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

Zhitong Xiong

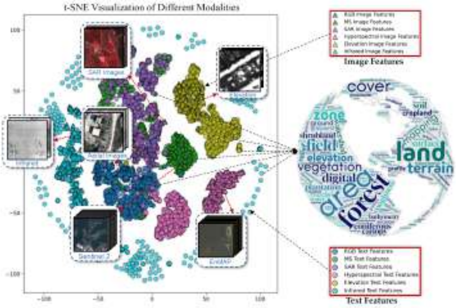
Technical University of Munich
Presented by: Zhitong Xiong

05.05.2025

The Challenge: Diverse Earth Observation (EO) Data

- Earth Observation (EO) data is vast and varied:
  - Optical (RGB)
  - Radar (SAR)
  - Multispectral (MSI)
  - Hyperspectral (HSI)
  - Infrared (IR)
  - Elevation (DEM)
- **Problem:** Heterogeneity makes unified analysis difficult.
- **Gap:** Prior models often focus on single modalities or lack robust language alignment for interpretation.

Goal: Bridge heterogeneous data via language



Bridging Heterogeneous Remote Sensing Data through Language Representations

Introducing GeoLangBind: A Unified Approach

- **Goal:** Develop a foundation model (**GeoLangBind**) to unify diverse EO modalities using language as a common bridge.
- **Core Idea:** Align different sensor data into a shared language embedding space.

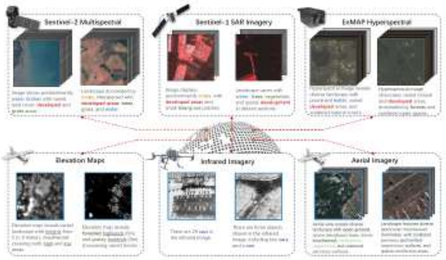
Introducing GeoLangBind: A Unified Approach

- **Goal:** Develop a foundation model (**GeoLangBind**) to unify diverse EO modalities using language as a common bridge.
- **Core Idea:** Align different sensor data into a shared language embedding space.
- **Key Contributions:**
    - New large-scale dataset: **GeoLangBind-2M**.
    - Novel Architecture: **Modality-aware Knowledge Agglomeration (MaKA)** & **Wavelength-aware Encoder**.
    - Efficient Training: **Progressive Weight Merging** strategy.
    - Result: State-of-the-art (SOTA) **zero-shot performance** across many EO tasks.

The GeoLangBind-2M Dataset

- **Need:** Lack of comprehensive, multimodal EO datasets with aligned text descriptions.
- **Solution:** Created **GeoLangBind-2M**.
- **Content:**
  - ~2 million image-text pairs.
  - Spans 6 key modalities:
  - RGB (Optical)
  - Sentinel-1 SAR (Radar)
  - Sentinel-2 MSI (Multispectral)
  - EnMAP HSI (Hyperspectral)
  - Infrared (Thermal)
  - Elevation (DEM)

Dataset Examples



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.

GeoLangBind Architecture Overview

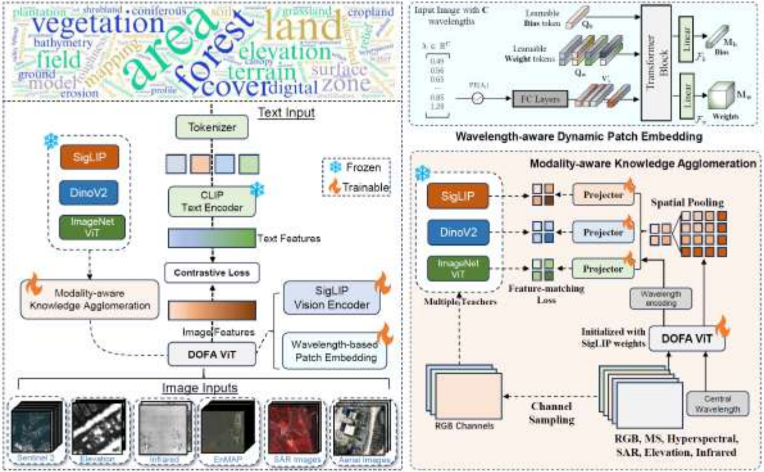


Figure: Overall model architecture of GeoLangBind. Key components highlighted.

Key Technique 1: Modality-aware Knowledge Agglomeration (MaKA)

- **Purpose:** Enhance fine-grained image understanding beyond simple contrastive alignment. Improve feature quality.

Key Technique 1: Modality-aware Knowledge Agglomeration (MaKA)

- **Purpose:** Enhance fine-grained image understanding beyond simple contrastive alignment. Improve feature quality.
- **How it works:**
  - Leverages multiple pre-trained 'teacher' models (SigLIP, DINOv2, ViT).
  - Uses input **wavelengths** as modality-specific condition to guide knowledge distillation.
  - Employs wavelength-aware prompts and conditional layer normalization.
  - Uses feature-matching loss (L_{match}) to align student (GeoLangBind) features with teacher features.
- This module is crucial for adapting powerful pre-trained models to the specifics of diverse EO data.
- (Refers to MaKA module shown in Figure 3 on previous slide).

Key Technique 2: Progressive Weight Merging

- **Problem Addressed:** Significant data imbalance (RGB vs. other modalities) can hinder joint training.

Key Technique 2: Progressive Weight Merging

- **Problem Addressed:** Significant data imbalance (RGB vs. other modalities) can hinder joint training.
- **Strategy:** More effective than simply mixing all data together.
  1. Train separate models: GeoLangBind on RGB (θ_{rgb}), GeoLangBind on other modalities (θ_{others}).
  2. Perform two-stage linear weight merging using task arithmetic:
     - Merge original SigLIP (θ_{siglip}) with RGB model:
       $\theta^* = (1 - m_1)\theta_{siglip} + m_1\theta_{rgb}$ (Optimal $m_1 = 0.9$)

Key Technique 2: Progressive Weight Merging

- **Problem Addressed:** Significant data imbalance (RGB vs. other modalities) can hinder joint training.
- **Strategy:** More effective than simply mixing all data together.
  1. Train separate models: GeoLangBind on RGB (θ_{rgb}), GeoLangBind on other modalities (θ_{others}).
  2. Perform two-stage linear weight merging using task arithmetic:
     - Merge original SigLIP (θ_{siglip}) with RGB model:
       $\theta^* = (1 - m_1)\theta_{siglip} + m_1\theta_{rgb}$ (Optimal $m_1 = 0.9$)
     - Merge intermediate model (θ^*) with 'Others' model:
       $\theta = (1 - m_2)\theta^* + m_2\theta_{others}$ (Optimal $m_2 = 0.5$)

Key Technique 2: Progressive Weight Merging

- **Problem Addressed:** Significant data imbalance (RGB vs. other modalities) can hinder joint training.
- **Strategy:** More effective than simply mixing all data together.
  1. Train separate models: GeoLangBind on RGB (θ_{rgb}), GeoLangBind on other modalities (θ_{others}).
  2. Perform two-stage linear weight merging using task arithmetic:
     - Merge original SigLIP (θ_{siglip}) with RGB model: $\theta^* = (1 - m_1)\theta_{siglip} + m_1\theta_{rgb}$ (Optimal $m_1 = 0.9$)
     - Merge intermediate model (θ^*) with 'Others' model: $\theta = (1 - m_2)\theta^* + m_2\theta_{others}$ (Optimal $m_2 = 0.5$)
- **Benefit:** Efficiently combines knowledge from different modality subsets and base models.

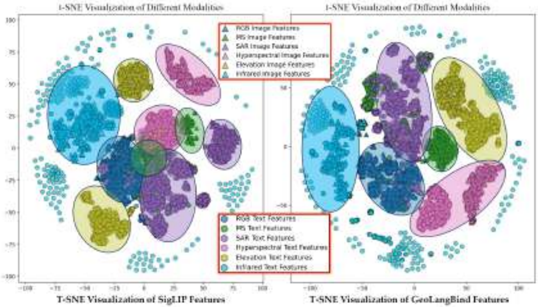
Result 1: Improved Cross-Modality Alignment

Finding: GeoLangBind significantly improves feature alignment across diverse EO modalities compared to standard vision-language models like SigLIP.

Explanation:

- Circles: Language features
- Triangles: Image features
- Colors: Different modalities
- Left (SigLIP): Gap between language and non-RGB modalities.
- Right (GeoLangBind): Tighter clustering, better alignment.

t-SNE Visualization Comparison



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.

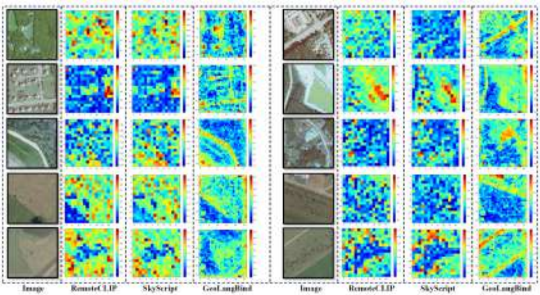
Result 2: Enhanced Feature Representation Quality

Finding: GeoLangBind produces features that capture finer spatial details and semantic structure compared to previous EO models.

Explanation:

- Heatmaps visualize model attention/features.
- GeoLangBind (left in pairs) shows sharper activation, better focus on objects (e.g., buildings, fields, cattle).
- Baselines (RemoteCLIP, SkyScript) show more diffuse activations.

Feature Heatmap Comparison



Visual comparison of deep features from RemoteCLIP, Skyscript, and GeoLangBind model.

Result 3: State-of-the-Art Zero-Shot Performance

- **Finding:** GeoLangBind achieves superior performance on downstream tasks *without* task-specific fine-tuning.
- **Evaluation Scope:** Tested on 23 diverse EO datasets.
- **Tasks:**
  - Scene Classification (e.g., EuroSAT, AID, RESISC45)
  - Fine-grained Classification (e.g., PatternNet)
  - Semantic Segmentation (e.g., Potsdam, Vaihingen via GeoBench)
  - Cross-Modal Retrieval (Image-Text and Text-Image)

Result 3: State-of-the-Art Zero-Shot Performance

- **Finding:** GeoLangBind achieves superior performance on downstream tasks *without* task-specific fine-tuning.
- **Evaluation Scope:** Tested on 23 diverse EO datasets.
- **Tasks:**
  - Scene Classification (e.g., EuroSAT, AID, RESISC45)
  - Fine-grained Classification (e.g., PatternNet)
  - Semantic Segmentation (e.g., Potsdam, Vaihingen via GeoBench)
  - Cross-Modal Retrieval (Image-Text and Text-Image)
- **Key Result Example (Scene Classification - Average Accuracy):**

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
	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**
Large	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	**76.83**	**75.50**	**59.04**	**29.10**	**70.16**	80.17	73.15	**51.62**	**64.45**	**61.83**	26.00	**81.36**

Conclusion

- **Summary:** GeoLangBind successfully unifies diverse EO data modalities using language as an effective bridge.

Conclusion

- **Summary:** GeoLangBind successfully unifies diverse EO data modalities using language as an effective bridge.
- **Key Innovations Recap:**
  - Large-scale multimodal **GeoLangBind-2M** dataset.
  - **Wavelength-aware encoder** for flexible input handling.
  - **MaKA** module for enhanced fine-grained features via distillation.
  - **Progressive weight merging** for efficient and effective training.

Conclusion

- **Summary:** GeoLangBind successfully unifies diverse EO data modalities using language as an effective bridge.
- **Key Innovations Recap:**
  - Large-scale multimodal **GeoLangBind-2M** dataset.
  - **Wavelength-aware encoder** for flexible input handling.
  - **MaKA** module for enhanced fine-grained features via distillation.
  - **Progressive weight merging** for efficient and effective training.
- **Impact:** Achieved State-of-the-Art **zero-shot performance**, providing a powerful and versatile tool for EO analysis.
- **Availability:** Dataset and pre-trained models will be made publicly available.

Thank You!

Questions?

Zhitong Xiong
zhitong.xiong@tum.de