

EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues

Sagar Soni^{*1}, Akshay Duhane^{*2}, Hiyam Debary^{*1}, Mustansar Fiaz^{*1}, Muhammad Akhtar Munir²
 Muhammad Sohail Danish², Paolo Fraccaro¹, Campbell D Watson¹, Levente J Klein¹
 Fahad Shahbaz Khan^{2,4}, Salman Khan^{2,3}

¹IBM Research ²Mohamed bin Zayed University of AI

³Australian National University ⁴Linköping University

Abstract

Automated analysis of vast Earth observation data via interactive Vision-Language Models (VLMs) can unlock new opportunities for environmental monitoring, disaster response, and resource management. Existing generic VLMs do not perform well on Remote Sensing data, while the recent Geo-spatial VLMs remain restricted to a fixed resolution and few sensor modalities. In this paper, we introduce EarthDial, a conversational assistant specifically designed for Earth Observation (EO) data, transforming complex, multi-sensory Earth observations into interactive, natural language dialogues. EarthDial supports multi-spectral, multi-temporal, and multi-resolution imagery, enabling a wide range of remote sensing tasks, including classification, detection, captioning, question answering, visual reasoning, and visual grounding. To achieve this, we introduce an extensive instruction tuning dataset comprising over 11.11M instruction pairs covering RGB, Synthetic Aperture Radar (SAR), and multispectral modalities such as Near-Infrared (NIR) and infrared. Furthermore, EarthDial handles bi-temporal and multi-temporal sequence analysis for applications like change detection. Our extensive experimental results on 44 downstream datasets demonstrate that EarthDial outperforms existing generic and domain-specific models, achieving better generalization across various EO tasks. Our source codes and pre-trained models are at <https://github.com/hiyamdebary/EarthDial>.

1. Introduction

Recent advancements in VLMs enable unified visual interpretation, where a single model can perform diverse tasks such as classification, localization, visual question-answering, counting, visual reasoning, and visual ground-

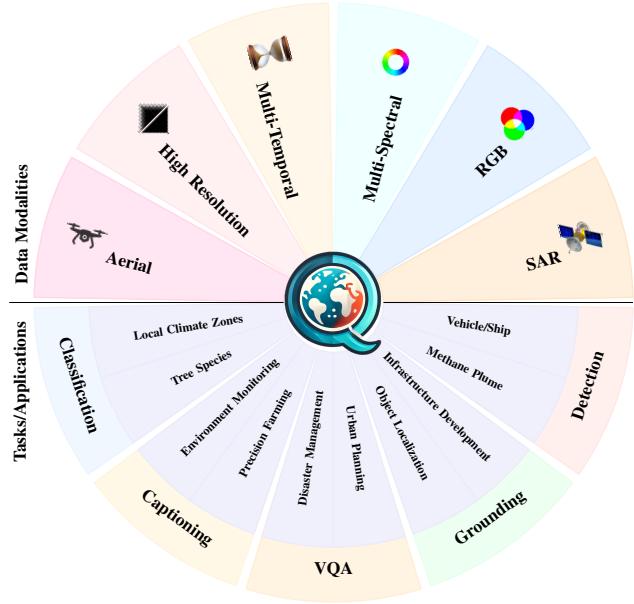


Figure 1. EarthDial is the first domain-specific VLM for earth observation data that can comprehensively interpret multi-sensor imagery. Specifically, our model covers visible RGB, SAR, multi-temporal, high-res satellite and aerial imagery available in varying spatial resolutions (*top half*). We develop the largest remote sensing image-text instruction dataset with over 11M samples. EarthDial can perform several multimodal understanding tasks: classification, detection, captioning, visual question-answering (VQA), and grounding (*bottom half*). This unlocks a number of downstream applications where EarthDial shows promising results.

ing [2, 10, 11, 39, 54, 56]. However, these generic VLMs do not scale well to Earth Observation (EO) data, which require specialized capabilities to handle the complex geospatial, spectral, and temporal dimensions of remote sensing (RS) data. Even state-of-the-art proprietary models like GPT-4V show low accuracies in domain-specific RS data [68], emphasizing the need for EO-specialized VLMs.

Recently, domain-specific VLMs have been developed

^{*}Equally contributing first authors.

Dataset	Type	# Samples	Tasks																		
			OS	MS	MT	MR	IC	RC	VQA	SC	MLSC	TSC	OD	VG	DA	BTCD	MTCD	M-TC	UHI	LCZ	TSC
RSICap [26]	Optical	2.6K	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
VHM [46]	Optical	180K	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
VRSBench [34]	Optical	205K	✓	✗	✗	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
GeoChat [28]	Optical	380K	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
MMRS [70]	Optical, SAR, IR	1.01M	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
SkyEye-968k [67]		0.97M	✓	✗	✗	✗	✓	✗	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
LHRS-Instruct [43]		81K	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗
Fit-RS [42]		1.8M	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗
EarthDial-Instruct (ours)		Optical, SAR, S2, IR, NAIP	11.11M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Overview of RS VLM datasets, their types, and supported tasks. OS: Open-Source, MS: Multi-Spectral, MT: Multi-Temporal, MR: Multi-Resolution. Tasks: IC: Image Captioning, RC: Region Captioning, VQA: Visual Question Answering, SC: Scene Classification, MLSC: Multi-Label Scene Classification, TSC: Temporal Scene Classification, OD: Object Detection, VG: Visual Grounding, DA: Disaster Assessment, BTCD: Bi-Temporal Change Detection, MTCD: Multi-Temporal Change Detection, M-TC: Multi-Task Conversation, MPD: Methane Plume Detection, UHI: Urban Heat Island, TSC: Tree Species Classification, LCZ: Local Climate Zones. Our dataset is 6× larger and covers diverse sensing modalities and provides annotations for a rich set of downstream tasks.

to understand EO data using generative multimodal models. RS-GPT fine-tuned the MiniGPT-4 model on 2.5k remote sensing instructions [26]. GeoChat represents an initial effort capable of performing image and region-level understanding as well as visual grounding in high-resolution remote sensing images [28]. Several approaches have focused on data scaling; e.g., LHRS-Bot uses crowd-sourced labels from OpenStreetMap to obtain 1.15M RS image-text pairs for multimodal alignment [43]. SkyEyeGPT curates a 968K sample instruction-following dataset for remote sensing conversational tasks [67]. However, these efforts are limited in high-resolution image processing and do not support multi-spectral, multi-temporal analysis.

In this work, we present EarthDial, aiming to develop the first unified model that can cohesively process multi-resolution, multi-spectral, and multi-temporal remote sensing imagery to unlock numerous downstream tasks. In all the above modalities, EarthDial can perform diverse tasks, including classification, object/change detection, question-answering, image and region captioning, and visual grounding. To achieve this goal, we propose the most extensive instruction tuning dataset to date, with over 11.11M instructions covering visible imagery with varying resolutions, SAR, and multispectral modalities, including NIR and infrared. Furthermore, EarthDial can handle bi-temporal and multi-temporal sequence analysis for applications such as change detection and temporal sequence classification.

Our main contributions are as follows:

- We propose EarthDial, a conversational VLM capable of processing multi-spectral, multi-temporal, and multi-resolution remote sensing imagery with natural text queries, addressing a wide range of EO tasks.
- We introduce the largest instruction tuning dataset for remote sensing, comprising over 11.11M instruction pairs across various modalities, enhancing the model’s understanding and generalization capabilities.
- Experimental results demonstrate that EarthDial performs well in comparison to existing domain-specific VLMs,

achieving higher accuracies and better generalization across 44 downstream EO tasks.

2. Related Work

Generic Vision-Language Models (VLMs): The development of generic VLMs like VisualGPT [9], BLIP [32], Flamingo [3], and Kosmos [27] has enabled advancements in showcasing multi-modal understanding by aligning visual and language data for diverse applications. Devoted efforts from researchers enabled the VLMs to perform a range of tasks, for example, OCR to diagram and infographics understanding to video analysis within a unified model [11, 30, 35, 56]. The continuous progress enabled the alignment of additional modalities such as audio, video, 3D point clouds [19, 55], audio-video grounding tasks [16], 3D visual grounding [61] as well as in fields such as LIDAR [36, 73] and robotics [23, 59]. Nevertheless, they struggle with the unique contextual complexities of remote sensing (RS) data, which requires specialized alignment for geospatial, spectral, and temporal information.

Geospatial VLMs: Recently, various efforts have been devoted towards domain-specific RS vision-language understanding to address the limitations of general VLMs [26, 28, 38, 43, 67, 70]. RemoteCLIP [38] employs contrastive learning over RS image-text pairs, illustrating the zero-shot classification and image-text retrieval capabilities. RS-GPT [26] fine-tuned over EVA-CLIP and Vicuna LLM demonstrates image captioning and VQA abilities while struggling over detection and visual grounding tasks. GeoChat [28], LHRS-Bot [43] and SkyEyeGPT [67] extend their capabilities to resolve multiple tasks such as region-level understanding as well as visual grounding in high-resolution RS images. However, these models do not cover multi-spectral and temporal modalities. More recently, EarthGPT [70] introduces the MMRS1M dataset to integrate optical, SAR, and infrared modalities, advancing multi-sensor RS comprehension. However, they do not cover other multi-spectral inputs and lack generalization to

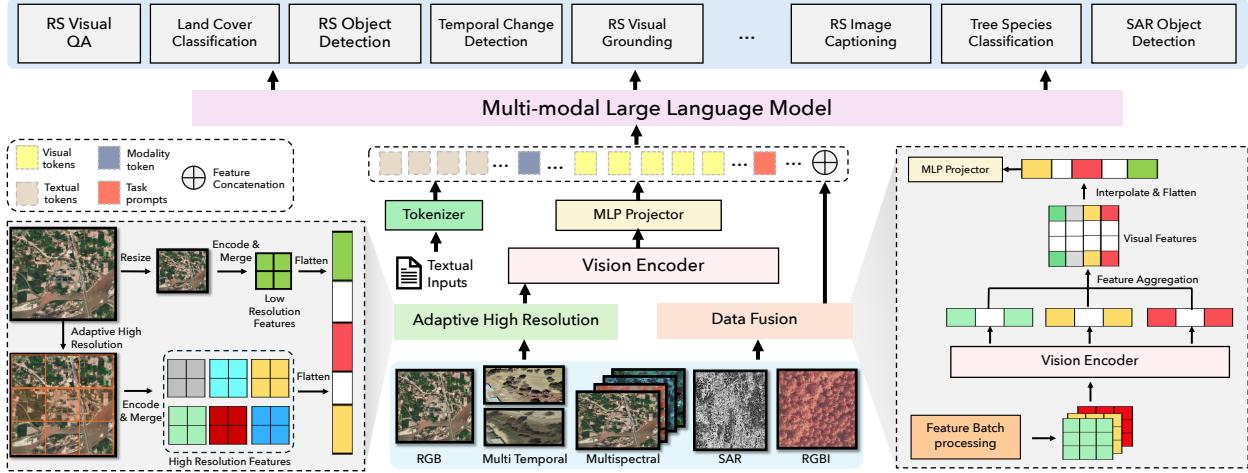


Figure 2. EarthDial Architecture: The model can take a diverse set of inputs ranging from RGB to multi-spectral and time-series images. Multi-resolution inputs are converted to tokens based on an adaptive high-resolution block [11] that includes both local and global features. The multi-channel inputs (multi-spectral/temporal) are converted to tokens via the data fusion block, which aggregates features across all channels. The resulting visual tokens are mapped to LLM input space using MLP projectors and concatenated with the textual inputs. We use special task and modality tokens to distinguish between several input modalities and downstream tasks (Table 2). The LLM is trained with multimodal inputs to perform a number of downstream tasks, ranging from VQA to detection, grounding and change detection.

multi-temporal and varying resolution inputs.

RS Instruction Datasets: A large number of instruction-following datasets have been introduced to train the RS-VLMs effectively. For example, GeoChat-Instruct [28], SkyEye-968k [67], and RS5M [71] provide extensive RS image-text pairs, supporting instruction-based VLM training on optical data. These existing datasets limit the model’s capabilities across different sensor modalities. On the other hand, MLLMs [70], provide image-text pairs from optical, SAR, and infrared images. Regardless of the large scale, the aforementioned datasets often lack diverse RS applications across various modalities, including multi-resolution, multi-spectral, and multi-temporal RS sensor data. The recently introduced dataset aims to enhance content richness with factual and deceptive questions, enabling VLMs to address a wider range of tasks in the RS domain [46]. However, there is no large-scale unified instruction-following dataset that can encapsulate the distinctive contextual complexities of diverse RS applications across different modalities, multi-resolution, multi-spectral, and multi-temporal RS sensor data. Our work is an effort to bridge this gap with an 11M instruction set to seamlessly integrate multiple earth observation modalities covering diverse spectral and with time-series imaging data for diverse RS applications.

3. EarthDial

Our goal is to develop a domain-specialized VLM that can handle complex geospatial, spectral, and temporal dimensions unique to RS imagery. As described above, the existing general and geospatial VLMs lack in understanding

high-resolution, multi-spectral, and multi-temporal RS imagery. To bridge this gap, we propose a comprehensive large-scale instruction tuning dataset for RS domain with over 11M instruction, covering diverse resolutions and geographical locations. Building on this dataset, we propose EarthDial, the first unified model capable of processing multi-resolution, multi-spectral, and multi-temporal RS data across a variety of tasks, from classification and visual grounding to change detection.

EarthDial leverages state-of-the-art vision-language models (VLMs) for natural images and provides a multi-stage finetuning recipe to progressively expand model capabilities. Our model architecture builds on InternVL [11, 12] with specific modifications to enable multi-spectral and multi-temporal processing (Sec. 3.1). We proposed a three-stage model training process to enhance the model’s capabilities across multiresolution, multispectral, and multi-temporal datasets. We first pretrain with remote sensing datasets, focusing on adapting state-of-the-art VLMs for EO-specific dialogues. In the next stage, output of pretrained encoders and LLM are adapted using RGB and temporal imagery for downstream tasks. Finally, an extended finetuning stage is specifically designed to improve its performance with multispectral and synthetic aperture radar (SAR) datasets to broadly cover additional applications. Next, we explain our model design in detail.

3.1. Model architecture

As illustrated in Fig. 2, EarthDial consists of three trainable components: a visual encoder, an MLP layer projec-

tor, and a large language model (LLM). Our model is relatively lightweight with only 4B parameters compared to the existing natural geospatial VLMs. The model is designed in such a way that it can take multi-resolution, multi-spectral, and time series datasets to generate various RS dialogues. As our visual encoder, we use InternViT-300M [12], a lightweight vision model distilled from the larger 6B InternViT that demonstrates strong visual encoding capability. Since our design goal is to have an efficient model, we use the Phi-3-mini pre-trained LLM [1]. To connect the visual encoder with the LLM, a simple MLP is used as the connector block to map visual tokens to the LLM space. As explained in Fig. 3, we tune the parameters of these three blocks systematically in different stages of training.

Furthermore, the model incorporates two key modules, Adaptive High Resolution and Data Fusion, that are crucial in applying EarthDial to different resolution inputs as well as multi-spectral and multi-temporal RS data.

Adaptive High Resolution: In remote sensing, images come in various sizes and resolutions, particularly high-resolution imagery where resizing for model can lead to the loss of critical pixel details. To address this, we adopt a dynamic resolution input strategy inspired by InternVL 1.5 [11], which enhances the model’s ability to capture fine-grained details. The approach dynamically selects an optimal aspect ratio from a set of pre-defined ratios, dividing the image into 448×448 -pixel tiles and creating a thumbnail for global context to help the model understand the overall scene. Depending on the input resolution, 1 to 12 tiles can be created during training and upto 40 during inference. This approach minimizes aspect ratio distortion and accommodates varying resolutions during training and evaluation.

Data Fusion: The data fusion module in EarthDial is designed to improve the model’s capability to process multi-temporal, multi-spectral, synthetic aperture radar (SAR), and RGBI/hyper-spectral datasets. For multi-spectral inputs of any type, it operates by iteratively processing three channels of data at a time, which are passed through the Vision Transformer (ViT), to extract features for each channel. The extracted features are then aggregated and reduced in size using bilinear interpolation via the AnyRes block [30] to ensure efficient handling of multi-spectral inputs. The AnyRes block splits the inputs into patches, encodes them, and uses bilinear interpolation to reduce tokens per patch, thus enabling the processing of multi-spectral inputs. These reduced visual embeddings are then concatenated with input corresponding text embeddings. The final step involves fusing these combined visual and textual features, which are passed to the LLM for further contextual processing. This fusion strategy allows EarthDial to integrate visual data from various modalities together with textual descriptions, improving its performance on complex RS tasks.

For RGB temporal images, we first pass each image

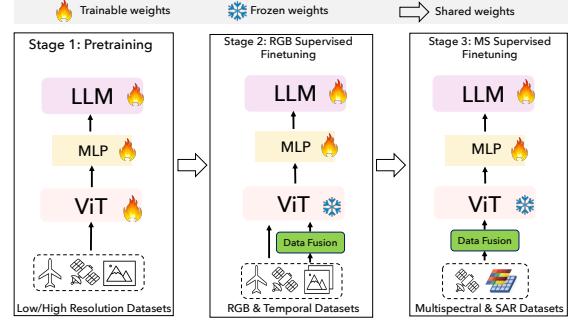


Figure 3. EarthDial training Strategy for different RS modalities. We first pretrain with RGB imagery of different resolutions to achieve better alignment. Thereafter only LLM and projectors are trained on RGB and temporal inputs. We then expand the model’s capability to multi-spectral and SAR imagery in Stage 3.

through the ViT to extract visual tokens, then stack and concatenate these tokens before passing the combined representation to the LLM. Next, we explain our model training.

3.2. Stage-wise Model Training

EarthDial is initialized with pretrained backbones explained in Sec. 3.1. It is then trained in three distinct stages, designed to systematically expand model’s capabilities.

Stage 1 - RS Conversational Pretraining: The main goal of pretraining is to enhance the vision encoder’s ability to learn visual features from various satellite imagery sources, such as Sentinel-2, Landsat 8, Sentinel-1, and aerial images, and to generate descriptions for remotely sensed objects and scenes. During this phase, we use instruction datasets described earlier data section, specifically the Satlas and Skyscript datasets, which contain 7.6M image-text pairs for different types of remote sensing imagery.

In the pretraining stage, the objective is to establish alignment between multiple sensor modalities in the RS domain with their corresponding natural language text descriptions. This is typically achieved using large datasets of image-text pairs, such as image and captions. The model learns to predict the corresponding text for a given image in an autoregressive manner, optimizing performance through a standard cross-entropy loss function. This process helps the model improve its ability to generate accurate text descriptions from remote sensing visual inputs. As shown in Fig. 3, the architecture during pretraining is deliberately kept simple, only comprising of the vision encoder, MLP and the LLM, without any data fusion. At the pre-training stage, model is not introduced temporal and multi-spectral datasets, to keep the task relatively simple and initially learn strong representations from single-image RGB datasets from our proposed EarthDial-Instruct dataset.

At this stage, all the learnable components were trained to ensure proper alignment of the RS imagery. For pretraining the EarthDial model, we utilized 8 NVIDIA A100-80G

GPUs. The training process employed an initial learning rate of $4e-5$, optimized using a cosine learning rate scheduler. The key hyperparameters included the use of thumbnails to capture local features, an adaptive patch size ranging from 1 to 6 for capturing more detailed high-level features, a batch size of 2, and a maximum sequence length of 4096 tokens. Additionally, a weight decay of 0.01 was applied to regularize the model.

Stage 2: RS RGB and Temporal finetuning: In the second stage of model training, we carefully selected high-quality instruction sets with prompts aimed at teaching the model to better understand user commands and successfully complete desired tasks on different satellite imagery. This approach allows Large Language Models (LLMs) to generalize to unseen tasks, improving zero-shot performance across a variety of remote sensing tasks.

During this stage, we utilized previously pretrained model encoders and fine-tuned the MLP and LLM layers to perform visual instruction tuning for diverse remote sensing tasks, such as image captioning, classification, detection tasks (e.g., grounding, referring, identification), visual question answering (VQA), and temporal change detection.

For multi-temporal images, we applied data fusion techniques, as explained in the model overview section, before passing the data to the feature extractor. Additionally, we used all the RGB and time-series datasets described in Tab. 2 during this phase of model fine-tuning.

Stage 3: RS Multispectral and SAR Finetuning: In the third stage of model training, we extended the model’s capabilities to work with multispectral, high-resolution RGBI, and SAR images. This was achieved by introducing a data fusion module, as explained earlier, to handle imagery with more than three or fewer than three spectral bands. The goal was to enable the model to learn from multispectral and SAR data for various remote sensing tasks.

We utilized the pretrained weights from the previous stage 2 and fine-tuned the MLP and LLM layers, while keeping the ViT parameters frozen from Stage 1. The same ViT+MLP+LLM architecture was employed, with the addition of the data fusion module (Sec. 3.1) to integrate information from multiple spectral channels. This training stage allows the model to handle a broader range of tasks, including land cover classification, species detection, Methane plume detection, UHI and SAR-based ship detection. Training hyperparameter details are given in Table 3.

4. EarthDial-Instruct Dataset

Pretraining Instruction Data: With EarthDial, our primary aim is to improve generalization performance on diverse downstream tasks, covering a wide range of modalities, multi-resolution, and multi-temporal data. Therefore, we curate high-quality pre-train question-answer (QA) instruction pairs from SkyScript [58] and SatlasPretrain [6]

Stages	Datasets	Number of QA pairs	Token format
Stage 1	NAIP	3,000,113	[hr.rgb_.05]
	Sentinel-2	2,749,511	[s2.rgb_.10]
	Landsat	1,671,437	[l8.rgb_.30]
	SkyScript	249,855	[s2.rgb_.10]
Stage 2	Classification	565,853	[hr.rgb_.05]
	Detection	22,624	[hr.rgb_.05]
	Visual Grounding	17,845	[hr.rgb_.05]
	Caption	202,530	[caption][hr.rgb_.05]
	VQA	630,768	[hr.rgb_.05]
	Change Detection	64,631	[changedet][hr.rgb.temp_.05]
	Disaster assessment	37,563	[hr.rgb.temp_.05]
Stage 3	Geochat	308,861	[hr.rgb_.05]
	Sentinel -1	1,668,043	[s1.vh_.10]
	Local Climate Zones	765,591	[s2.ms_.30]
	Tree Species	38,527	[treeclassify][hr.rgb_.05]
	Methane Plume	6,849	[hyper.rgb_.3]
	Urban Heat Island	1,296	[uhi][l8.ms_.30]

Table 2. Summary of the number of QA instruction pairs used during each stage, the image sources, and token formats.

Stages	QA pairs	Epoch	Train-time	Gpus	Batchsize	Grad.	Accu.	Lr	Decay	Adaptive patch sizes
2	1.8M	1	4 (hours)	4A100 80G	2	64	$4e-5$	0.05		1 to 6
3	2.4M	1	6 (hours)	4A100 80G	2	64	$4e-5$	0.05		1 to 6

Table 3. Details of stages, hyperparams, & training duration

data, which includes Sentinel-2 (S2), Sentinel-1 (SAR), NAIP, and Landsat imagery along with labels. We choose InternLM-XComposer2 [22] for generating instructions using labels. Our data curation process involves filtering to ensure data quality. First, we filter out samples with sparse labels (<3). Second, we apply luminance and coverage-based filtering to remove cloudy and low spatial coverage images. Third, we prompt the InternLM-XComposer2 to generate QA instruction pairs based on the key attributes (points, polygons, object category, and position) specified in the inputs and labels. The details can be found in the supplementary material. The curated instruction stats across different imagery sources are presented in Tab. 2.

Downstream Tasks Image-text Instruction: While pre-training focuses on enhancing generalization capabilities, we also need task-specific fine-tuning with diverse data types to improve downstream performance. To handle this, we carefully curate a large number of instruction-following datasets that cover ten diverse downstream tasks (e.g., scene classification, object detection, visual question answering, image captioning, change detection, Methane plume detection, tree species classification, local climate zones, urban heat islands, and disaster assessment), six visual modalities (Optical, SAR, S2, Infrared, NIR, and Hyperspectral), and two visual temporal modalities (Optical and SAR). Details are in the supplementary.

Scene Classification: We construct scene classification instructions with nine standard scene classifications, one multilabel scene classification (BigEarthNet [50]), and one temporal scene classification (FMoW [17]). We limit the sequences to 4 images to handle multitemporal scene classification. We also use local climate zones (LCZ) [75] and

Model	AID [60] (RGB)	UCMerced [63] (RGB)	WHU-19 [18] (RGB)	BigEarthNet [50] (RGB)	xBD Set 1 [24] (Temporal)	fMoW [17] (Temporal)
GPT-4o	74.73	88.76	91.14	49	67.95	21.43
InternVL-8B [12]	60.4	58.23	79.3	19.73	51.44	21.04
GeoChat [28]	72.03	84.43	80.09	20.35	53.32	59.2
EarthDial	88.76	92.42	96.21	68.82	96.37	70.03

Table 4. Comparison of classification accuracy across various datasets. EarthDial indicates a significant improvement in classification accuracy over other existing generic and specialized VLMs.

Method	BigEarthNet [50] (MS)	SoSAT-LCZ42 [75] (MS)	TreeSatAI [4] (RGBI)	Ship Dataset (SAR Imagery)				
				Small	Medium	Large	Single	Multiple
GPT-4o	49	15.53	16.73	0.70	0.90	3.20	1.20	0
EarthDial	69.94	60.72	56.61	12.14	26.02	35.56	26.03	6.06

Table 5. Performance evaluation of EarthDial across diverse modalities for multi-class classification and referred object detection tasks on SAR imagery. For MS modality, EarthDial achieves an average 32.5% improvement in classification accuracy compared to GPT-4o. For the RGBI modality, EarthDial achieves 40.2% higher accuracy than GPT-4o. A similar trend is observed for SAR imagery, where EarthDial delivers higher mAP@0.5, even when detecting multiple objects, highlighting the advantages of leveraging multi-modal inputs.

TreeSatAI-Time-Series [4] datasets to determine the LCZs and botanical tree species, respectively.

Object Detection: We curate instruction-following tasks utilizing three tags i.e., *refer*, *identify*, and *grounding* to perform region-level captioning, referring expressions, and grounded description for object detection datasets from various remote imaging modalities like optical, SAR, and infrared. We include visual grounding [52, 66] datasets for region-level captioning. Following [28], we compute the key attributes in the image such as the object’s category, bounding box, color, relative position, and relative size present in the detection dataset. We present the box as $[x_{min}, y_{min}, x_{max}, y_{max}, \theta]$. Here, (x_{min}, y_{min}) denotes the top left corner point while (x_{max}, y_{max}) presents the bottom right corner of the bounding box. The angle θ represents the rotation angle of the bounding box.

Visual Question Answering (VQA) & Image Captioning: We create VQA and image captioning instructions by including six VQA and five image captioning datasets.

Change Detection: We integrate three binary change detection datasets and one multitemporal (MUDS [62]) dataset. The original MUDS dataset has masks. Thus, to generate the instructions, we manually analyzed the images and masks and generated five captions for each sequence.

Methane Plume Detection: For the Methane plume detection, we utilize the STARCOPI [49] dataset which presents labeled hyperspectral (groundtruth mask and emission rate) data. We conversationally prompt three questions; (i) is there any Methane plume present in the input, (ii) what is the location of the plume, and (iii) what is its emission rate?

Urban Heat Island (UHI): We compute land surface temperature (LST) and normalized difference vegetation index (NDVI) maps from S2 and Landsat imagery. Based on the data, we prompt to classify the underlying region into cooler, mildly hot, and extremely hot regions. We also generate instructions about what underlying land use caused the

temperature and how to mitigate it.

Disaster Assessment: We use the xBD [24] dataset for building disaster assessment by utilizing bitemporal pre- and post-disaster images. We also include the QuakeSet [8] dataset and prompt to determine if an earthquake occurred between the bi-temporal SAR images and its magnitude.

5. Experiments

Here, we discuss the experimental results of the proposed EarthDial across a diverse set of applications, including RGB, multispectral, SAR, infrared, and thermal imagery. Our evaluation covers various tasks such as scene classification, referred object detection, region captioning, grounding descriptions, VQA, image captioning, change detection, and methane plume detection.

Scene classification: For zero-shot evaluation, we compare EarthDial with RGB datasets in Tab. 4, whereas BigEarthNet (RGB), xBD Set 1, and fMoW are supervised. We also compare multi-spectral (MS) datasets (BigEarthNet, SoSAT-LCZ42), and the RGB-Infrared TreeSatAI dataset as shown in Tab. 5. We notice that EarthDial shows consistent performance gain against existing generic and specialized VLMs. Moreover, EarthDial outperforms temporal scene classification FMoW as well as over xBD test-set 1 for disaster assessment [24] as in Tab. 4. Further results are in a supplementary document.

Object Detection: Following [28], we address three sub-tasks: referred object detection, region captioning, and grounding description. We consider existing generic VLMs like GPT-4o, InternVL2-4B while specialized GeoChat for the comparison. As existing InternVL2 doesn’t provide the rotated bounding boxes, for fair comparison, we finetune the InternVL2 on GeoChat-Instruct and compared it with our EarthDial. Table 6, 7, and 8 depict our EarthDial is a clear winner and consistently outperforms the all other compared VLMs by a large margin. Especially on grounding

Model	GeoChat-Instruct [28]					NWPU VHR-10 [13] (ZS)					Swimming Pool Dataset (ZS)					Urban Tree Crown Detection [65] (ZS)				
	Small	Medium	Large	Single	Multiple	Small	Medium	Large	Single	Multiple	Small	Medium	Large	Single	Multiple	Small	Medium	Large	Single	Multiple
GeoChat [28]	2.9	13.6	21.7	16	4.3	2.5	3.2	14.7	13.23	1.9	-	3.1	7.3	1.2	0.6	-	1.8	8.9	2.9	3.1
InternVL2-4B [12]	6.3	24.37	37.38	24.96	11.72	7.1	12.68	25.48	22.96	8.1	0.6	6.6	8.9	4.5	0.865	-	3.17	13.41	5.9	3.1
InternVL2-8B [12]	7.20	23.76	31.99	25.77	9.30	4.26	11.85	20.72	21.66	5.86	0.3	4.7	18.27	7.6	0.514	0.6	3.99	17.1	7.9	3.94
EarthDial	11.43	31.76	39.07	34.29	13.41	11.66	14.21	23.12	25.37	8.9	1.04	7.4	24.90	8.4	1.04	1.1	7.01	25.67	11.13	6.7

Table 6. Comparison of our EarthDial for referred object detection tasks across various datasets. Small, medium, and large denote the object size, while single and multiple denote the number of objects. Here, ZS means zero-shot evaluation.

Model ZS=Zero-Shot	GeoChat-Instruct [28]			HIT UAV [53] (ZS)			NWPU VHR 10 [7] (ZS)			SAR-Ship Dataset [57]			SRSDD-v1.0 [29]			Swimming Pool (ZS)			UCAS AOD [74] (ZS)			Urban Tree Crown [65] (ZS)		
	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT			
GPT-4o	9.41	7.6	8.02	10.96	9.02	8.23	17.68	11.81	9.63	7.49	7.24	7.07	6.9	6.67	7.94	13.94	10.19	7.91	-	-	-	11.63	10.11	7.12
InternVL2-8B [12]	10.58	9.06	8.5	11	9.53	8.4	11.88	9.63	7.7	9.67	8.67	8.19	10.55	8.84	8.94	14.63	12	6.95	14.52	10.43	8.59	11.89	9.8	6.79
GeoChat [28]	72.77	72.74	61.9	59.85	59.85	51.31	65.02	65.02	53.31	57.15	57.15	52.2	63.72	63.72	57.31	64.73	64.73	51.47	65.03	65.03	52.4	60.52	60.52	50.48
EarthDial	73.38	73.34	62.72	61.83	61.83	52.80	72.14	72.14	60.01	63.1	63.1	54.83	68.8	68.8	62.45	61.96	61.96	47.42	64.03	64.03	52.82	63.47	54.09	

Table 7. Comparison of our EarthDial with existing generic and specialized VLMs for region captioning task across various datasets.

Model	HIT UAV [53] (ZS)						NWPU VHR 10 [13] (ZS)						Swimming Pool Dataset (ZS)						UCAS AOD [74] (Zero-Shot)					
	@0.5	@0.25	R-1	R-L	MT	@0.5	@0.25	R-1	R-L	MT	@0.5	@0.25	R-1	R-L	MT	@0.5	@0.25	R-1	R-L	MT				
GPT-4o	0.1	0.7	14.20	10.56	7.16	0.7	6.1	14.72	10.82	9.41	0.1	1.2	12.87	10.07	7.79	0.1	1.3	14.71	11.14	5.97				
InternVL2-4B [12]	0.6	6.4	28.1	27.68	23.94	10.6	29.87	30.67	29.09	21.92	0.8	4.2	28.3	28.08	24.64	4.6	31.8	21.01	20.01	11.65				
GeoChat [28]	0.8	8.0	22.82	22.22	22.27	2.2	15.27	21.46	20.74	21.38	1.8	8.8	21.45	21.15	23.94	1.45	13.63	20.02	18.81	14.22				
EarthDial	2.61	13.86	28.31	28.06	22.25	17.07	41.00	27.05	26.35	23.12	1.9	7.4	29.7	29.31	22.77	8.5	34.02	21.17	20.28	13.01				

Table 8. Comparison of EarthDial with existing generic and specialized VLMs on the grounding description task across multiple datasets.

Model	NWPU RESISC45 Captions [14]			RSCID Captions [41]			RSITMD Captions [64] (Zero-shot eval)			Sydney Captions [47]			UCM Captions [47]			
	Rouge1	Rouge-L	Meteor	Rouge1	Rouge-L	Meteor	Rouge1	Rouge-L	Meteor	Rouge1	Rouge-L	Meteor	Rouge1	Rouge-L	Meteor	
MiniGPTv2	19.43	14.86	28.16	20.53	15.59	26.03	18.31	14.22	24.83	14.52	12.55	18.87	25.77	20.58	33.18	
Qwen-VL [5]	38.57	67.59	61.00	55.35	Qwen-VL [5]	66.44	60.41	63.06		15.71	13.8	19.69	22.9	17.91	28.61	
InternVL2-8B [12]	20.69	15.64	30.18	21.59	16.13	28.17	18.91	14.65	26.02	12.0	11.26	10.63	14.4	13.22	14.27	
GeoChat [28]	14.86	12.54	15.21	13.48	11.59	12.39	13.41	11.5	12.33							
LHRS-Bot [43]	88.51	90.00	89.07	89.19	EarthGPT [70]	62.77	79.53	72.06								
EarthDial	92.58	92.75	94	92.70	EarthDial	58.89	83.11	72.45								

Table 10. Comparison of EarthDial with existing VLMs for visual question answering task (left: RSVQA-LRBEN, right: RSVQA-HRBEN). Comp: Comparison, R/U: Rural/Urban.

description tasks, where existing methods struggle to detect/localize the objects, our EarthDial achieves higher mAP. Also, improved results on SAR imagery datasets showcase the multi-modal data processing capability of our EarthDial.

Image Captioning and Visual Question Answering: Our EarthDial outperforms the existing generic and specialized VLMs by clear margins over image captioning datasets as shown in Tab. 9. In addition, for the VQA task, we utilize datasets RSVQA-LRBEN and RSVQA-HRBEN (zero-shot) for evaluation, following [28]. Tab. 10 presents the VQA accuracy of existing models compared to the proposed EarthDial, outperforming most of the categories.

Change Detection: To demonstrate the temporal data processing capability of the proposed EarthDial, we evaluate its performance on a change detection task. We applied the data fusion strategy (discussed in Sec. 3.1) to merge tokens

Model	Dubai CC			LEVIR MCI [37]			MUDS [62]			SYSU [44] (zero-shot)		
	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT	R-1	R-L	MT
GPT-4o	8.81	7.45	18.68	10.33	8.4	22.05	14.18	11.02	20.92	16.48	12.32	17.49
InternVL2-4B [12]	7.31	6.38	21.12	8.88	7.43	22.14	10.25	7.9	17.73	13.27	9.98	14.36
GeoChat [28]	14.21	14.19	28.91	17.15	35.42	12.35	12.28	12.23	15.98	13.45	12.02	13.96
EarthDial	31.94	30.66	55.83	33.78	30.47	74.8	28.16	24.03	33.56	18.03	17.42	14.98

Table 11. Comparison of EarthDial with existing generic and specialized VLMs on the change detection task.

obtained from multiple images across time. The results in Tab. 11 highlight EarthDial’s strong ability to interpret and respond effectively to temporal data.

Temporal Disaster Assesment: Here, we show the capability of our EarthDial in processing temporal data. First, we consider the benchmark xBD dataset belonging to the disaster assessment task for the experiment. xBD dataset has two images: pre-disaster and post-disaster. Thus, from xBD dataset, we prepare eight sub-tasks covering temporal image captioning, region classification, image classification, object detection, and referred object detection applications. Briefly described as below:

Image Captioning: Compared to the pre-disaster image, describe the damage observed in the post-disaster image.

Region Classification: We have two sets here. In Test Set-1, classification of the level of damage in the user-marked region is included. In Test Set-2, binary classification of the user-marked region into ‘damage’ or ‘no-damage’ class is

Model	Image Captioning			Region Classification		Image Classification			Object Detection		Referred Object Detection	
	R-1	R-L	MT	Test Set-1	Test Set-2	Test Set-1	Test Set-2	Test Set-3	mAP@0.5	mAP@0.25	mAP@0.5	mAP@0.25
GPT-4o	14.21	10.35	19.52	51.68	71.62	67.95	75.45	70.41	0.2	2.15	0	0
InternVL2-8B	13.89	10.37	14.92	14.39	58.33	51.44	61.52	51.12	0.6	1.07	0	0.7
GeoChat	14.18	10.67	12.20	25.30	57.65	53.32	52.19	49.51	1.15	7.2	0.2	3.09
EarthDial	87.26	87.26	88.53	53.7	83.09	96.37	82.85	54.01	7.6	21.11	5.1	13.09

Table 12. Comparison of our EarthDial for various tasks on the xBD dataset (temporal). R-1, R-L, and MT denote ROUGE-1, ROUGE-L, and METEOR scores, respectively. **Image Captioning:** Describes the damage observed in the post-disaster image. **Region Classification:** Test Set-1, model classifies the level of damage in the user-marked region. Test Set-2, binary classification of user marked region into 'damage' or 'no-damage' class. **Image Classification:** Test Set-1, classifies the image into the type of disaster. Test Set-2, binary classification of image into 'damage' or 'no-damage' class. Test Set-3, classifies the image based on affected building count (none, one, two, or many). Classification recall is used for the evaluation of Region and Image Classification tasks. **Object Detection:** Locates all large buildings in the post-disaster image. **Referred Object Detection:** Locates the user-referred damage region/object in the post-disaster image. It is clearly observed that our EarthDial significantly improves performance for all the tasks.

added.

Image Classification: We have three sets here. In test Set-1, the classification of the image into the type of disaster (volcano, fire, earthquake, flood, tsunami, wind) is included. In Test Set-2, binary image classification into the 'damage' or 'no-damage' class is added. In Test Set-3, examples classify the image based on the affected building count (none, one, two, or many).

Object Detection: Locates all large buildings in the post-disaster image.

Referred Object Detection: Locates the user-referred damage region/object in the post-disaster image.

In the main paper, we report results for *Image Classification test Set-1*. Here, Tab. 12 presents a summary of the performance results of both generic and specialized VLMs across the sub-tasks discussed above. EarthDial consistently outperforms all other existing VLMs by a significant margin, demonstrating its capability to effectively process temporal data for the desired task. Tab. 12 presents a summary of the performance results of both generic and specialized VLMs across the sub-tasks discussed above. EarthDial consistently outperforms all other existing VLMs by a significant margin, demonstrating its capability to effectively process temporal data for the desired task. Additionally, we evaluated our method on QuakeSet [8] for earthquake prediction using SAR imagery. We evaluate the model based on binary classification to determine whether an earthquake event occurred or not between the input SAR imagery. While GPT-4o achieves a classification accuracy of 55.86, our method outperforms it with an accuracy of 57.53.

Multi-modal Data Processing: To showcase the capability of EarthDial in processing multi-modal data, we examine multi-spectral (MS), RGB-infrared, and SAR imagery for classification and referred object detection tasks. Comparative results of the proposed EarthDial and existing GPT-4o are given in Tab. 5. Our EarthDial outperforms GPT-4o by a significant margin on both multi-spectral, RGBI, and SAR

imagery. Significant improvement in the performance highlights the effectiveness of our multi-band fusion strategy.

Urban Heat Island: For UHI, we prompt to classify the underlying region into cooler, mildly hot, and extremely hot regions. We achieve an accuracy of 56.77% in identifying the temperature trends from user-input Landsat8 bands, whereas GPT-4o achieves an accuracy of 22.68% in the same task. This shows the capability of our EarthDial in processing multi-modality data effectively compared to the existing generic GPT-4o.

Methane Plume Classification: We consider STARCOP dataset for the evaluation, which has a four-channel image (RGB + mag1c band). Our EarthDial processes these RGBM bands to predict the presence of methane plumes as "Yes" or "No". We compare our accuracy with the GPT4o model. We simply give RGB and Mag1c bands to the GPT4o and collect the response. With this, GPT4o achieves an accuracy of 40.93%, while our EarDial achieves 77.09% i.e., improves it by 32.16%.

6. Ablation Study

Here, we demonstrate the effectiveness of our multi-stage pre-training, and multi-spectral band fusion strategy.

Effect of Multi-Stage Pre-Training: We assess the performance of the EarthDial model on a complex detection task, both before and after multi-stage pre-training. For this analysis, we focus on the referred object detection task using the Geochat-Instruct dataset. With multi-stage pre-training, EarthDial's mAP @0.5 improves by 5%, showing a notable boost in detecting multiple referred objects compared to its performance without pre-training.

Effect of Multispectral Band Fusion: In this setup, we evaluate EarthDial's performance with respect to the data fusion module (using average pooling and bilinear interpolation) for multi-spectral band fusion, on the classification tasks for BigEarthNet and TreeSat AI. Table 13 shows that the bilinear interpolation fusion strategy is more effective than simple average pooling, enhancing model's av-



Figure 4. Illustration of our versatile **EarthDial** model that performs across multi-modalities, multi-resolution, multispectral, and multi-temporal data from diverse remote sensing applications. **EarthDial** extends its capabilities to a range of tasks such as scene classification, image/region-captioning, referring expression, VQA, referring expression, object detection, temporal change/disaster detection, Methane plume detection, tree species classification, UHI, and LCZs detection across multi-modalities, multi-resolution remote sensing data.

Referred object detection task (GeoChat-Instruct dataset [28])						Multi-band classification			
AHR	Pre-training	Small	Medium	Large	Single	Data Fusion	BigEarthNet	TreeSat MS	AI
\times	\times	6.06	22.97	36.26	24.75	10.3	Average	47.66	49.09
\times	\checkmark	10.85	32.05	36.15	33.73	11.91	Max-pooling	34.62	29.84
\checkmark	\checkmark	11.75	33.33	42.60	34.78	15.03	Bilinear	67.01	56.93

Table 13. Ablation study about the effect of multi-stage training (sec. 3.2) and spectral band fusion strategy (sec. 3.1). The proposed multi-stage pre-training approach improves EarthDial’s mAP @0.5 for detecting multiple referred objects by 5%. Additionally, the use of our bilinear fusion strategy enhances the average classification accuracy by 9.5%.

verage accuracy by 13.5%. The impact of multi-spectral data combined with our multi-band fusion strategy is evident in Table 5. EarthDial achieves a 1.75% improvement in classification accuracy on the Multi-Spectral (MS) version of BigEarthNet compared to its RGB counterpart. This

demonstrates EarthDial’s ability to leverage complementary information from multispectral bands, enhancing performance in multi-class classification tasks.

7. Conclusion

We present EarthDial, a conversational assistant purpose-built for Earth Observation (EO) data, capable of transforming complex, multi-sensory Earth observations into interactive, and natural language dialogues. EarthDial supports multi-spectral, multi-temporal, and multi-resolution imagery as input and addresses a wide spectrum of remote sensing tasks, including classification, detection, captioning, question answering, visual reasoning, and visual grounding. To enable this versatility, we developed a comprehensive instruction-tuning dataset containing over 11M instruction pairs, encompassing diverse modalities such as

RGB, S2, Synthetic Aperture Radar (SAR), Near-Infrared (NIR), and infrared. Additionally, EarthDial excels in handling bi-temporal and multi-temporal sequence analysis, making it highly effective for applications like change detection/disaster assessment. Our experiments across 44 downstream tasks highlight EarthDial’s superior performance over existing generic and domain-specific models, demonstrating its robust generalization capabilities and its potential to set a new standard in EO task automation.

Appendix

Here, we first provide details about the EarthDial-Instruct dataset used to train our model, in three stages. Second, we conduct an ablation study comparing the performance of the EarthDial model fine-tuned with LoRA against the fully fine-tuned version, evaluating both models on zero-shot detection datasets. Last, we provide more qualitative analysis of our EarthDial model, compared to recent state-of-the-art VLMs, demonstrating its better generalization across multi-modalities, multi-resolution, and multi-temporal downstream EO tasks.

A. EarthDial-Instruct Dataset

The fundamental objective of constructing domain-specific VLM is to improve generalization performance on diverse downstream tasks, covering a wide range of modalities, multi-resolution, and multi-temporal data. Therefore, we curate high-quality pre-train question-answer (QA) instruction pairs from SkyScript [58] and SatlasPretrain [6] data, which includes Sentinel-2 (S2), Sentinel-1 (SAR), NAIP, and Landsat imagery along with labels. Specifically, we choose InternLM-XComposer2 [22] as an instruction generator after evaluating its generation outputs against state-of-the-art leading VLMs at the time of selection, where it demonstrated superior efficiency in handling large-scale data for generating vision QA instruction pairs. The methodology involved multiple steps of filtering to ensure the quality of the data, as depicted in Fig. A.1. In step I we proceed with a label-based filtering, where we filter out samples that are associated with at least three labels, ensuring that each image contained enough descriptive content to support meaningful instruction samples. In step II, an image-based filtering is applied, where we apply luminance and coverage-based filtering to remove cloudy images as well as low spatial coverage images. More specifically, we apply a threshold on the average luminance and remove images with insufficient coverage. In step III, we prompt the InternLM-XComposer2 to generate QA instruction pairs based on the key attributes (points, polygons, object category, and position) specified in the inputs and labels. These attributes, before being input in the processing pipeline, undergo formatting to natural language to be

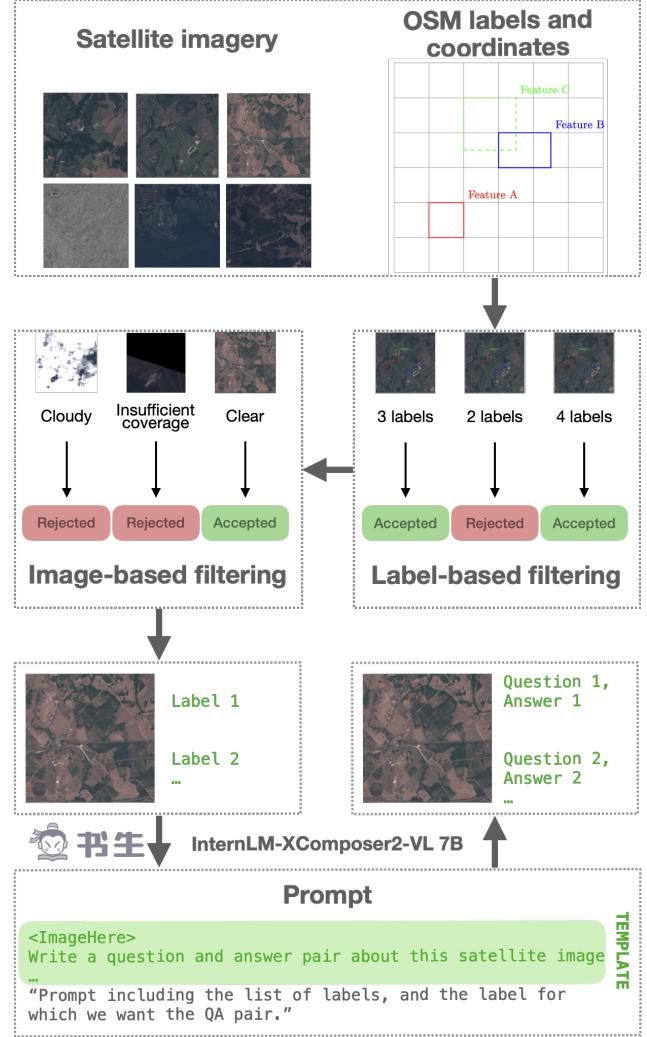


Figure A.1. Overview of the data preparation and filtering pipeline used in the QA instruction dataset generation. The process begins with the pairing of OpenStreetMap (OSM) labels and their corresponding different sources of satellite imagery. The data goes through a label-based filtering process selecting only images with 3 labels or above, and then this data undergoes a second filtering process which is image-based to remove low-quality images. The high quality images remaining are then passed to the InternLM-XComposer2-VL model to generate question-answer pairs based on the associated reliable labels from OSM.

understood by the VLM. When processing a sample, we prompt the model multiple times, asking for a QA instruction set for each attribute specifically. Each prompt also contains information about all the other attributes detected in the image. Furthermore, in the same prompt, we provide an example of a satisfactory QA instruction set, sampled from a list of predefined instruction sets. The generation is repeated up to 5 times, if the expected format is not respected. We present the workflow explicitly below:

1. A satisfactory QA instruction set example: *Subject: parking lot. Question: How does the parking lot contribute to environmental sustainability? Answer: The parking lot in the lower left seems to be equipped with solar panel canopies, promoting renewable energy use.*
2. The prompt: *Write a question and answer pair about this satellite image. For example, on another image, a satisfactory pair is: `satisfactory_qa.instruction`. The current image has been annotated with the following keywords: `attribute_1`, `attribute_2`, Generate the pair for the following subject: `attribute_1`, which is visible in the satellite image. The question or answer must refer to the `attribute_1`, and must refer to either its position, interaction with other elements in the image, characteristics, or function. The answer must be objective, based on visible elements in the image, and require the image to answer. Avoid any assumptions or extrapolations that are not clearly supported by the image.*
3. The template: <ImageHere>*the prompt*.

We manually verify randomly drawn parts of the instruction sets to validate the quality of generated instructions.

Downstream Tasks Image-text Instruction

Though pre-training enhances the generalization capabilities, we also need task-specific fine-tuning with diverse data types to improve downstream performance as shown in Tab. A.1 and Fig. A.2. We curate a large number of instruction-following datasets that include ten diverse downstream tasks: scene classification, object detection, visual question answering, image captioning, change detection, Methane plume detection, tree species classification, local climate zones, urban heat islands, and disaster assessment. It covers seven diverse visual modalities that include Optical, SAR, S2, Infrared, NIR, Landsat8, and Hyperspectral, and two visual temporal modalities (Optical and SAR).

B. Ablation on LoRA vs Full Fine-tuning

It is interesting to understand how different adaptation mechanisms can influence the performance after Stage 1 model pretraining. Here we explore Low-rank adaptation (LoRA) in comparison to full finetuning. LoRA is interesting to explore since it allows finetuning the model with minimal memory requirements, adds only a few additional tunable weights and helps retain knowledge acquired during the previous training stages. Specifically, for LoRA, we retain the pre-trained weights from Stage 1 and instead of full finetuning, only train the low-rank adapter weights which are then added to the original pretrained weights.

For the LoRA fine-tuning, we used a LoRA rank of 128, a batch size of 2, and a learning rate of 4e-5. This setup updated approximately 201M parameters in comparison to the EarthDial model’s 4 billion total parameters while keeping the Vision Transformer (ViT), MLP, and LLM components

frozen. The fine-tuning leveraged thumbnail images to capture global features and utilized an adaptive patch size ranging from 1 to 6 to capture more detailed high-level features.

The LoRA fine-tuning was performed on 2 NVIDIA A100 GPUs (80 GB each) and the model was then evaluated on zero-shot detection datasets. Compared to the fully finetuned model, the LoRA fine-tuned model exhibited lower performance, as summarized in Table A.2. The LoRA fine-tuned model exhibited lower performance compared to the fully fine-tuned model due to its limited parameter updates, frozen components, and constrained adaptability for complex zero-shot detection tasks.

As seen from Table A.2, the results indicate that EarthDial (Ours) significantly outperforms EarthDial-Lora across all metrics. Specifically, EarthDial (Ours) achieves a substantial improvement in detecting multiple objects (from 2.6 to 6.7) and large objects (from 9.2 to 25.67) on the Urban Tree Crown Detection dataset. A similar trend is observed on the Swimming Pool dataset, showcasing Earthdial (full-finetuning) model’s superior performance in handling the referred object detection task effectively.

B.1. Qualitative Analysis:

In Fig. A.3, we present a qualitative analysis of EarthDial. We compare our method with existing state-of-the-art InternVL-4B [12], GPT-4o [45], and GeoChat [28] VLMs. We notice that EarthDial shows better capability to detect the object for the SAR and infrared imagery, especially in crowded scenes. For the multi-label scene classification, our model outputs multi-labels whereas other compared models output limits to a single label. For bi-temporal and multi-temporal change detection, we observe that our model shows better capability to identify the semantic changes in the complex scenes and indicates the newly constructed roads and buildings. For disaster assessment, over optical and SAR imagery, our model has better capability to identify the underlying structure and performs better for disaster understanding. In addition, over RGB+NIR and S2 imagery, we compare our model with GPT-4o while InternVL-4B and GeoChat do not support multi-spectral data processing. The qualitative comparison shows that our model has better capability to handle multi-spectral imagery data and performs better. Our qualitative comparison demonstrates the merits of EarthDial by consistently showing better performance on challenging scenarios across different modalities, multi-resolution, and multi-temporal imagery data. In Fig. A.4, we also present the failure cases where EarthDial fails under complex scenarios. For instance, identifying green medium tree at the left is difficult because there are many green trees in the input. Similarly, prompting to identify the ship provided with the bounding box may cause failure because the training set includes limited ship information compared to the vehicles. Introducing more SAR

Task	Dataset	Split	Type	QA Examples
Scene Classification	AID [60]	test	Optical	
	UCMerced-LandUse [63]	test	Optical	
	WHU-RS19 [18]	test	Optical	
	EuroSat [25]	test	Optical, S2	
	BigEarthNet [50]	train/val/test	Optical, S2	
	NWPU-RESISC45 [14]	train	Optical	
	PatternNet [72]	train	Optical	
	RS-CD [31]	train	Optical	
	RSI-CD256	train	Optical	
	FMoW [17]	train/val	Optical	
Object Detection	FGSCR-42 [20]	train	Optical	
	TreeSatAI-Time-Series [4]	train/val/test	Optical, NIR	
	SoSAT-LCZ42 [75]	train/val/test	S2	
	DOTA [21]	train/test	Optical	
	DOIR [33]	train/test	Optical	
Visual Grounding	FAIR-1M [51]	train/test	Optical	
	HIT-UAV [53]	train/test	Infrared	
	UCAS-AOD [74]	train/val/test	Optical	
	NWPU-VHR10 [13]	test	Optical	
	SAR-Ship [57]	train/test	SAR	
	SRSDD-v1.0 [29]	train/test	SAR	
	Swimming Pool & Car	test	Optical	
	Urban Tree Crown Detection [65]	test	Optical	
Visual Question Answering	DIOR-RSVG [66]	train/val/test	Optical	User: An airplane on the upper left. EarthDial: [bbox].
	RSVG [52]	train/val/test	Optical	
	FloodNet [48]	train	Optical	User: Is a parking present?
	MQVQA [69]	train	Optical	EarthDial: Yes.
	RSVQA-LRBEN [40]	train/test	Optical	User: What is the amount of residential buildings?
	RSVQA-HRBEN [40]	test	Optical	EarthDial: 14.
Image Captioning	Sydney Caption [47]	train/test	Optical	User: What elements in the picture suggest its theme?
	UCM Caption [47]	train/val/test	Optical	EarthDial: Runway.
	RSITMD [64]	train/test	Optical	
	RSCID [41]	train/val/test	Optical	
	NWPU-Captions [15]	train/val/test	Optical	User: Could you provide the caption for input image? EarthDial: Many white planes were parked at the airport.
Change Detection	Sydney Caption [47]	train/test	Optical	
	UCM Caption [47]	train/val/test	Optical	
	LEVIR-MCI [37]	train/val/test	Optical	User: Are there any semantic changes detected in images?
	SYSU-CC [44]	test	Optical	EarthDial: Two houses are built at the top of the scene.
Methane Plume Detection	Dubai-CC	train/val/test	Optical	
	MUDS [62]	train/test	Optical	
	STARCOP [49]	train/test	Hyperspectral	User: Does this image have a methane plume? EarthDial: Yes. User: Give me the location of the methane plume. EarthDial: [bbox]. User: What is the emission rate of methane plume? EarthDial: The emission rate is 11239kg/h.
Urban Heat Islands				User: What is the temperature trend in the input? EarthDial: mildly hot.
	UHI-AD	train/test	Landsat8	User: What factors are responsible for the temperature? EarthDial: Urbanization and few water bodies. User: What sustainable practices can mitigate UHI effect? EarthDial: Introduce fountains, green corridors, and ponds.
Disaster Assesment	QuakeSet [8]	train/val/test	SAR	User: Do input images present earthquake effects? EarthDial: Yes. User: Could you tell the magnitude of the earthquake? EarthDial: 5.58mb.
	xBD [24]	train/test	Optical	User: Identify the type of disaster that occurred. Options: flood, wind, fire, tsunami, earthquake, volcano? EarthDial: Volcano. User: Are there any buildings affected due to disaster? EarthDial: Yes. User: Identify major-damaged building located at center. EarthDial: [bbox]. User: Is the building at [bbox] affected due to disaster? EarthDial: Yes. User: Describe the damage observed in the post-disaster image. EarthDial: There has been a volcano disaster that resulted in many damaged buildings. User: How many building are affected? EarthDial: Many. User: Locate all large buildings in the post-disaster image. EarthDial: [bbox], [bbox], [bbox]. User: Give the level of damage for [bbox]. EarthDial: Destroyed.

Table A.1. Overview of the downstream datasets that include various tasks, splits, types (modalities), and the generated question-answer pair (QA-pair) examples from the respective datasets. Here, split means that we generate QA-pairs for each split separately. The [bbox] indicates the bounding box of the object as $[x_{min}, y_{min}, x_{max}, y_{max}, \theta]$.

Model	Swimming Pool Dataset (ZS)					Urban Tree Crown Detection [65] (ZS)				
	Small	Medium	Large	Single	Multiple	Small	Medium	Large	Single	Multiple
GeoChat [28]	-	3.1	7.3	1.2	0.6	-	1.8	8.9	2.9	3.1
InternVL2-4B [12]	0.6	6.6	8.9	4.5	0.865	-	3.17	13.41	5.9	3.1
EarthDial-Lora	1.3	2.6	9.45	4.3	0.7	0.2	2.6	9.2	4.1	2.6
EarthDial (Ours)	1.04	7.4	24.90	8.4	1.04	1.1	7.01	25.67	11.13	6.7

Table A.2. Comparison of our EarthDial for referred object detection tasks across various datasets. We use mAP@0.5 as the evaluation metric. Small, medium, and large denote the object size, while single and multiple denote the number of objects. Here, ZS means zero-shot evaluation.

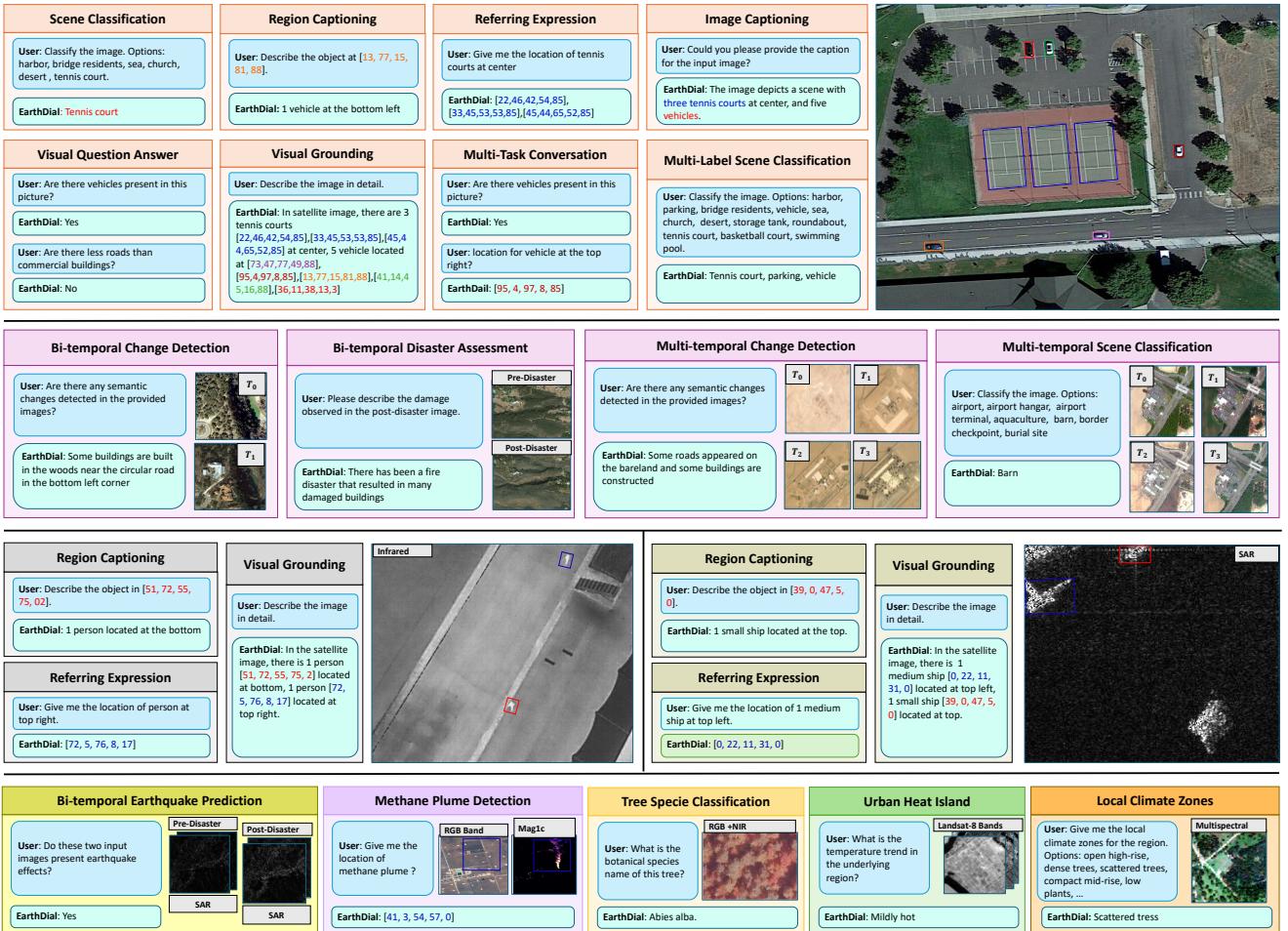


Figure A.2. Illustration of our versatile **EarthDial** model that performs across multi-modalities, multi-resolution, multispectral, and multi-temporal data from diverse remote sensing applications. **EarthDial** extends its capabilities to a range of tasks such as scene classification, image/region-captioning, referring expression, VQA, referring expression, object detection, temporal change/disaster detection, Methane plume detection, tree species classification, UHI, and LCZs detection across multi-modalities, multi-resolution remote sensing data.

ship QA-pairs in the training set might improve the performance. On the other hand, detecting subtle change regions is difficult due to the nature of small semantic changes. For temporal scene classification, since the office building and

multi-unit residential are similar in nature, therefore model might fail under such complex scenes. Nevertheless, our model encapsulates the distinctive contextual complexities of diverse RS applications and performs better compared

 InternVL-4B: 1 airplane at the right GPT-4o: The runway is on bare land next to the grass. GeoChat: 1 airplane at the right EarthDial: A plane is on the runway beside the grass.	 InternVL-4B*: 1 vehicle at left GPT-4o: 1 car at left GeoChat: 1 vehicle at left EarthDial: 1 ship at the left	 InternVL-4B*: 1 vehicle at bottom left GPT-4o: 1 car at right GeoChat: 1 vehicle at bottom left EarthDial: 1 medium car at the bottom left	 InternVL-4B: N/A GPT-4o: Pinus nigra GeoChat: N/A EarthDial: Picea abies
 InternVL-4B: No GPT-4o: No GeoChat: Yes EarthDial: Yes	 InternVL-4B*: 1 building at center GPT-4o: Some building at center GeoChat: Some building at center EarthDial: 1 medium ship at the center	 InternVL-4B*: 1 vehicle at the top GPT-4o: 1 vehicle at the top GeoChat: 1 vehicle at the top right EarthDial: 1 small car at the top right	 InternVL-4B: Non-irrigated arable land GPT-4o: Mixed forest GeoChat: Mixed forest EarthDial: Non-irrigated arable land, coniferous forest, mixed forest
 InternVL-4B: Many houses are built along the roads. GPT-4o: The two scenes seem identical. GeoChat: Houses are built at the right. EarthDial: The forest disappears and many houses and roads appear.	 InternVL-4B: There are no semantic changes detected in the provided images GPT-4o: The area includes a mix of natural and man-made features, roads. GeoChat: Few houses are build. EarthDial: Roads and buildings have taken the place of grassland across the main road.	 InternVL-4B: N/A GPT-4o: Coniferous forest GeoChat: N/A EarthDial: Coniferous forest, Mixed forest, Transitional woodland/shrub	 InternVL-4B: N/A GPT-4o: Scattered trees GeoChat: N/A EarthDial: Dense trees
 InternVL-4B: No GPT-4o: No GeoChat: No EarthDial: Yes	 InternVL-4B: There is a disaster GPT-4o: Many buildings have disappeared. GeoChat: Many buildings in the images. EarthDial: There has been an earthquake that damaged many buildings.	 InternVL-4B*: Some building at the top GPT-4o: [28, 20, 42, 32, 0] GeoChat: [30, 32, 45, 45, 90] EarthDial: [34, 16, 39, 20, 0]	 InternVL-4B: Aquaculture GPT-4o: Flooded road GeoChat: Flooded road EarthDial: Port

Figure A.3. Illustration of the qualitative comparison of our **EarthDial** with state-of-the-art VLMs (InternVL-4B [12], GPT-4o [45], GeoChat [28]). It demonstrates the merits of our approach by performing better under challenging scenarios across multi-modalities, multi-resolution, and temporal input data. Here, InternVL-4B* indicates that it is trained over GeoChat-Instruct. As existing InternVL2 doesn't provide the rotated bounding boxes, for a fair comparison, we finetune the InternVL2-4B on GeoChat-Instruct and compared it with our EarthDial (only detection-related tasks).

 EarthDial: [28, 61, 35, 75, 2]	 EarthDial: Several buildings are around a square.	 EarthDial: 1 vehicle at the center	 EarthDial: Quercus rubra
 EarthDial: The two scenes seem identical	 EarthDial: There has been a wind disaster, however it did not effect the buildings	 EarthDial: Dump sites, continuous urban fabric	 EarthDial: Multi-unit residential

Figure A.4. Illustration of the failure cases of our **EarthDial**. Our method fails under ambiguous and complex scenarios. For example, prompting the model to provide the medium tree with the input of many green trees. Similarly, for the change detection task, the model fails to detect the subtle changes that occurred at the bottom right of the scene due to variations in texture that are not easily distinguishable.

to existing generalized and domain-specific VLMs across different modalities, multi-resolution, multi-spectral, and multi-temporal RS sensor data.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree,

- Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 4
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [4] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. *arXiv preprint arXiv:2404.08351*, 2024. 6, 12
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7
- [6] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinand, 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. 5, 10
- [7] Luxuan Bian. Nwpv vhr-10, 2023. 7
- [8] Daniele Rege Cambrin and Paolo Garza. Quakeset: A dataset and low-resource models to monitor earthquakes through sentinel-1. *arXiv preprint arXiv:2403.18116*, 2024. 6, 8, 12
- [9] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 2
- [10] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1
- [11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 2, 3, 4
- [12] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 4, 6, 7, 11, 13, 14
- [13] Gong Cheng, Junwei Han, Peicheng Zhou, and Lei Guo. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98: 119–132, 2014. 7, 12
- [14] 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. 7, 12
- [15] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpv-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 12
- [16] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. *arXiv preprint arXiv:2407.01851*, 2024. 2
- [17] 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. 5, 6, 12
- [18] Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and remote sensing letters*, 8(1):173–176, 2010. 6, 12
- [19] Zhichao Deng, Xiangtai Li, Xia Li, Yunhai Tong, Shen Zhao, and Mengyuan Liu. Vg4d: Vision-language model goes 4d video recognition. *arXiv preprint arXiv:2404.11605*, 2024. 2
- [20] Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4):747, 2021. 12
- [21] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7778–7796, 2021. 12
- [22] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 5, 10
- [23] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 2
- [24] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019. 6, 12
- [25] 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. 12

- [26] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 2
- [27] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2
- [28] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 2, 3, 6, 7, 9, 11, 13, 14
- [29] Songlin Lei, Dongdong Lu, Xiaolan Qiu, and Chibiao Ding. Srsdd-v1. 0: A high-resolution sar rotation ship detection dataset. *Remote Sensing*, 13(24):5104, 2021. 7, 12
- [30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 4
- [31] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large-scale remote sensing image classification benchmark using crowd-sourced data. *Sensors*, 20(6):1594, 2020. 12
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [33] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 12
- [34] 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. 2
- [35] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 2
- [36] Guibiao Liao, Jiankun Li, and Xiaoqing Ye. Vlm2scene: Self-supervised image-text-lidar learning with foundation models for autonomous driving scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3351–3359, 2024. 2
- [37] Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 7, 12
- [38] Fan Liu, Delong Chen, Zhangqiyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Re-moteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1
- [40] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. 12
- [41] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017. 7, 12
- [42] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024. 2
- [43] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *arXiv preprint arXiv:2402.02544*, 2024. 2, 7
- [44] Mubashir Noman, Noor Ahsan, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Cdchat: A large multimodal model for remote sensing change description. *arXiv preprint arXiv:2409.16261*, 2024. 7, 12
- [45] OpenAI. Gpt-4 technical report. *arXiv preprint*, abs/2303.08774, 2023. Available at <https://doi.org/10.48550/arXiv.2303.08774>. 11, 14
- [46] C. Pang, X. Weng, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, S. Wang, L. Feng, G.S. Xia, and C. He. VHM: Versatile and Honest Vision Language Model for Remote Sensing Image Analysis. *arXiv*, 2024. 2, 3
- [47] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, pages 1–5. IEEE, 2016. 7, 12
- [48] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debrat Varshney, Masoud Yari, and Robin Robertson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 12
- [49] Vit Ruzicka, Gonzalo Mateo-Garcia, Luis Gomez-Chova, Anna Vaughan, Luis Guanter, and Andrew Markham. Semantic segmentation of methane plumes with hyperspectral machine learning models. *Scientific Reports*, 13(1):19999, 2023. 6, 12
- [50] 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. 5, 6, 12
- [51] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sens-

- ing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 12
- [52] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 404–412, 2022. 6, 12
- [53] Jiashun Suo, Tianyi Wang, Xingzhou Zhang, Haiyang Chen, Wei Zhou, and Weisong Shi. Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*, 10(1):227, 2023. 7, 12
- [54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [55] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024. 2
- [56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2
- [57] Yuanyuan Wang, Chao Wang, Hong Zhang, Yingbo Dong, and Sisi Wei. A sar dataset of ship detection for deep learning under complex backgrounds. *remote sensing*, 11(7):765, 2019. 7, 12
- [58] 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. 5, 10
- [59] Zidan Wang, Rui Shen, and Bradly Stadie. Solving robotics problems in zero-shot with vision-language models. *arXiv preprint arXiv:2407.19094*, 2024. 2
- [60] 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. 6, 12
- [61] Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. *arXiv preprint arXiv:2410.13860*, 2024. 2
- [62] Charig Yang, Weidi Xie, and Andrew Zisserman. Made to order: Discovering monotonic temporal changes via self-supervised video ordering. *arXiv preprint arXiv:2404.16828*, 2024. 6, 7, 12
- [63] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 6, 12
- [64] 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, 12
- [65] Pedro Zamboni, José Marcato Junior, Jonathan de Andrade Silva, Gabriela Takahashi Miyoshi, Edson Takashi Matsubara, Keiller Nogueira, and Wesley Nunes Gonçalves. Benchmarking anchor-based and anchor-free state-of-the-art deep learning methods for individual tree detection in rgb high-resolution images. *Remote Sensing*, 13(13):2482, 2021. 7, 12, 13
- [66] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 6, 12
- [67] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv preprint arXiv:2401.09712*, 2024. 2, 3
- [68] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv: 2401.17600*, 2024. 1
- [69] Meimei Zhang, Fang Chen, and Bin Li. Multi-step question-driven visual question answering for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 12
- [70] 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. 2, 3, 7
- [71] 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. 3
- [72] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 12
- [73] Yijie Zhou, Likun Cai, Xianhui Cheng, Zhongxue Gan, Xiangyang Xue, and Wenchao Ding. Openannotate3d: Open-vocabulary auto-labeling system for multi-modal 3d data. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9086–9092. IEEE, 2024. 2
- [74] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE international conference on image processing (ICIP)*, pages 3735–3739. IEEE, 2015. 7, 12
- [75] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, et al. So2sat lc242: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019. 5, 6, 12