Supplementary Material
EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues

Sagar Soni*[1], Akshay Dudhane*[2], Hiyam Debary*[1], Mustansar Fiaz*[1], Muhammad Akhtar Munir[2]
Muhammad Sohail Danish[2], Paolo Fraccaro[1], Campbell D Watson[1], Levente J Klein[1]
Fahad Shahbaz Khan[2,4], Salman Khan[2,3]
[1]IBM Research [2]Mohamed bin Zayed University of AI
[3]Australian National University [4]Linköping University

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.

1. 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 [32] and SatlasPretrain [2] data, which includes Sentinel-2 (S2), Sentinel-1 (SAR), NAIP, and Landsat imagery along with labels. Specifically, we choose InternLM-XComposer2 [12] 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. 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 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.

Task	Dataset	Split	Type	QA Examples
Scene Classification	AID [33]	test	Optical	**User**: Classify the given image in one of the classes. Options: ground track field, chaparral, harbor, desert, ship, railway station, meadow, bridge, ... **EarthDial**: Railway station.
	UCMerced-LandUse [35]	test	Optical	
	WHU-RS19 [9]	test	Optical	
	EuroSat [14]	test	Optical, S2	
	BigEarthNet [27]	train/val/test	Optical, S2	
	NWPU-RESISC45 [6]	train	Optical	
	PatternNet [40]	train	Optical	
	RS-CD [17]	train	Optical	
	RSI-CD256	train	Optical	
	FMoW [8]	train/val	Optical	
	FGSCR-42 [10]	train	Optical	
	TreeSatAI-Time-Series [1]	train/val/test	Optical, NIR	
	SoSAT-LCZ42 [42]	train/val/test	S2	
Object Detection	DOTA [11]	train/test	Optical	**User**: Where is silver boeing737 airplane? **EarthDial**: [bbox]. **User**: What object is in this location [bbox]? **EarthDial**: 1 baseball field at the top right. **User**: Describe this image in detail. **EarthDial**: In the image, two white motorboats [bbox, bbox] are positioned close to each other on the left side.
	DOIR [18]	train/test	Optical	
	FAIR-1M [28]	train/test	Optical	
	HIT-UAV [30]	train/val/test	Infrared	
	UCAS-AOD [41]	test	Optical	
	NWPU-VHR10 [5]	test	Optical	
	SAR-Ship [31]	train/test	SAR	
	SRSDD-v1.0 [16]	train/test	SAR	
	Swimming Pool & Car	test	Optical	
	Urban Tree Crown Detection [37]	test	Optical	
Visual Grounding	DIOR-RSVG [38]	train/val/test	Optical	**User**: An airplane on the upper left. **EarthDial**: [bbox].
	RSVG [29]	train/val/test	Optical	
Visual Question Answering	FloodNet [25]	train	Optical	**User**: Is a parking present? **EarthDial**: Yes. **User**: What is the amount of residential buildings? **EarthDial**: 14. **User**: What elements in the picture suggest its theme? **EarthDial**: Runway.
	MQVQA [39]	train	Optical	
	RSVQA-LRBEN [20]	train/test	Optical	
	RSVQA-HRBEN [20]	test	Optical	
	Sydney Caption [24]	train/test	Optical	
	UCM Caption [24]	train/val/test	Optical	
Image Captioning	RSITMD [36]	train/test	Optical	**User**: Could you provide the caption for input image? **EarthDial**: Many white planes were parked at the airport.
	RSCID [21]	train/val/test	Optical	
	NWPU-Captions [7]	train/val/test	Optical	
	Sydney Caption [24]	train/test	Optical	
	UCM Caption [24]	train/val/test	Optical	
Change Detection	LEVIR-MCI [19]	train/val/test	Optical	**User**: Are there any semantic changes detected in images? **EarthDial**: Two houses are built at the top of the scene.
	SYSU-CC [22]	test	Optical	
	Dubai-CC	train/val/test	Optical	
	MUDS [34]	train/test	Optical	
Methane Plume Detection	STARCOP [26]	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	UHI-AD	train/test	Landsat8	**User**: What is the temperature trend in the input? **EarthDial**: mildly hot. **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 [3]	train/val/test	SAR	**User**: Do input images present earthquake effects? **EarthDial**: Yes. **User**: Could you tell the magnitude the earthquake? **EarthDial**: 5.58mb.
	xBD [13]	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 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 $[x_{min}, y_{min}, x_{max}, y_{max}, \theta]$.

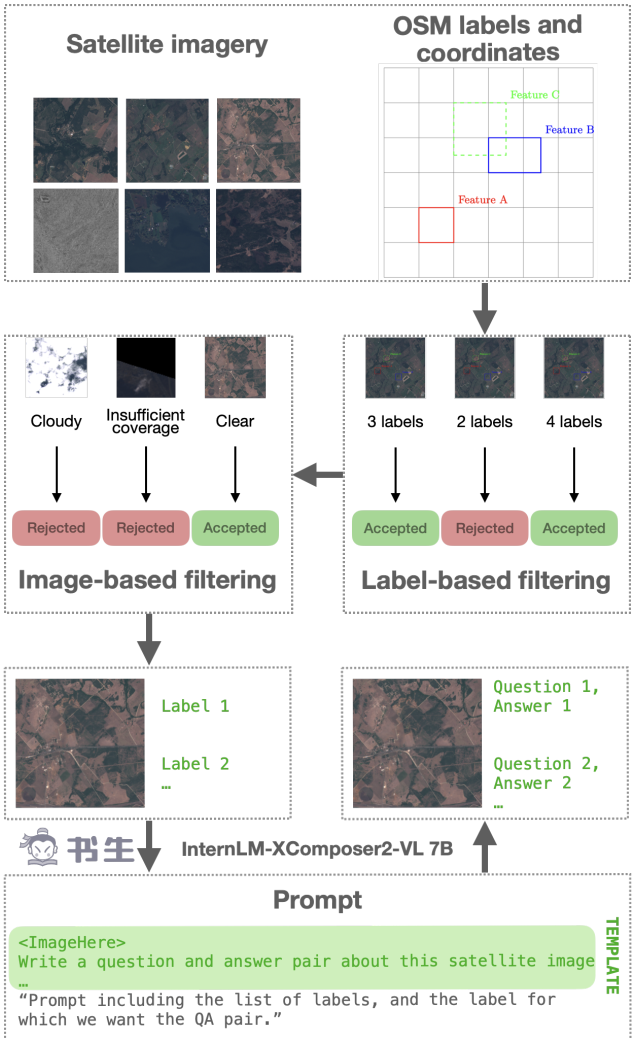

Figure 1. Overview of the data preparation and filtering pipeline used in the QA instruction dataset generation. The process begins with the pairing of OpenStreetMap (OSM) labels and their corresponding different sources of satellite imagery. The data goes through a label-based filtering process selecting 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.

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. 1 and Fig. 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).

2. 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 fine-tuned model, the LoRA fine-tuned model exhibited lower performance, as summarized in Table 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 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-finetunning) model's superior performance in handling the referred object detection task effectively.

2.1. Qualitative Analysis:

In Fig. 3, we present a qualitative analysis of EarthDial. We compare our method with existing state-of-the-art InternVL-4B [4], GPT-4o [23], and GeoChat [15] 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

Model	Swimming Pool Dataset (ZS)					Urban Tree Crown Detection [37] (ZS)				
	Small	Medium	Large	Single	Multiple	Small	Medium	Large	Single	Multiple
GeoChat [15]	-	3.1	7.3	1.2	0.6	-	1.8	8.9	2.9	3.1
InternVL2-4B [4]	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 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.

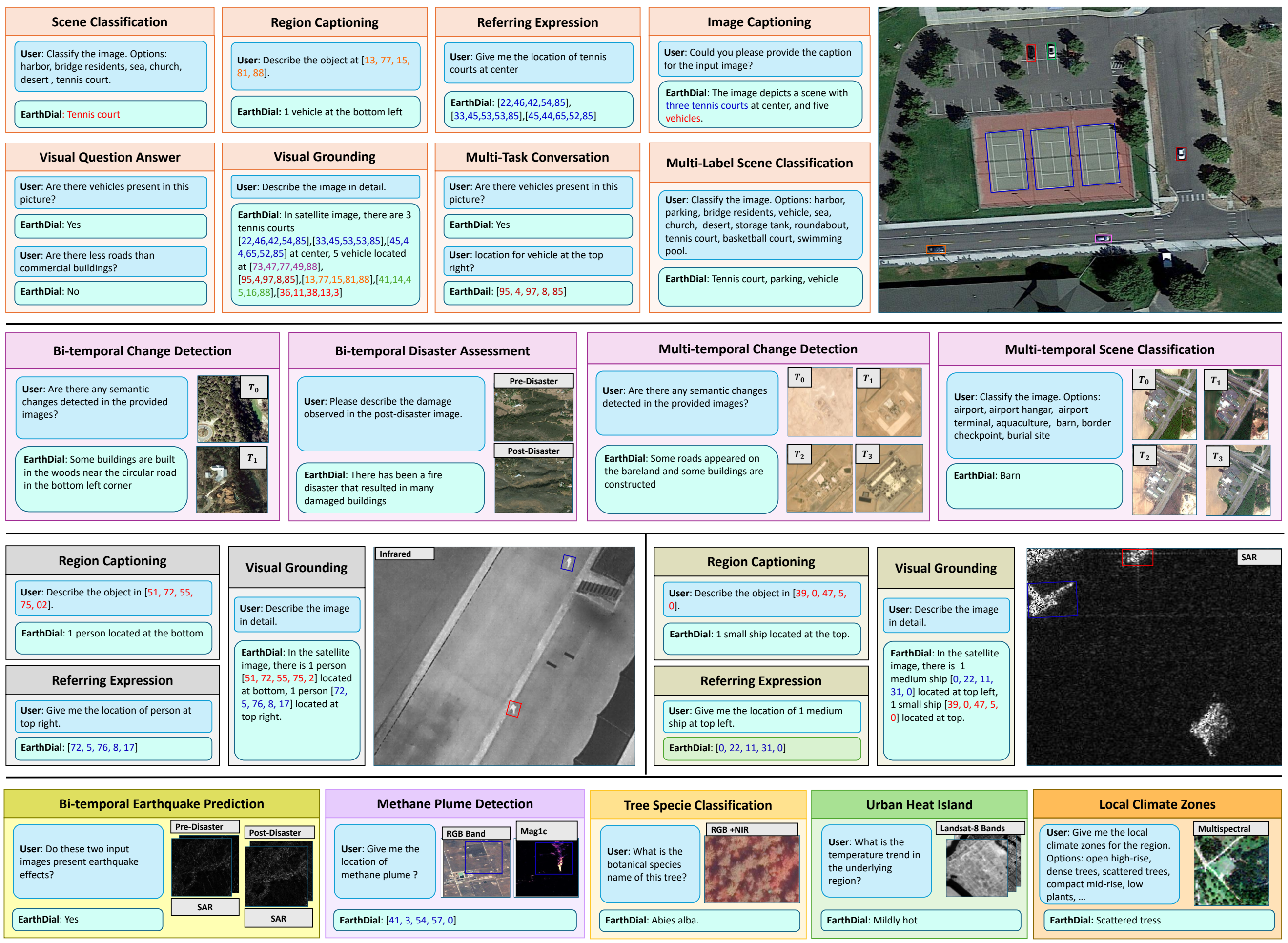

Figure 2. Illustration of our versatile **EarthDial** model that performs across multi-modalities, multi-resolution, multispectral, and multi-temporal data from diverse remote sensing applications. **EarthDial** extends its capabilities to a range of tasks such as scene classification, image/region-captioning, 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.

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 im-

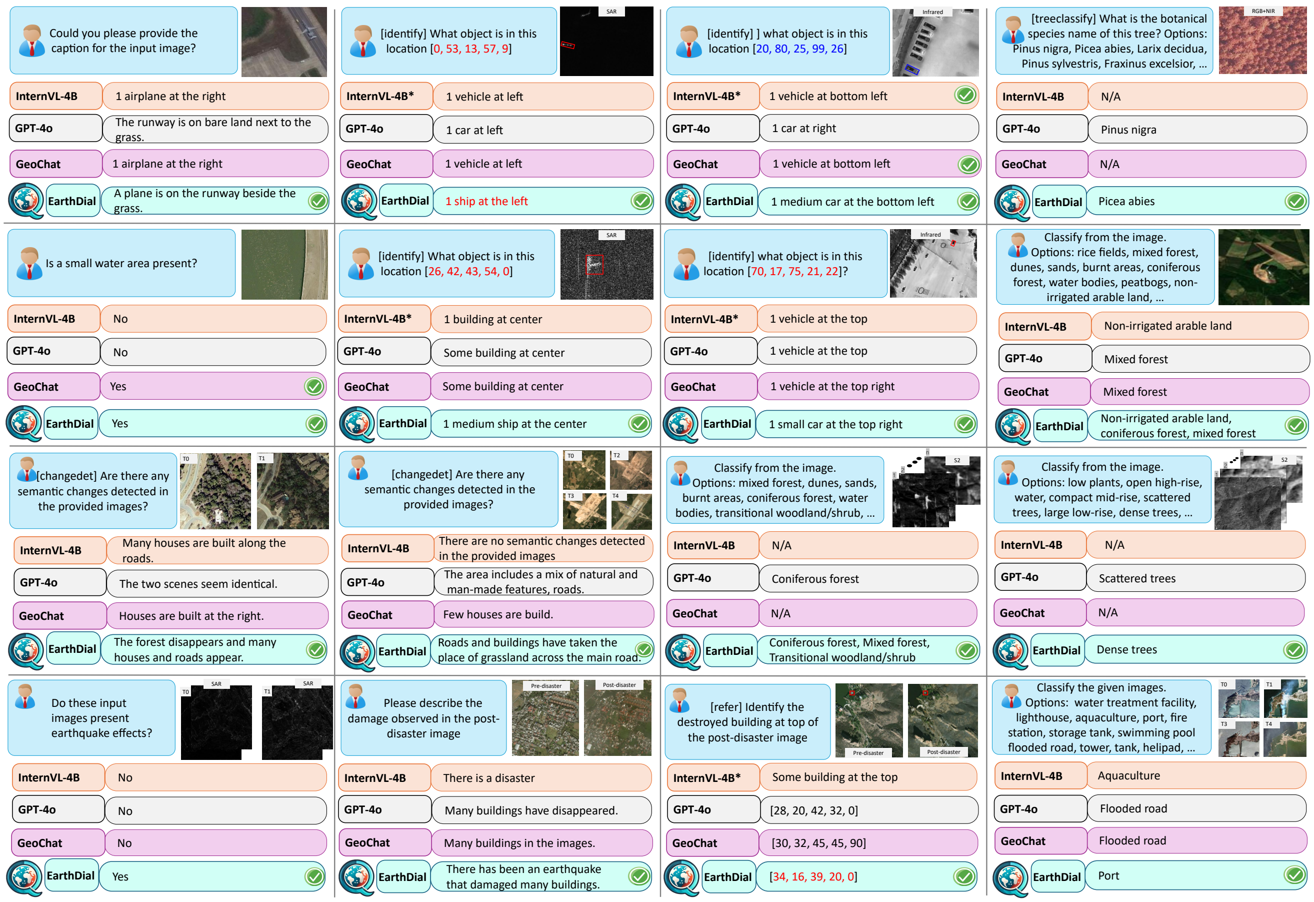

Figure 3. Illustration of the qualitative comparison of our **EarthDial** with state-of-the-art VLMs (InternVL-4B [4], GPT-4o [23], GeoChat [15]). It demonstrates the merits of our approach by performing better under challenging scenarios across multi-modalities, multi-resolution, and temporal input 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).

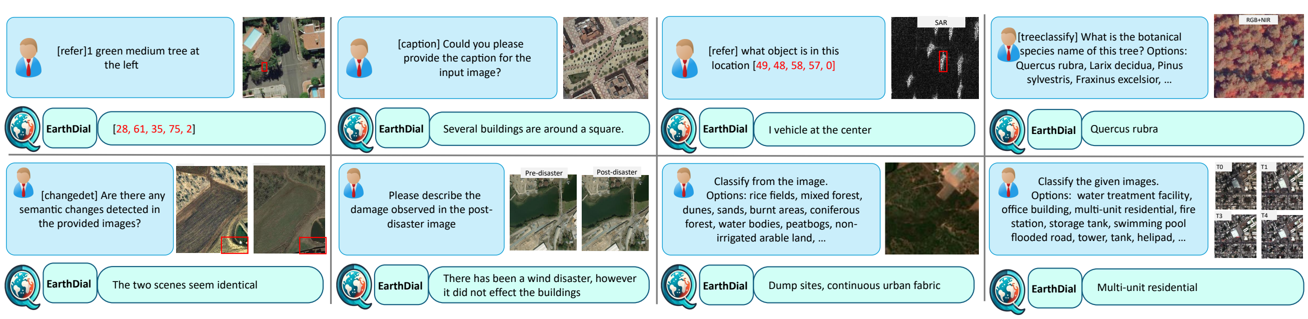

Figure 4. Illustration of the failure cases of our **EarthDial**. Our method fails under ambiguous and complex scenarios. For example, prompting the model to provide the medium tree with the input of many green trees. Similarly, for the change detection task, the model fails to detect the subtle changes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

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

References

[1] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. *arXiv preprint arXiv:2404.08351*, 2024. 2

[2] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 1

[3] 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. 2

[4] Zhe Chen, Jiannan Wu, Wenhai 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, 5

[5] 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. 2

[6] 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. 2

[7] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 2

[8] 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. 2

[9] 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. 2

[10] 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. 2

[11] 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. 2

[12] 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. 1

[13] 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. 2

[14] 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. 2

[15] 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. 3, 4, 5

[16] 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. 2

[17] 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 crowdsourced data. *Sensors*, 20(6):1594, 2020. 2

[18] 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. 2

[19] 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. 2

[20] 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. 2

[21] 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. 2

[22] 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. 2

[23] OpenAI. Gpt-4 technical report. *arXiv preprint*, abs/2303.08774, 2023. Available at https://doi.org/10.48550/arXiv.2303.08774. 3, 5

[24] 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. 2

[25] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 2

[26] 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. 2

[27] 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. 2

[28] 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 sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 2

[29] 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. 2

[30] 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. 2

[31] 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. 2

[32] 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. 1

[33] 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. 2

[34] 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. 2

[35] 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. 2

[36] 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. 2

[37] 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. 2, 4

[38] 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. 2

[39] 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. 2

[40] 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. 2

[41] 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. 2

[42] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, et al. So2sat lcz42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019. 2