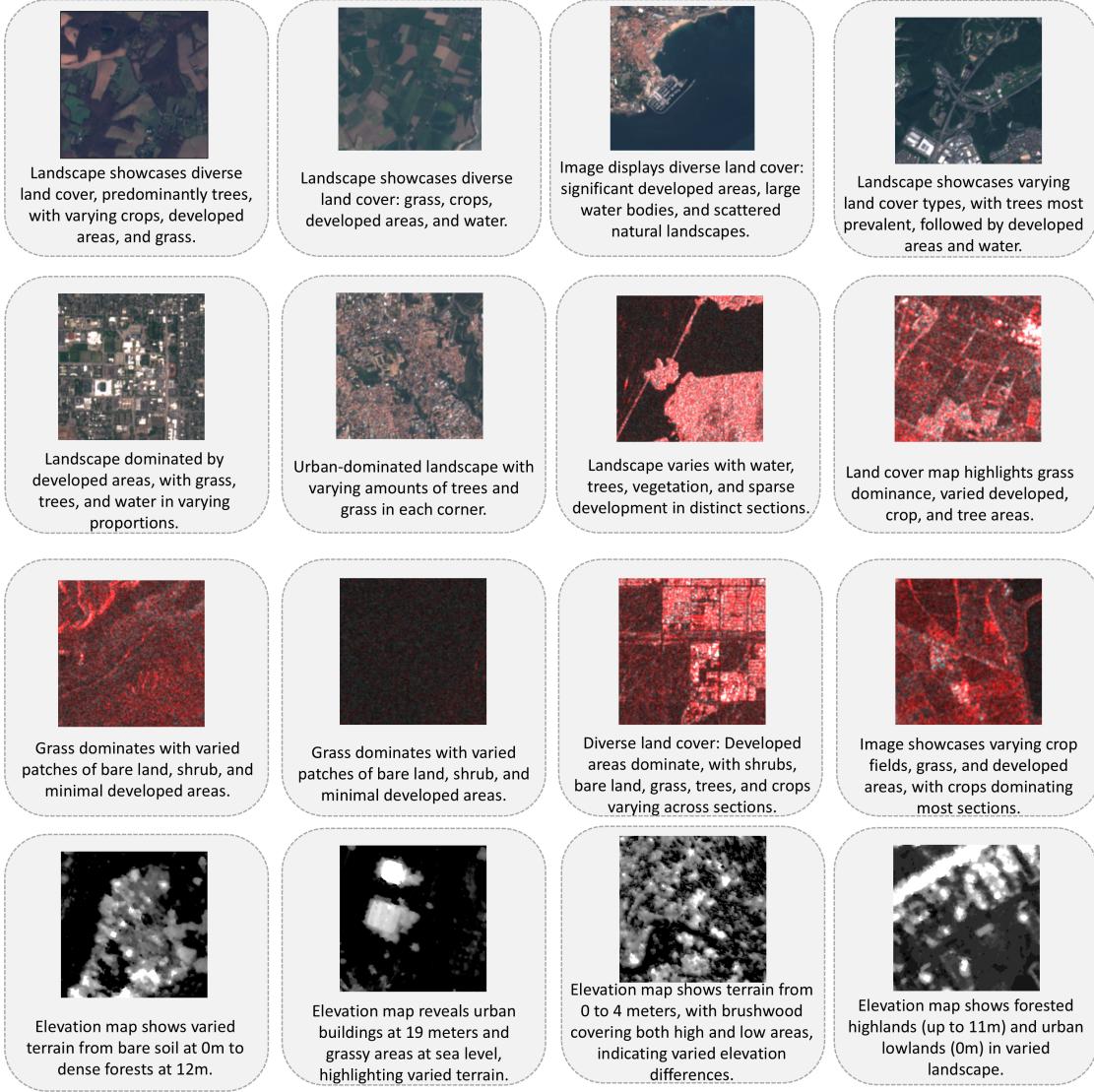


Figure 16. More examples of the image-text pair of GeoLangBind-2M, our multimodal image-text dataset.

provide a better overview of our dataset.

A.4. Ablation studies on distillation loss terms

Table 11. Distillation loss balancing weights vs. accuracy

α_s (SigLIP)	α_d (DinoV2)	α_v (ViT)	aid (Top-1)	eurosat (Top-1)
0	0	0	59.10	32.96
1.0	0.0	0.0	61.42	33.63
1.0	1.0	0.0	65.36	35.26
1.0	1.0	1.0	<u>68.75</u>	38.63
2.0	0.0	0.0	64.85	33.19
2.0	0.5	0.5	64.70	25.30
2.0	1.0	1.0	69.20	36.67

Tab. 11 presents an ablation study on the impact of different distillation loss balancing weights (α_s , α_d , α_v) on clas-

sification accuracy for the AID and EuroSAT datasets. The baseline model (no distillation) achieves 59.10% (AID) and 32.96% (EuroSAT). Adding SigLIP distillation ($\alpha_s = 1.0$) improves performance, while combining SigLIP and DinoV2 ($\alpha_s = 1.0$, $\alpha_d = 1.0$) further enhances results, especially for AID (65.36%). Introducing ViT distillation ($\alpha_v = 1.0$) boosts EuroSAT accuracy, peaking at 38.63% when $\alpha_s = \alpha_d = \alpha_v = 1.0$. In this work, we adopt the best overall configuration, $\alpha_s = 2.0$, $\alpha_d = 1.0$, $\alpha_v = 1.0$, achieves the highest AID accuracy (69.20%) and strong EuroSAT performance (36.67%). This demonstrates the effectiveness of multi-teacher distillation.

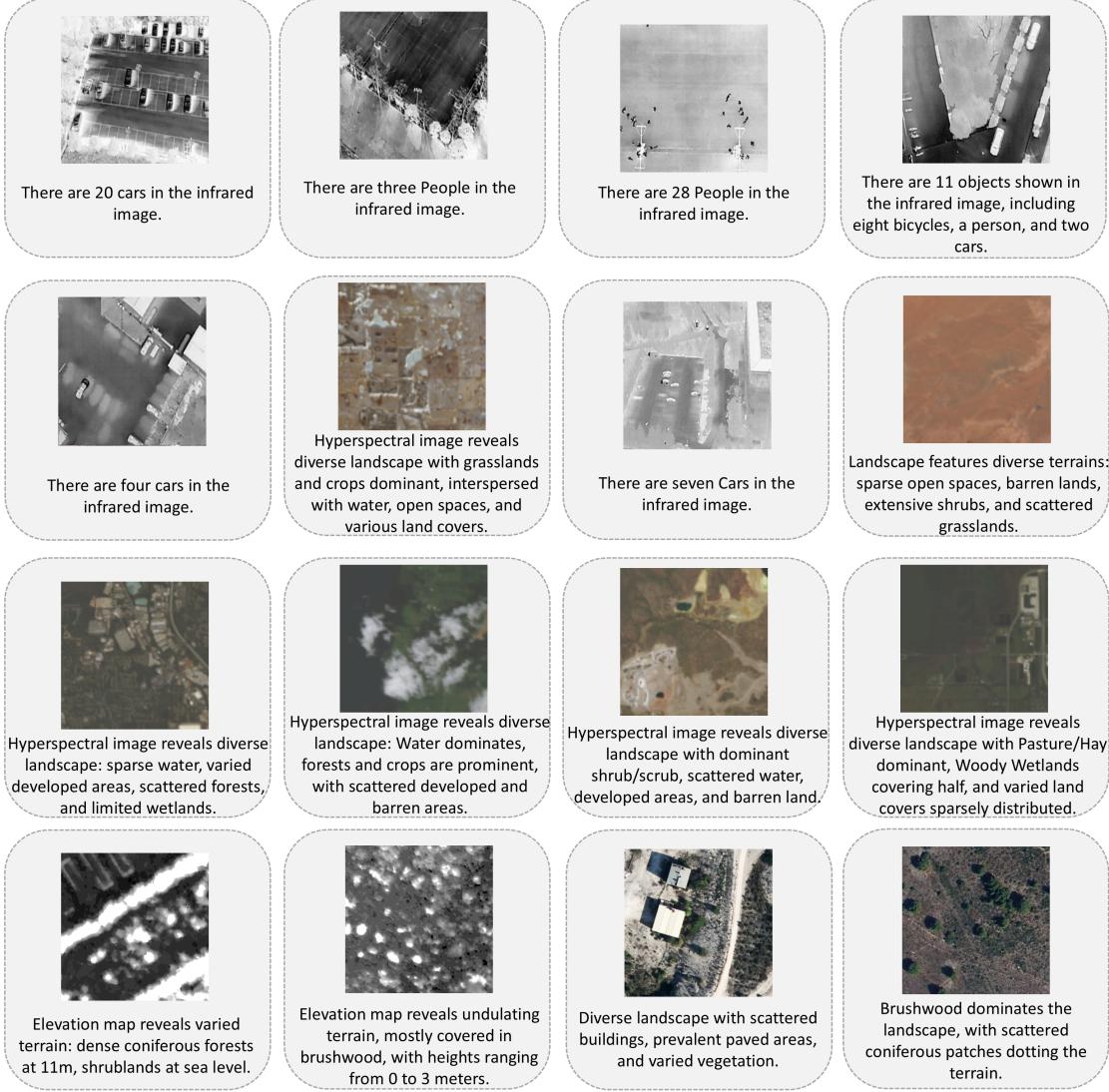


Figure 17. More examples of the image-text pair of GeoLangBind-2M, our multimodal image-text dataset.

A.5. Confusion matrices of zero-shot classification tasks

Figs. 18 to 25 present the confusion matrices of our GeoLangBind-L-384 model on all eight datasets for zero-shot classification. These confusion matrices reveal that common model mistakes largely occur on classes for which humans would also have trouble distinguishing, or for which classes have similar semantic meaning. For instance, the model commonly confuses “sparse residential” and “medium residential” (RESISC45), “annual crop” and “permanent crop” (EuroSAT), “aquaculture land” and “pond” (SkyScript), “chaparral” and “desert” (RESISC45), “mine” and “quarry” (SkyScript, Million-AID), “recreational facility” and “stadium” and “playground” (fMoW, AID), and

“ship” and “harbor” and “ferry terminal” (RESISC45, PatternNet).

In Fig. 26, we show the comparison of confusion matrices between SkyScript (ViT-L) and GeoLangBind-L-384 on fine-grained zero-shot classification tasks. It can be clearly seen that our model performs better.

A.6. More visualization of features

To better understand the representations learned by different models, Fig. 27 visualizes the feature maps extracted from RemoteCLIP (ViT-L), SkyScript (ViT-L), and GeoLangBind-L-384 across two datasets: m-chesapeake and m-nz-cattle from GEO-Bench. The m-chesapeake dataset primarily focuses on land cover classification, while the m-nz-cattle dataset captures cattle in remote sensing images. The feature

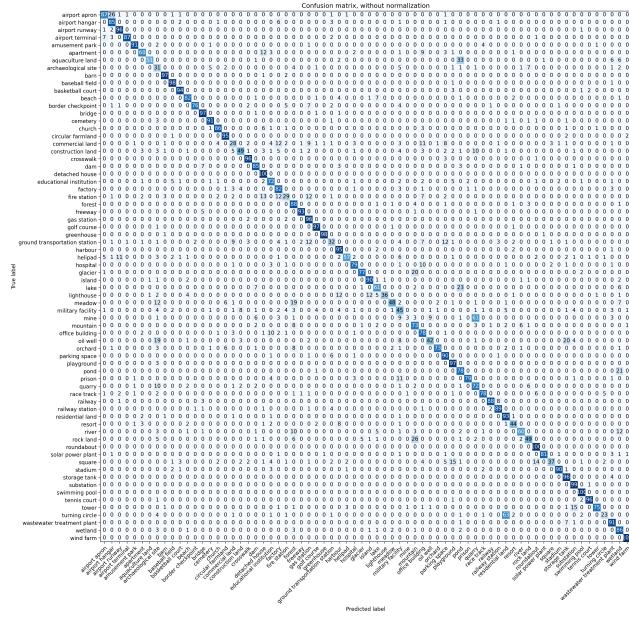


Figure 18. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on SkyScript dataset.

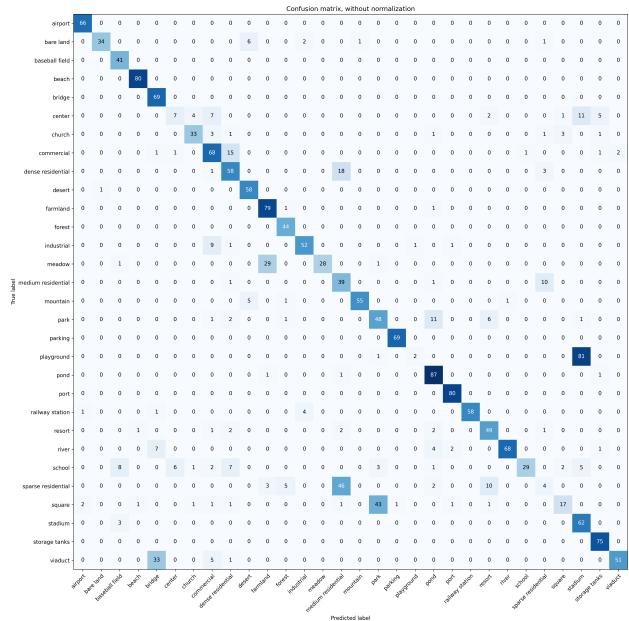


Figure 19. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on AID dataset.

visualization highlights key differences in spatial structure, consistency, and detail preservation across models.

Spatially structured features GeoLangBind produces features with higher spatial resolution, enabling finer spatial distinctions compared to RemoteCLIP and SkyScript. The feature maps exhibit well-defined structures that align with

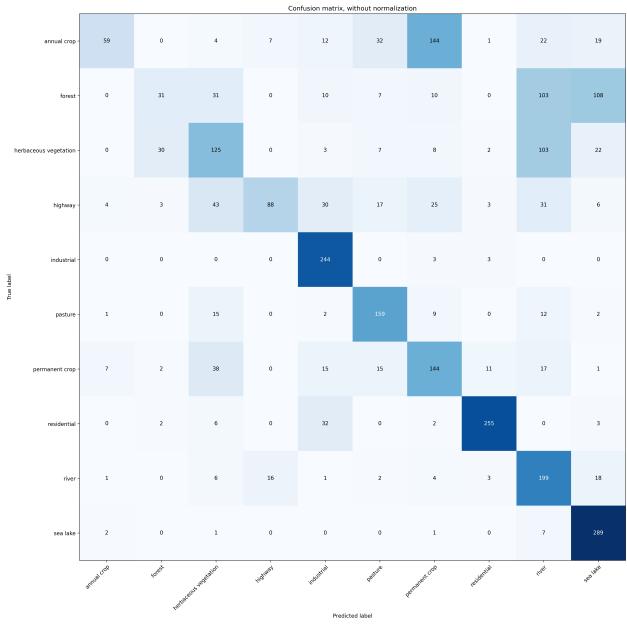


Figure 20. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on EuroSAT dataset.

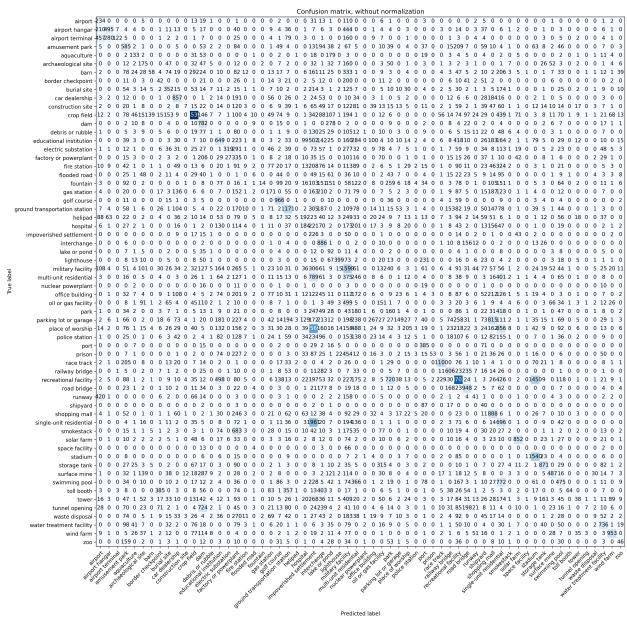


Figure 21. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on fMoW dataset.

meaningful spatial regions in the original images, such as roads, buildings, and vegetation.

Consistency across semantic classes The feature maps from GeoLangBind demonstrate greater consistency across images with similar semantic content. As observed, buildings and roads consistently exhibit high activations, while

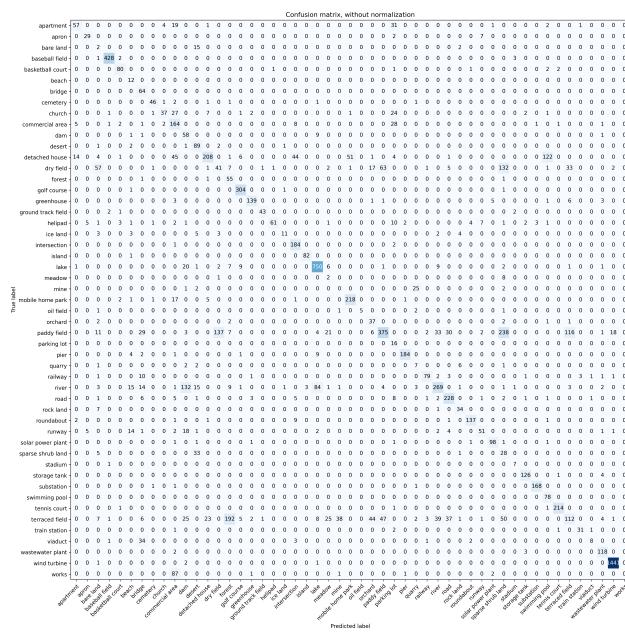


Figure 22. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on Million-AID dataset.

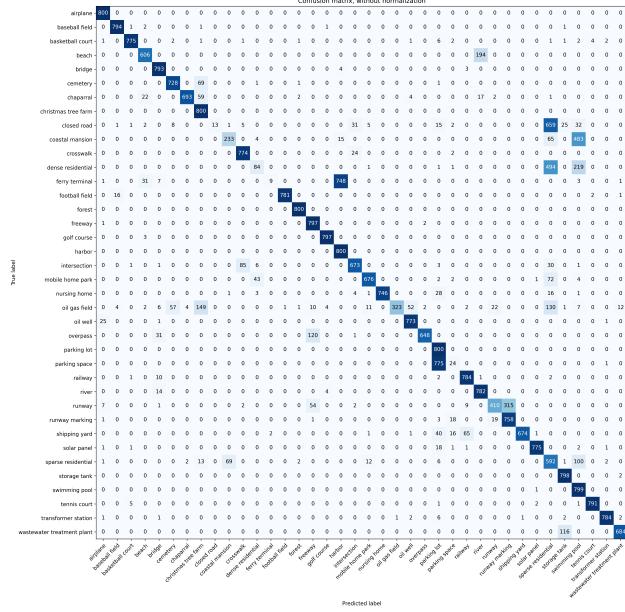


Figure 23. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on PatternNet dataset.

tree-covered areas show lower activations. This structured activation pattern is apparent across multiple samples, suggesting that GeoLangBind learns more stable and semantically aligned representations.

Preservation of details GeoLangBind retains finer details within its feature maps, capturing small-scale variations and

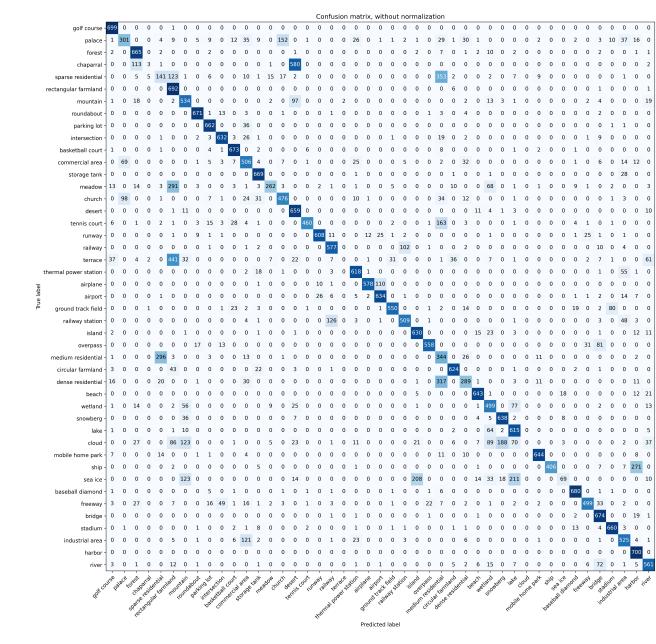


Figure 24. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on the RESISC45 dataset.

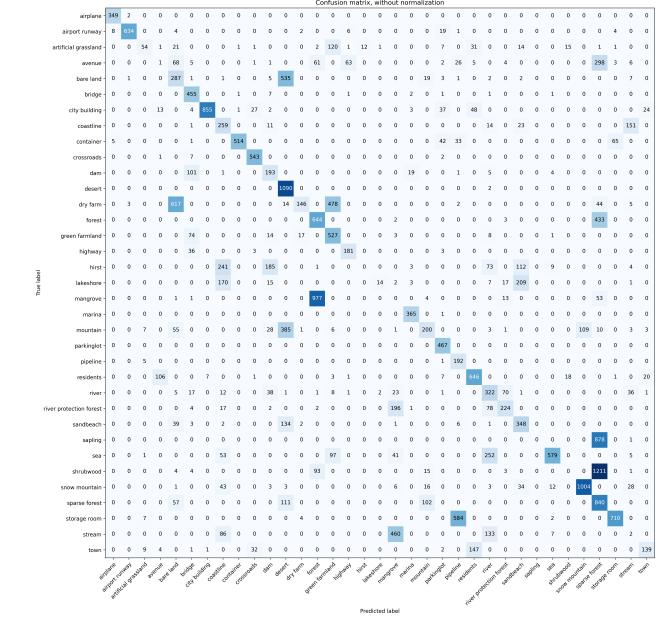


Figure 25. Confusion matrix of the zero-shot classification results of GeoLangBind-L-384 on RSICB dataset.

object boundaries more effectively than the baselines. This advantage is particularly evident in the m-nz-cattle dataset, where cattle are localized with sharper feature activations, as highlighted by the red bounding boxes. RemoteCLIP and SkyScript, on the other hand, exhibit more diffused and less distinct feature responses, making object-level distinctions

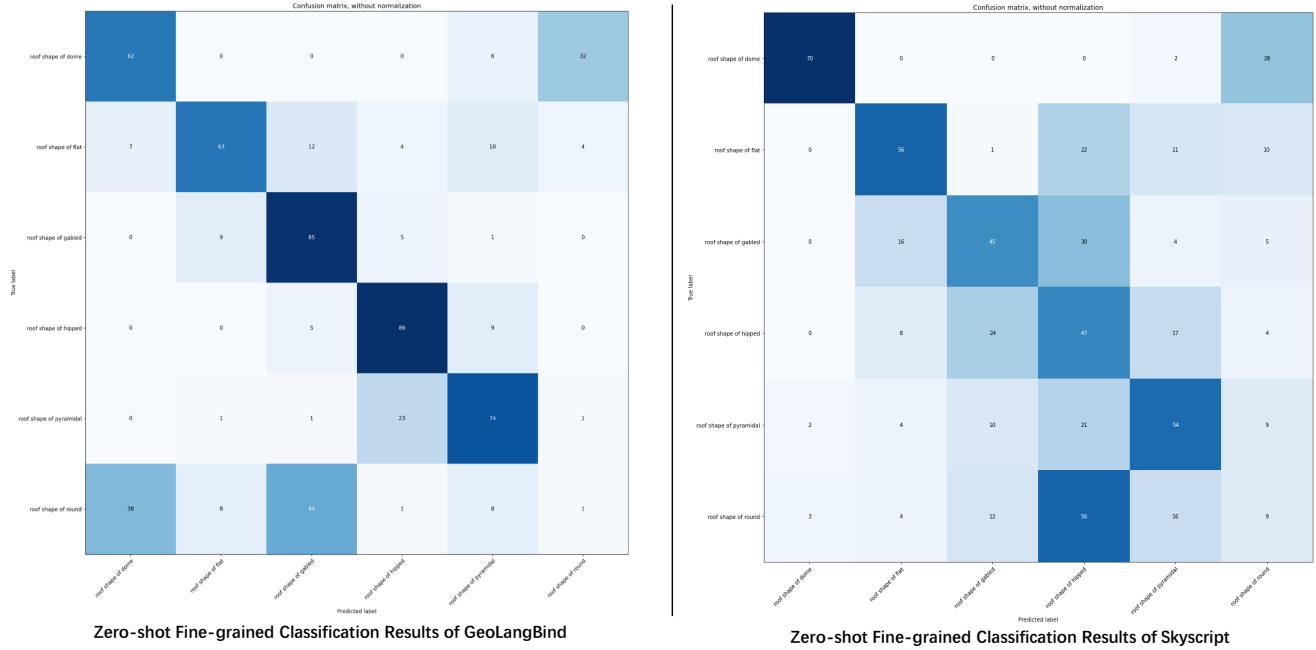


Figure 26. Comparison of confusion matrices between SkyScript-Large and GeoLangBind-L-384 on the fine-grained zero-shot classification task.

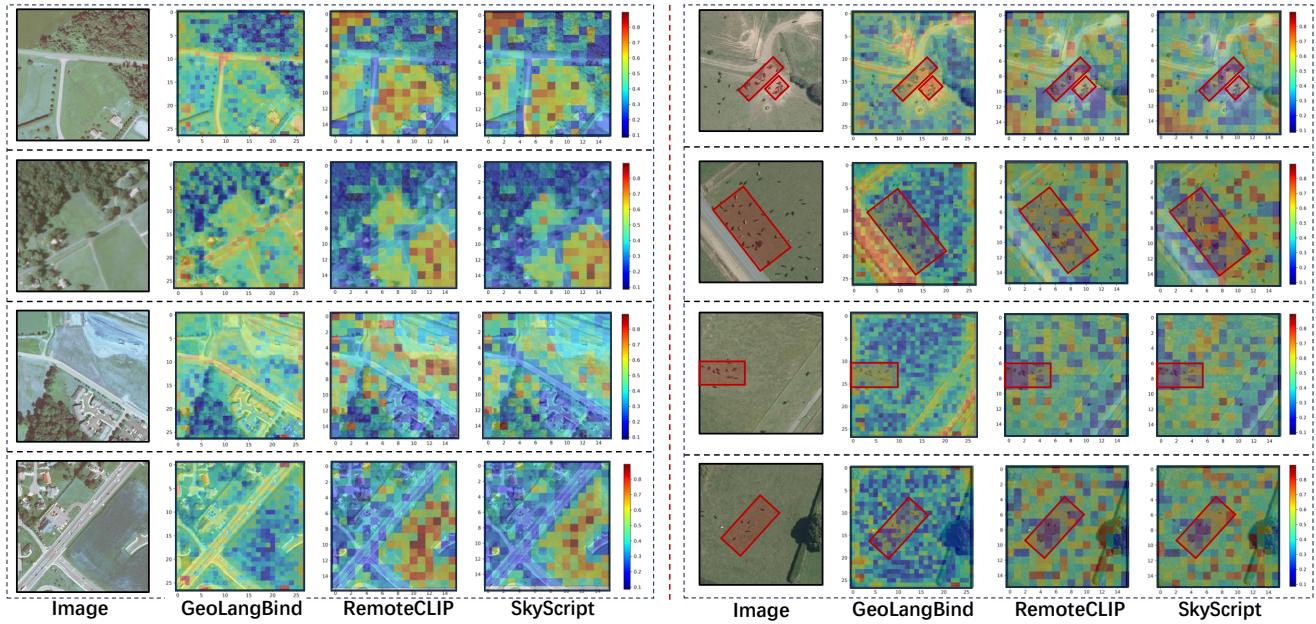


Figure 27. Visualization of features from RemoteCLIP (ViT-L), SkyScript (ViT-L), and GeoLangBind-L-384 on the m-chesapeake and m-nz-cattle datasets from GEO-Bench.

less precise.