GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks

Supplementary Material

This supplementary material includes the dataset table (S1), dataset verification details (S2), Additional comparisons (S3) and quantitative results to illustrate multiple cases for assessing model responses (S4). It also provides results on multispectral images (S5) along with a comparison between bi-temporal and multi-temporal approaches (S6). Additionally, a geographical analysis is included (S7) to observe the span of data coverage across locations, followed by a detailed description of the word cloud (S8).

S1. Datasets

The datasets we use in our evaluation cover a wide range of geospatial tasks, showing the variety and depth of challenges in geospatial analysis. As shown in Table A1, these datasets include tasks like scene understanding, spatial relation, instance counting, temporal understanding, referring expression segmentation, and working with non-optical data. This diversity enables us to create versatile question-answer pairs tailored to each specific task. The inclusion of datasets from recent years ensures that our evaluation tackles recent challenges and uses up-to-date information.

These datasets also offer a rich variety of annotation types, sensor data, and spatial resolutions, reflecting the diverse nature of geospatial data. The annotation types range from class labels and bounding boxes to semantic and instance masks, giving different levels of detail for model evaluation. The sensor data includes RGB images, Multispectral Imaging (MSI), and Synthetic Aperture Radar (SAR), with resolutions from fine to coarse scales. This heterogeneity allows us to test models under different imaging conditions and resolutions, fostering robustness and generalizability. For example, datasets like FAIR1M [21], DIOR [5], and DOTA [24] provide high-resolution RGB images with bounding box annotations, which are critical for tasks like object detection and understanding spatial relationships in complex scenes. Temporal understanding datasets such as fMoW [10], xBD[9], PASTIS[16], FPCD[22], and GVLM[26] are crucial for tracking changes over time, helping in tasks such as disaster assessment and monitoring urban development. Non-optical datasets like So2Sat [28] and QuakeSet [15] introduce SAR data, expanding our analysis to situations where optical imagery isn't available due to weather or lighting conditions. Scene Understanding datasets like AiRound [13] and RESICS45 [4] offer class annotations that help categorize large-scale scenes, essential for land use and land cover classification.

S2. Dataset Verification Details

The benchmark was developed using a structured pipeline with semi-automated curation and manual verification. Two annotators manually annotated spatial relationships, and two others handled validation. All tasks were based on referenced datasets and cross-validated by all four. GPT assisted in drafting QA pairs, but all samples were manually verified. For captioning, we provided spatial and annotated inputs to generate fine-grained descriptions. In total, 750 human hours were spent on data curation, manual spatial-relation annotation, validation, and refinement, covering issues such as ambiguity, hallucinations, and distractor quality.

S3. Additional Results

We include two additional result, first, we report the (Fig. A1 left) performance of Claude-Sonnet 3.7 on classification tasks . The model shows strong performance in several categories, including Scene Classification (0.76), Spatial Relationship (0.66), and Water Bodies Counting (0.55), with an overall classification average of 0.2957 across 15 tasks. Second, to provide a more informative measure of performance on counting tasks, we compute mean absolute error (MAE) across all models (Fig. A1 right). LLaVA-OneVision and GPT-4o show lower MAEs (63.6 and 69.5 respectively), reflecting more precise quantitative estimation than what is captured by categorical accuracy alone. Including MAE offers a finer-grained understanding of model behavior, especially in tasks involving numerical reasoning.

S4. Qualitative Results

The images in Fig. A2 show patterns in how the models performed on geospatial tasks relevant to scene understanding. Models perform well in identifying scenes with distinct features, such as "interchange", where most models succeeded except Ferret [25], RS-LLaVA [3], and Sphinx [11], respectively. In the third image, all models except Ferret correctly identified the "stadium", demonstrating notable contextual understanding. For the fourth image, only a few models correctly identified "mixed cereal" crops, with failures attributed to the ambiguous nature of crop patterns. The first image in the second row shows dense greenery, indicating a moist environment, with the fire risk correctly classified as "low". The second image in the second row benefits from clear context, aiding classification as a water treatment facility. In contrast, the third image in the same row lacks context, making it prone to misclassification. This compar-

Name	Task	Annotation Type	Sensor (Res)	Year
AiRound[13]			RGB, Sentinel-2 (10m)	2020
RESICS45[4]			RGB	2017
PatternNet[27]	Scene Understanding, Object Classification	Class	RGB	2018
MtSCCD[12]			RGB (1m)	2024
FireRisk[18]			RGB (1m)	2023
FGSCR[8]			RGB	2021
FAIR1M[21]	Spatial Relation Classification, Referring Expression Detection, Captioning	Bounding Box	RGB (0.3–0.8m)	2021
DIOR[5]			RGB	2020
DOTA[24]			RGB (0.1–1m)	2021
Forest Damage[1]			RGB	2021
Deforestation[6]			RGB	2024
COWC[14]	Counting	Bounding Box	RGB (15 cm)	2016
NASA Marine Debris[17]			RGB (3m)	2024
The RarePlanes Dataset[19]			RGB (0.3m)	2020
fMoW[10]		Class	RGB (1m)	2018
xBD[9]	Temporal Understanding	Bounding Box, Instance Mask, Class	RGB (0.8m)	2019
PASTIS[16]		Semantic Mask	MSI (10m)	2021
FPCD[22]		Semantic Mask	RGB (1m)	2022
GVLM[26]		Class	RGB (0.6m)	2023
DeepGlobe Land Cover[7]	Referring Expression Segmentation	Semantic Mask	RGB (0.5m)	2018
GeoNRW[2]			RGB (1m)	2021
So2Sat[28]	Non-Optical	Class	SAR, MSI (10m)	2020
QuakeSet[15]		Number	SAR (10m)	2024

Table A1. Comprehensive overview of geospatial datasets utilized for evaluating Vision-Language Models (VLMs) across diverse tasks, including Scene Understanding, Spatial Relation, Object Classification, Spatial Relation Classification, Referring Expression Detection, Captioning, Temporal Understanding, Referring Expression Segmentation, and Non-Optical tasks. The datasets are categorized by annotation types (e.g., class, bounding box, semantic mask) and sensor types (e.g., RGB imagery, Multispectral Imaging (MSI), Synthetic Aperture Radar (SAR)), highlighting their versatility for a wide range of geospatial applications.

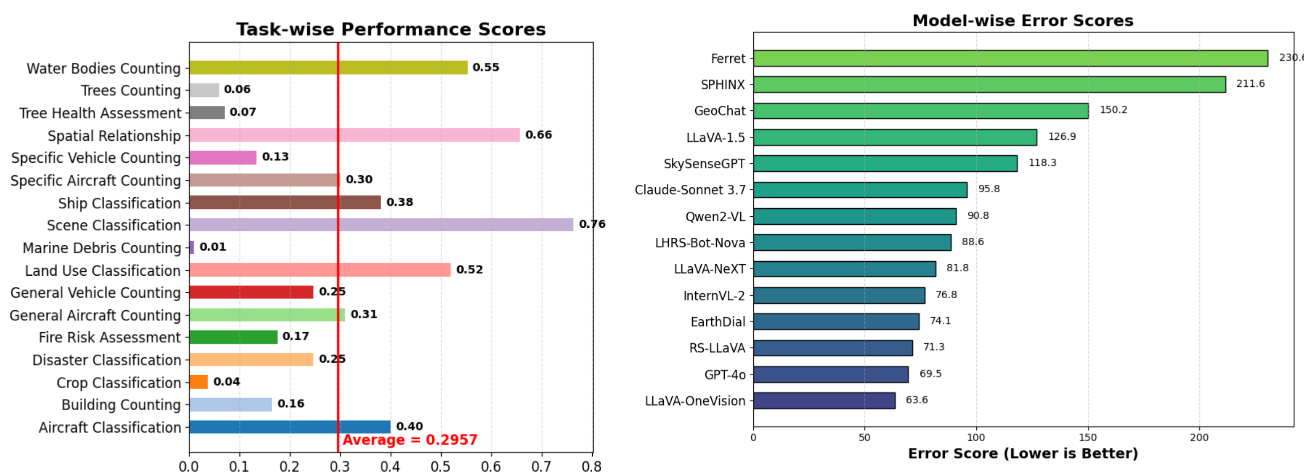


Figure A1. **Additional Comparisons.** (*left*) Claude-Sonnet 3.7 task-wise performance scores. (*right*) Model-wise error scores (lower is better).

ison highlights the importance of contextual information for accurate scene classification. In the last image of Fig. A2, ambiguous scenes such as the "ferry terminal" where all models except EarthDial [20] failed, the misclassification is likely due to overlapping visual cues. The visual similarity between a harbor and a ferry terminal makes it challenging for models to differentiate between these categories. For the counting tasks in Fig.A3, almost all models struggled, with wrongly estimating due to difficulty in differentiating objects in complex environments.

Fig. A4 shows that the models performed well in the first two images, probably due to familiar contextual clues. The “atago-class destroyer” and “small civil transport/utility” aircraft are common object types with distinct characteristics, making them easier for models to recognize. However, in the last two images, none of the models successfully identified the “murasame-class destroyer” or “garibaldi aircraft carrier” which are rarer categories. The failure is likely due to insufficient exposure to these specific classes in training datasets, coupled with the overlapping features of the objects that require advanced fine-grained recognition.

As shown in Fig. A5, models performed well on disaster assessment with relatively clear indicators, such as “fire” damage. For the second image, depicting “flooding”, Ferret and LHRS-Bot-Nova struggled. The third image depicts “tsunami” damage, characterized by disrupted layouts, scattered debris, and damaged buildings, which are often visually similar to flooding. Models may misclassify this due to overlapping features, and insufficient tsunami-specific training data. For the last image, only Qwen2VL identified the “seismic activity”, as others likely misclassified it due to overlapping features with “precipitation-related events”.

In Fig. A6, for the first image, a few models performed well because the objects are close to each other, easy to identify, and have minimal visual complexity. In the remaining images, models struggled because the objects were farther apart, making it harder to identify their spatial relationships. The cluttered environments and larger spatial gaps made it difficult for the models to accurately understand the relationships between the objects. Fig. A9 shows an aerial image alongside its ground truth caption and responses from various models. The ground truth provides a detailed and accurate description of the scene, while the model generated captions vary in capturing key elements such as urban & natural features, pathways, and architectural structures. This comparison highlights differences in model responses for image captioning tasks.

S5. Multi-spectral

In this section, we compare how GPT-4o and Qwen2-VL [23] perform on Crop Type Classification and Land Use Classification tasks using RGB and multispectral (MS) data (Fig. A7). The models perform much better with RGB inputs because they are designed and trained specifically for RGB images. The accuracy drops significantly for multispectral data, especially in crop-type classification. To use MS data with these models, Sentinel-2 bands were combined into three channels sequentially to mimic RGB inputs. For land use classification, which depends more on spatial patterns than detailed spectral information, the drop in performance is smaller. These results show the need for improved methods to adapt MS data for such tasks.

S6. Bi-temporal vs. Multi-temporal

We compare bi-temporal and multi-temporal image classification performance for Crop Type and Land Use classification tasks (Fig. A8). Multi-temporal data outperforms bi-temporal data for land use classification, suggesting that more timestamps are sufficient to capture key temporal changes. For crop classification, multi-temporal inputs, reflect the improvement in GPT-4o and LLaVA-OneVision.

S7. Geographical Analysis

In this section, we detail the geographic distribution of benchmarking datasets used in the studied geospatial tasks. It categorizes datasets into global/diverse datasets and regional/localized datasets. Global datasets provide extensive coverage with samples from over 100 countries or diverse regions worldwide. On the other hand, regional and localized datasets, are tailored to specific tasks. The map in Fig. A10 highlights that our benchmark is well represented across the globe.

S8. Word Cloud

The breakdown in Fig. A11a leverages the word cloud as part of evaluating VLMs in geospatial tasks, with image captioning being one of the key areas of interest. The word cloud highlights terms commonly used in captions describing aerial or geospatial imagery. Words like “aerial”, “surrounding” and “residential” reflect spatial and contextual elements frequently addressed in such descriptions, while terms such as “harbor”, “ship”, “tennis court” and “greenery” represent specific features often observed in geospatial data. This provides a basis for understanding the capabilities and limitations of these models in capturing spatial relationships and identifying key features within geospatial tasks.

The word cloud in Fig. A11b shows the terms used in MCQs. Keywords such as “large vehicle”, “transport utility”, “harbor”, “bridge” and “small vehicle” emphasize important categories and features frequently mentioned in the questions. Additional terms like “aircraft carrier”, “runway”, “basketball court” and “helicopter” represent a blend of transportation, infrastructure, and activity-based elements often linked to geospatial data. The use of varied and domain-specific vocabulary ensures the MCQs encompass a broad spectrum of scenarios for testing model capabilities.

References

[1] Swedish Forest Agency. Forest damages – larch casebearer 1.0, 2021. National Forest Data Lab. Dataset. 2

[2] Gerald Baier, Antonin Deschemps, Michael Schmitt, and Naoto Yokoya. Synthesizing optical and sar imagery from land cover maps and auxiliary raster data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 2

[3] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9): 1477, 2024. 1

[4] 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. 1, 2

[5] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *CoRR*, abs/2110.01931, 2021. 1, 2

[6] CSE499DeforestationSatellite. Deforestation-satellite-imagery dataset. `https://universe.roboflow.com/cse499deforestationsatellite/deforestation-satellite-imagery-335n4`. 2

[7] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *CoRR*, abs/1805.06561, 2018. 2

[8] Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4), 2021. 2

[9] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric T. Heim, Howie Choset, and Matthew E. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *CoRR*, abs/1911.09296, 2019. 1, 2

[10] Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, Juan M. Lavista Ferres, and Jennifer Marcus. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation, 2024. 1, 2

[11] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 1

[12] Jinglei Liu, Weixun Zhou, Haiyan Guan, and Wenzhi Zhao. Similarity learning for land use scene-level change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:6501–6513, 2024. 2

[13] Gabriel L. S. Machado, Edemir Ferreira, Keiller Nogueira, Hugo N. Oliveira, Pedro H. T. Gama, and Jefersson A. dos Santos. Airound and cv-brct: Novel multi-view datasets for scene classification. *CoRR*, abs/2008.01133, 2020. 1, 2

[14] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision – ECCV 2016*, pages 785–800, Cham, 2016. Springer International Publishing. 2

[15] Daniele Rege Cambrin and Paolo Garza. Quakeset: A dataset and low-resource models to monitor earthquakes through sentinel-1. *Proceedings of the International ISCRAM Conference*, 2024. 1, 2

[16] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *ICCV*, 2021. 1, 2

[17] A. Shah, L. Thomas, and M. Maskey. Marine debris dataset for object detection in planetscope imagery, 2021. 2

[18] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning, 2023. 2

[19] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes dataset, 2020. 2

[20] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. *arXiv preprint arXiv:2412.15190*, 2024. 2

[21] Xian Sun, Peijin Wang, Zhiyuan Yan, F. Xu, Ruiping Wang, W. Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, M. Weinmann, S. Hinz, Cheng Wang, and K. Fu. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *Isprs Journal of Photogrammetry and Remote Sensing*, 2021. 1, 2

[22] Chintan Tundia, Rajiv Kumar, Om Damani, and G. Sivakumar. Fpcd: An open aerial vhr dataset for farm pond change detection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 862–869. SCITEPRESS - Science and Technology Publications, 2023. 1, 2

[23] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[24] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1, 2

[25] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1

[26] Xiaokang Zhang, Weikang Yu, Man-On Pun, and Wenzhong Shi. Cross-domain landslide mapping from large-scale re-

mote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:1–17, 2023. 1, 2

[27] Weixun Zhou, Shawn D. Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *CoRR*, abs/1706.03424, 2017. 2

[28] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. 1, 2

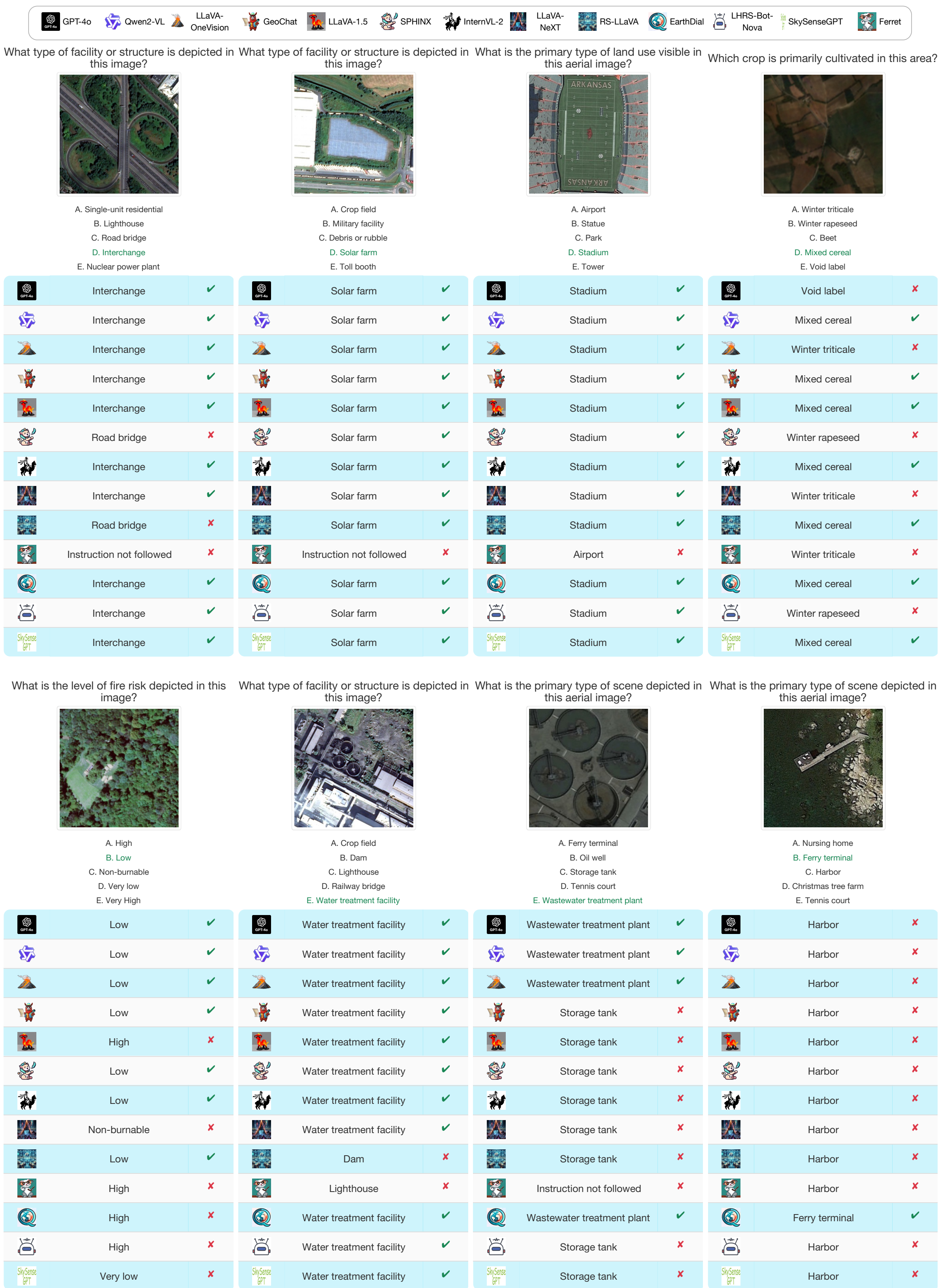


Figure A2. Scene Understanding: This illustrates model performance on geospatial scene understanding tasks, highlighting successes in clear contexts and challenges in ambiguous scenes. The results emphasize the importance of contextual reasoning and addressing overlapping visual cues for accurate classification.

Figure A3. Counting: The figure showcases model performance on counting tasks.

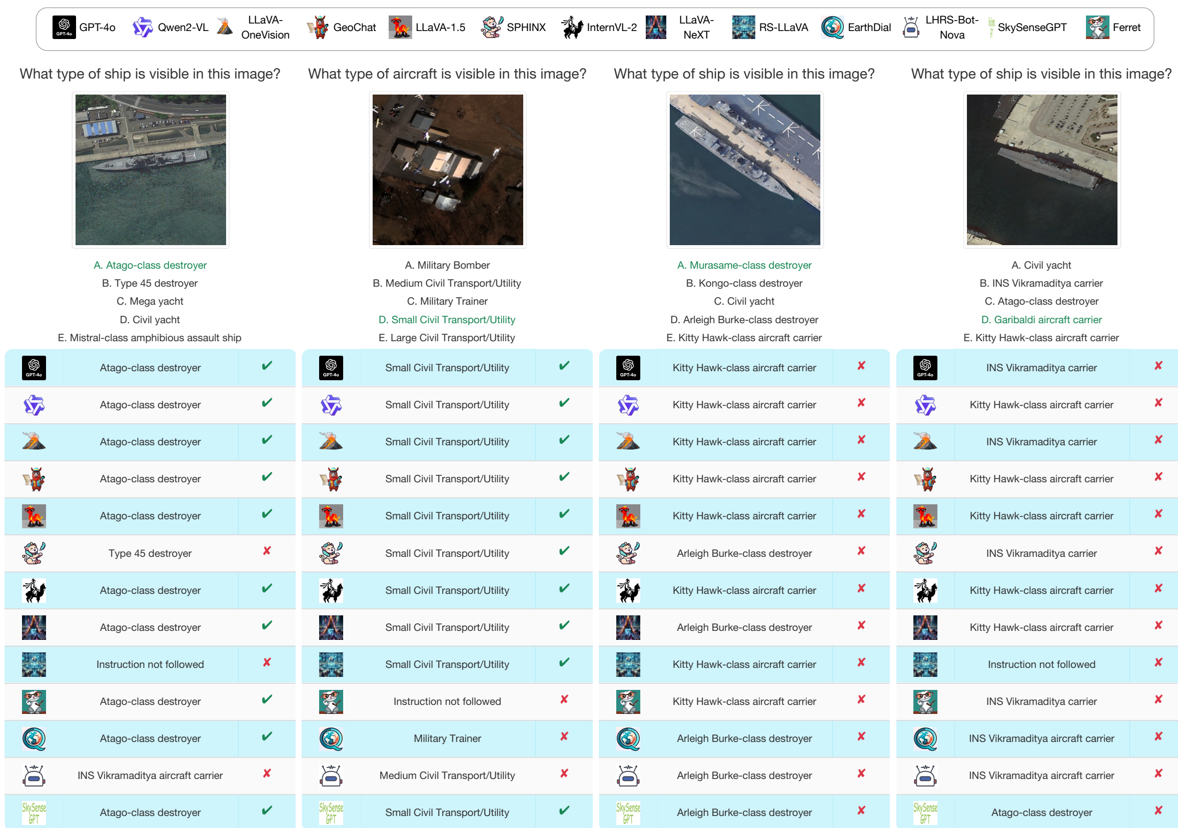


Figure A4. Object Classification: The figure highlights model performance on object classification, showing success with familiar objects like the "atago-class destroyer" and "small civil transport/utility" aircraft. However, models struggled with rarer objects like the "murasame-class destroyer" and "garibaldi aircraft carrier" indicating a need for improvement on less common classes and fine-grained recognition.

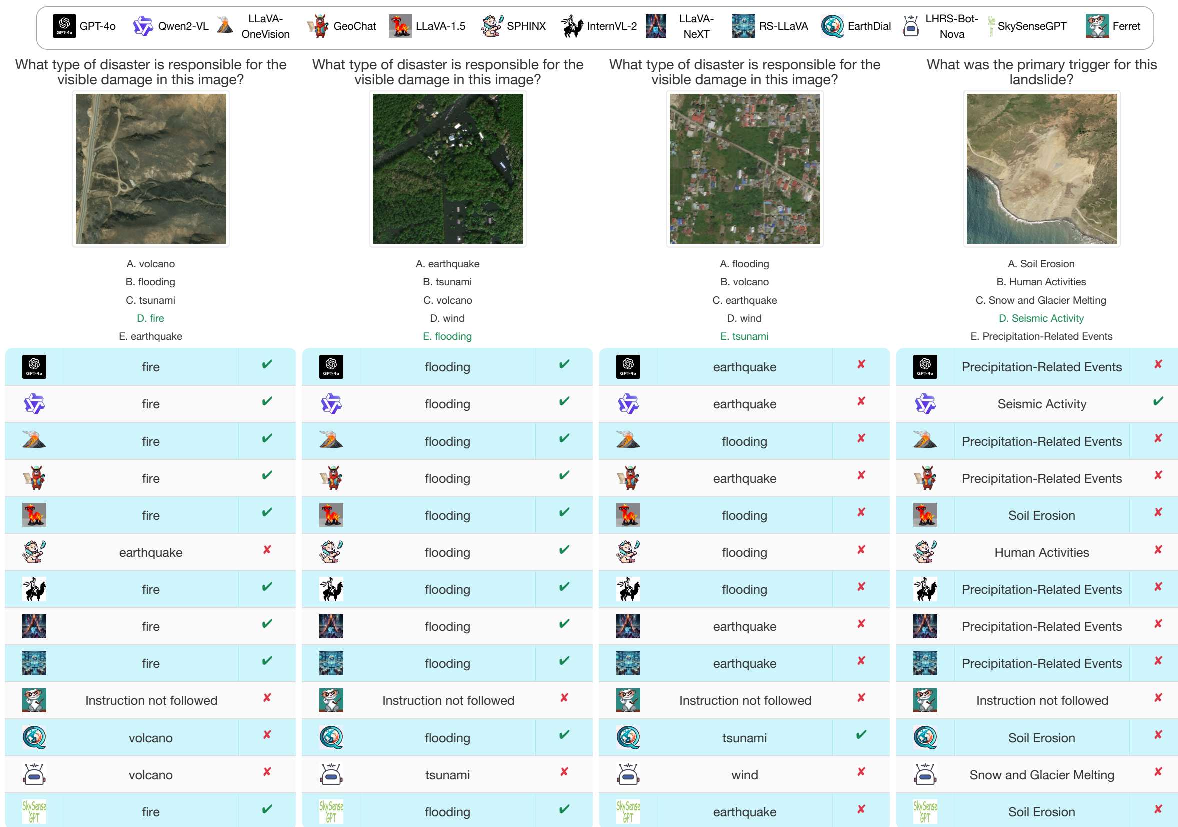


Figure A5. Event Detection: Model performance on disaster assessment tasks, with success in scenarios like 'fire' and 'flooding' but challenges in ambiguous cases like 'tsunami' and 'seismic activity'. Misclassifications highlight limitations in contextual reasoning and insufficient exposure on overlapping disaster features.

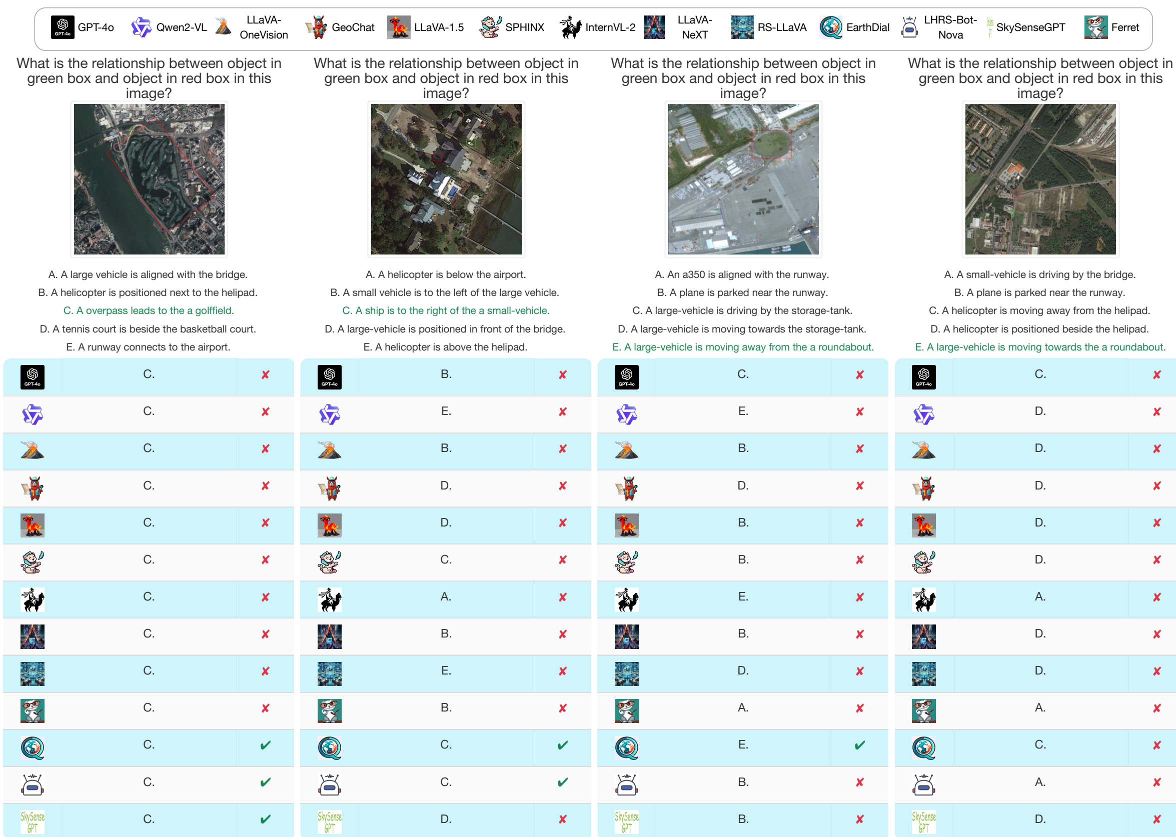


Figure A6. Spatial Relations: The figure demonstrates model performance on spatial relationship tasks, with success in close-object scenarios and struggles in cluttered environments with distant objects.

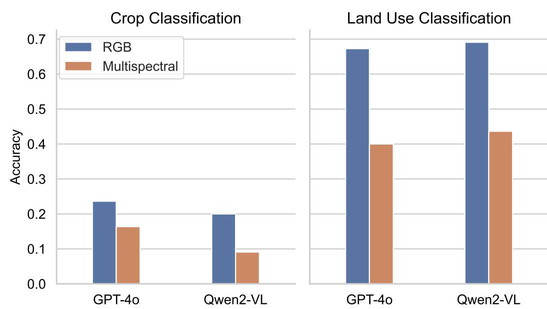


Figure A7. It compares RGB and multispectral performance for Crop Type and Land Use Classification, showing a performance drop in multispectral accuracy.

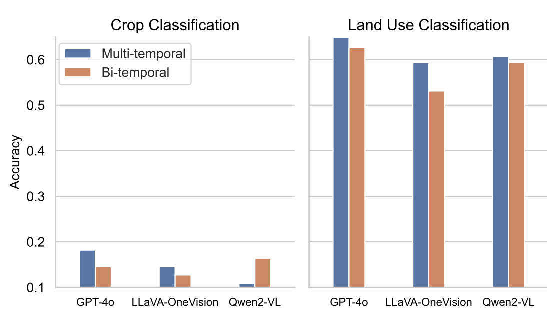


Figure A8. Bi-temporal and Multi-temporal performance

Ground Truth GPT-4o Qwen2-VL LLaVA-OneVision GeoChat LLaVA-1.5 SPHINX LLaVA-NeXT InternVL-2 Ferret

The aerial image showcases a blend of urban and natural elements. In the top left, a large building with an adjacent parking lot is visible, where several vehicles, including vans and small cars, are parked, suggesting a bustling area. Nearby, a landscaped park features winding paths and lush greenery, creating a serene environment. The center of the park includes open grassy areas with neatly arranged seating, ideal for gatherings or outdoor events. To the right, a building with a green roof is situated next to recreational courts, enhancing the communal functionality of the space. At the bottom, a road borders the park, leading to a larger structure with curved architectural features, resembling an outdoor stadium or amphitheater. The organized layout of paths and landscaped sections throughout the park highlights a thoughtful integration of natural beauty and urban design.

GPT-4o

The image is an aerial view of a landscaped area and surrounding structures, likely a botanical garden or park. Key features in this image include: 1. **Pathways**: There are several winding pathways throughout the landscaped area, designed for walking and exploration. 2. **Gardens and Green Spaces**: The area is filled with various patches of greenery, including lawns and planted garden sections. These are likely filled with shrubs, trees, and other plants, although specific types cannot be discerned from this height.3. **Amphitheater**: A small, semicircular set of tiered seating suggests the presence of an outdoor amphitheater, often used for events or performances.4. **Building and Structures**: - There is a building with a turquoise roof, possibly a pavilion or visitor center. - Additional buildings are located nearby, one with a black roof and another with a white roof. **Adjacent Parking**: Several parking lots are visible around the edges of the landscaped area, indicating accessibility for visitors arriving by car.6. **Nearby Roads**: A road runs along the edge of the property, providing access to the park and parking areas.7. **Sports Field**: There appears to be a marked-off area, likely a small field or court, suggesting sports or recreational activities. Overall, this image depicts a well-maintained recreational or botanical space, designed for public use and enjoyment.

The image shows an aerial view of a park or garden area situated in an urban environment. Below is a detailed description: Park Layout: The park is encircled by buildings and roads, emphasizing its location within an urban setting. At its center, the park features a large, open grassy field that serves as the primary green space. A network of pathways weaves through the park, connecting various sections and adding structure to the layout.
Pathways: The pathways are distinctly visible and seem to be constructed from concrete or a similar material, ensuring durability and accessibility. They are strategically designed to guide visitors throughout the park, enhancing the overall navigability and user experience.

The image is a bird's eye view of a large park with a variety of landscaping elements. The park features a large swimming pool, surrounded by trees and grass. There are several benches scattered throughout the park, providing seating for visitors.
In addition to the swimming pool, the park also includes a tennis court, a soccer field, and a basketball court. The park is well-maintained and offers a pleasant environment for outdoor activities.
The parking lot is located near the park, with several cars parked in the vicinity. The overall scene is a lively and inviting outdoor space for people to enjoy.

The image is an aerial photograph of a campus or park-like setting. The landscape is characterized by a mix of grassy areas, pathways, and parking lots. There are several trees scattered throughout the area, providing a sense of greenery. The paths are curved and meander through the landscape, suggesting a pedestrian-friendly environment. In the foreground, there is a large building with a flat roof, which could be a sports facility or a similar structure. Adjacent to this building, there is a parking lot with several cars parked. The building and parking lot are surrounded by grassy areas and pathways.In the middle ground, there is a large open space with a patterned surface, which could be a tennis court or a similar outdoor recreational area. The court is bordered by a pathway and is surrounded by grass.

In the satellite image, there are <p>two olive small cars</p> {<10><29><10><33>|<60>}<delim>{<10><32><10><36>|<63>} and <p>a white small car</p> {<10><35><10><39>|<63>} parked near each other on the left side. On the right side, there are <p>two gray small cars</p> {<102><109><100><100>|<56>}<delim>{<102><100><100><100>|<56>} parked close to each other. In the satellite image, there are two olive small cars {<10><29><10><33>|<60>}{<10><32><10><36>|<63>} and a white small car {<10><35><10><39>|<63>} parked near each other on the left side. On the right side, there are two gray small cars {<102><109><100> <100>|<56>}{<102><100><100><100>|<56>} parked close to each other.

The image is an aerial view of a park-like area with various features. The park has several walking paths that meander through the greenery, creating a network of trails for visitors to explore. There are multiple grassy areas interspersed with trees and shrubs, providing a natural and serene environment. In the center of the park, there is a large open space with a few benches arranged in a semi-circle, possibly for gatherings or performances. To the right of this central area, there is a building with a green roof, which could be a pavilion or a facility related to the park's operations. To the left of the park, there is a parking lot with several cars parked in it. Adjacent to the parking lot, there is a road running parallel to the park's boundary. On the far right side of the image, there is a circular structure that appears to be a sports field or an amphitheater, with rows of seating arranged in a semi-circular pattern. Overall, the image depicts a well-maintained public park with various amenities for visitors to enjoy, including walking paths, seating areas, and facilities for events or sports activities.

This image appears to show a satellite or aerial view of a large, paved area surrounded by greenery and a body of water. The image also includes a residential or commercial building and structures associated with a golf course or public recreational area. **Objects in the Image**: 1. **Golf Course:** - A prominent feature in the image is the golf course, located toward the center. The layout visible includes several holes and paths intertwined with tall grass and trees. - There is a specific structure resembling a tee area or a part of the golf course infrastructure, usually where tees are placed at the start of each hole. - Surrounding the golf course, there are smaller patches of trees, bushes, and some sand bunkers.2. **Bodies of Water:** - There is a large water body placed adjacent to the bottom right corner of the image and near the bottom left section. The presence of this body of water indicates that water hazards could possibly be part of the golf course layout.3. **Paved Areas:** - The bottom left part of the image features a large paved section which appears to be a parking lot. Cars are visible, indicating they are parked. - Next to this parking lot, there is a building with a distinctive roof structure that is likely a clubhouse or another recreational facility associated with the golf course.4. **Residential or Commercial Building:** - On the left side of the image, there is a large grey building with a sloped roof. This building appears to be a residential or commercial structure.5. **Green Space and Landscaping:** - The area between the golf course and the roads is landscaped with numerous trees, shrubs, large patches of grass, and a well-maintained garden bed. The landscaped areas suggest a maintenance of nature and natural environments typical of a good golf course or park.6. **Numerous Pathways and Walkways:** - These paths seem designed to provide convenient walkways for patrons to move around and enjoy the green spaces. - There are sidewalks meandering through the green areas, providing easy access to different parts of the facility.7. **Parking Lot:** - The parking lot adjacent to the clubhouse/pavement area has clearly marked spots for cars, indicating organized and possibly private parking for visitors.8. **Adjacent Road and Sidewalks:** - Adjacent to the parking lot, there is a street with sidewalks clearly visible. These roads are likely part of the surrounding urban infrastructure.9. **Other Structures and Buildings:** - Along the edges of the golf course and pavement, scattered within the natural terrain, are various small buildings and possibly maintenance structures.10. **Vegetation:** - Extensive tree cover is visible around the golf course and the pavement areas, indicating this might be a region with varied vegetation types.

The image features a bird's eye view [23, 20, 966, 966] of a campus [23, 20, 966, 966] with various landscaping elements [198, 471, 356, 742]. The campus includes a large green park with a pond [500, 201, 572, 253] [485, 187, 583, 271] located near the center of the landscape. Around the park, there are several buildings, possibly part of a university or other educational institution. In the surrounding area, numerous cars can be seen parked or moving along the edges of the campus. There are at least 13 cars of different sizes, with some closer to the buildings and others near the park. Additionally, a couple of benches [630, 499, 691, 554] can be found in the scene, providing seating options for those visiting the park or resting near the buildings.

The image showcases a large, well-maintained park with a circular pathway that winds through it. The pathway is surrounded by a variety of greenery, including trees and grass, creating a serene and inviting atmosphere. The park is situated in a residential area, with a parking lot nearby.There are several cars parked in the area, indicating that the park is accessible to visitors. In addition to the cars, there are a few trucks parked in the vicinity, suggesting that the park might be used for recreational activities or event.

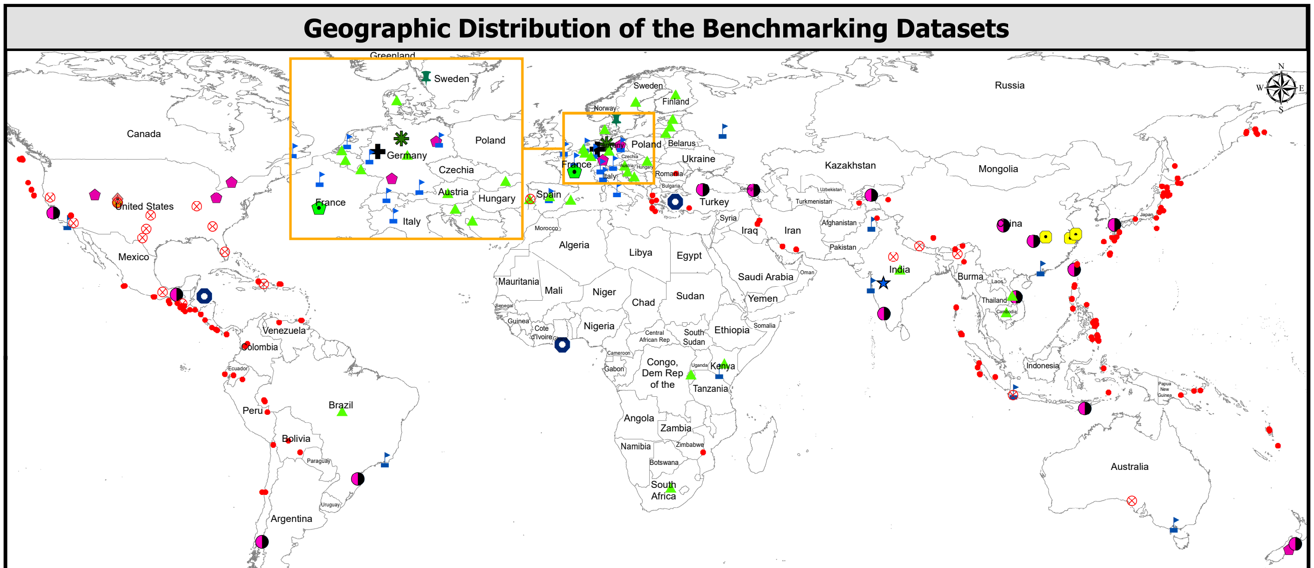
SkySense GPT

The scene depicts a park with a spacious green field and a few trees. In the park, there are multiple cars parked in different areas. One car is parked near the edge of the field, while another car is parked closer to the trees. Additionally, there is a car parked near the center of the field. The cars are positioned in a way that suggests they are parked in designated parking areas within the park. The presence of these cars indicates that visitors are enjoying the park and its facilities.

The image shows a park with a large green lawn, surrounded by trees and buildings. There is a parking lot nearby, and a road runs through the park. The park has a tennis court, a basketball court, and a playground. The buildings in the background have different colors, including white, gray, and brown.

In the satellite image, there are some buildings [[0, 10, 21, 22, 90]] located at the top left corner of the scene. At the bottom right, there are some buildings [[70, 78, 82, 100, 90]] with trees surrounding them.

Figure A9. Image Captioning: Example response of different models.



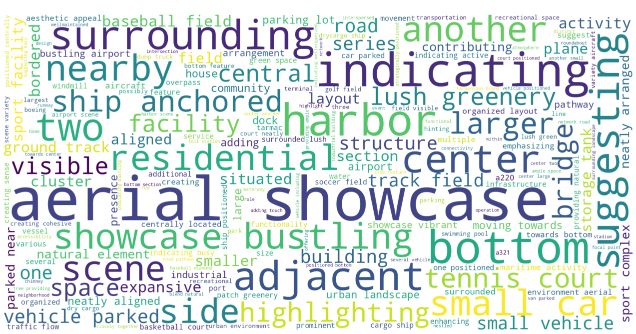
Global/Diverse Datasets

Name	Description	Name	Description
FAIR1M	Airports,harbors,cities across continents	resisc45	Data from > 100 countries worldwide
DIOR	Object detection from global regions	FGSCR	Images of 40 ports around the world
DOTA	Urban areas globally high-res images	fMoW	Diverse images from > 200 countries
RarePlanes	Real+synthetic locations more than 120	DeepGlobe	Land cover classification data globally
AiRound	Samples cover diverse regions globally		

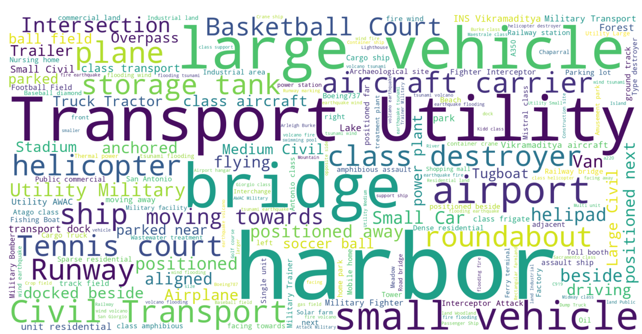
Regional or Localized Datasets

QuakeSet	PASTIS	MtSCCD_LUSC	COWC
xBD	FPCD	Forest Damage	Patternnet
FireRisk	so2sat_eval	TreeSatAI	MarineDebris
FieldsofTheWorld	GeoNRW	GVLM	

Figure A10. Figure shows the geographic distribution of benchmarking datasets, highlighting global coverage and regional specialization.



(a) Word cloud from captions used in geospatial image description tasks.



(b) Word cloud from MCQs designed for geospatial task evaluation.

Figure A11. Word clouds showcasing terms used in evaluating VLMs on geospatial tasks, with the first focusing on image captions and the second on MCQs.