A. Additional Overview of Our Benchmark

Benchmark Creation. To build an EO benchmark for VLMs, we focus on three broad categories of capabilities in our initial release: scene understanding, localization and counting, and change detection. Within each category, we construct evaluations based on applications ranging from animal conservation to urban monitoring (Figure 2). Our goals are to (1) evaluate the performance of existing VLMs, (2) provide insights into prompting techniques suitable for repurposing existing VLMs to EO tasks, and (3) implement an interface of data and models for flexible benchmark updates and evaluations of future VLMs. Our categories and tasks are:

- *Scene Understanding*: To evaluate how VLMs combine high-level information extracted from images with latent knowledge learned through language modeling, we construct three datasets: (1) a new aerial landmark recognition dataset to test the model's ability to recognize and geolocate landmarks in the United States; (2) the RSICD dataset [27] to evaluate the model's ability to generate open-ended captions for Google Earth images; (3) the BigEarthNet dataset [39] to probe the model's ability to identify land cover types in medium-resolution satellite images, and (4) the fMoW-WILDS [6] and PatternNet [59] datasets to assess the model's ability to classify land use in high-resolution satellite images.
- *Localization & Counting*: To evaluate whether VLMs can extract fine-grained information about a specific object and understand its spatial relationship to other objects, we assemble three datasets: (1) the DIOR-RSVG dataset [55] to assess Referring Expression Comprehension (REC) abilities, in which the model is required to localize objects based on their natural language descriptions; (2) the NEON-Tree [46], COWC [29], and xBD [13] datasets to assess counting small objects like cluttered trees, cars, and buildings in aerial and satellite images; (3) the aerial animal detection dataset [9] to gauge counting animal populations from tilted aerial images taken by handheld cameras.
- *Change Detection*: To evaluate if VLMs can identify differences between multiple images and complete user-specified tasks based on such differences, we repurpose the xBD dataset [13]. We show the model two high-resolution images taken before and after a natural disaster and ask it to assign damaged buildings to qualitative descriptions of damage categories.

We note that a number of capabilities desired for EO data remain unattainable by current-generation VLMs due to their inability to ingest multi-spectral, non-optical, or multi-temporal images. This is unlikely to be addressed by the vision community while its focus remains on natural images. Furthermore, available VLMs do not yet perform image segmentation, although we expect this to change in the near future.

Model Selection. Following the existing knowledge benchmarks of instruction-following VLMs by Yue et al., we select five top-performing models at the time of our evaluation, including GPT-4V(ision) [51], InstructBLIP-FLAN-T5-xxl [8], InstructBLIP-Vicuna-13b [8], LLaVA-v1.5 [25], and Qwen-VL-Chat [3]. Among our selected models, GPT-4V is the most capable model in terms of training recipe, training dataset, and model weights, but it is a closed model. LLaVA connects text and image modalities through a simple linear layer and trains both the vision encoder and language decoder on their curated instruction fine-tuning dataset. InstructBLIP [8] uses an instruction-aware Q-Former to connect vision and language modalities and perform instruction fine-tuning on their curated datasets. Qwen-VL-Chat [3] uses a single-layer cross-attention module to connect the visual features from OpenCLIP ViT with the LLM backbone. Our selection represents state-of-the-art models widely adopted by the researchers and practitioners in VLMs.

Summary of Findings. Below, we summarize insights from our evaluations, with a focus on GPT-4V, as it is generally the best-performing VLM across Earth observation tasks. We elaborate on the results in Sections 2, 3, and 4. Based on our findings, we conclude that existing instruction-following VLMs are not prepared for applications in EO data involving fine-grained image understanding and reasoning.

- *Scene Understanding*:
  1. On our new aerial landmark recognition task, GPT-4V achieves an overall accuracy of 0.67 (Table 1), surpassing open models by a large margin and demonstrating its comprehensive world knowledge. There appear to be regional disparities, with GPT-4V generally performing better in coastal states. In addition, although GPT-4V sometimes generates sophisticated reasoning paths, the reasoning can be incorrect despite a correct final answer.
  2. On RSICD image captioning, GPT-4V achieves a RefCLIPScore of 0.75 (Table 2), which measures both image-text semantic similarity and caption-reference similarity. Although GPT-4V does not achieve high similarity between generated and reference captions, our qualitative assessment is that it generates even more detailed captions than the humans employed in RSICD.
  3. On land cover/land use classification tasks, GPT-4V performance varies depending on image resolution, label ambiguity, and label granularity. On fMoW-WILDS, the average F1-score is 0.19 (Table 3); on PatternNet, average F1-score is 0.71

(Table 4), and on BigEarthNet, average F1-score is 0.38 (Table 5). High performance on PatternNet can be attributed to high image resolution and disambiguated labels. Low performance on fMoW-WILDS is largely due to ambiguous labels, which we discuss in Section 2.3.

- *Localization & Counting*:

1. On DIOR-RSVG object localization, GPT-4V obtains a mean intersection-over-union (IoU) of 0.16; only 7.6% of the test images have an IoU > 0.5, while a model that specializes in outputting bounding boxes achieves a mean IoU of 0.68 (Table 6).
2. While GPT-4V achieves moderate accuracies on the COWC vehicle counting ($R^2 = 0.61$, Table 8) and xBD building counting ($R^2 = 0.68$, Table 10) tasks, it fails on NEON-Tree counting ($R^2 = 0.20$, Table 7) and aerial animal detection ($R^2 = 0.08$, Table 9).

- *Change Detection*: On xBD change detection, GPT-4V fails to count and categorize the damaged buildings, with $R^2 = 0.10$ for buildings in the "destroyed" category (Table 10). Although GPT-4V can count the number of buildings before a disaster with moderate accuracy, it systematically fails to assess the building damage by contrasting before and after images. This systematic failure makes it unusable for disaster relief applications that require counting abilities.

Recommended Usage. We conclude that existing instruction-following VLMs are not prepared for applications in EO data involving fine-grained image understanding and reasoning. Although they achieve remarkable quantitative and qualitative performance on high-level scene understanding tasks like landmark recognition, image captioning, and certain land use classification tasks, current VLMs fail to deliver satisfactory performance in bounding box generation, counting, and change detection tasks. Systematic efforts are still needed to improve the within-image spatial awareness and between-image change understanding, including but not limited to model architecture, pretraining methodologies, datasets, and alignment techniques.

A.1. Related Works

VLM Applications in EO. With the development of instruction-following VLMs on large-scale image-text datasets, a variety of works also try to explore their applications in EO data. For example, Hu et al. fine-tune INSTRUCTBLIP [8] on remote sensing image captioning tasks to improve the model's capabilities in VQA. Roberts et al. probe GPT-4V(ISION) to interpret geospatial information from natural and remote sensing images. Tan et al. prompted GPT-4V with sample questions from geography, environmental science, agriculture, and urban planning domains. In addition, Kuckreja et al., Zhan et al., Zhang et al., and Muhtar et al. create datasets to fine-tune instruction-following vision-language models on EO data. However, despite the potential of VLMs to make EO data analysis much more accessible, there have been very few comprehensive benchmarks to assess the capabilities of instruction-following VLMs on EO data quantitatively.

Benchmarks for VLMs. Given natural language instructions and images as a prompt, an instruction-following VLM performs user-specified tasks such as image classification, visual question answering (VQA), image captioning, object localization [50], semantic and instance segmentation [33], etc. Because they are trained on a large corpus of text and images, VLMs are successful at a wide array of applications, including manufacturing defect detection [51], radiology report generation [51], damage assessment in auto insurance [51], and animal species identification [10]. In order to measure the progress of VLMs, benchmarks like MMMU [54], SEED-Bench [20], MM-Vet [53], MM-Bench [26], and LLaVA-Bench [24, 25] have been proposed to assess scene understanding and visual reasoning on natural images.

B. Limitations & Future Work

While we try to provide a comprehensive evaluation of the capabilities of instruction-following VLMs on EO data, we acknowledge the following limitations in our benchmark:

- **Potential data contamination.** As the pretraining recipes for GPT-4V and certain open models remain obscure, it is almost impossible to determine whether the model was pretrained or fine-tuned on our evaluation data. As the community develops VLMs for EO data, data contamination detection techniques [37] might be needed to ensure the benchmark continues to be fair and effective.
- **Limited error analysis.** Although we have provided the reader with failure examples in this work, a more systematic analysis that categorizes the failure cases into lack of knowledge, incorrect reasoning, perceptual error, and textual misunderstanding would deepen our understanding of the capabilities of current VLMs.

- **Static nature of the benchmark.** Dynamic updates may be required to ensure the benchmark remains relevant and challenging as models become more capable. Future work could involve establishing a data engine for sourcing new test examples across tasks and creating tasks that evaluate newer VLMs with segmentation capabilities [33].

C. Data Sheet

We follow the documentation frameworks provided by Gebru et al. [11] to foster transparency and accountability about the datasets utilized in this benchmark.

C.1. Motivation

For what purpose was the dataset created?

- We create this collection this benchmark to evaluate the effectiveness of instruction-following Vision-Language Models (VLMs) in performing crucial tasks related to Earth Observation (EO) data, specifically satellite and aerial images. These types of images are not commonly found in the training data of existing VLMs, which has led to uncertainty about the models' capabilities in handling them. The benchmark aims to assess VLMs' proficiency in scene understanding, localization and counting, and change detection tasks. This is particularly relevant for real-world urban monitoring, disaster relief, land use, and conservation applications.
- For

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- This benchmark is created by Chenhui Zhang and Sherrie Wang of the Earth Intelligence Lab at the Massachusetts Institute of Technology.
- We also acknowledge the entities who create and maintain the source datasets used in our benchmark.
  - Location Recognition: National Agriculture Imagery Program (NAIP), by U.S. Department of Agriculture (USDA)
  - Image Captioning: RSICD [27] by the University of Chinese Academy of Sciences and the Chinese Academy of Sciences.
  - Land Use & Land Cover Classification: fMoW [6] by Johns Hopkins University and DigitalGlobe, fMoW-WILDS [17] by Stanford University, PatternNet [59] by Wuhan University and University of California, Merced, and BigEarthNet [39] by TU Berlin.
  - Localization: DIOR-RSVG [55] by the Northwestern Polytechnical University of China.
  - Counting: NeonTreeEvaluation [46] by Weecology, COWC [29] by Lawrence Livermore National Laboratory, aerial animal detection [9] by Wageningen University & Research, and xBD [13] by the Defense Innovation Unit of the Department of Defense and Carnegie Mellon University.
  - Change Detection: xBD [13] by the Defense Innovation Unit of the Department of Defense and Carnegie Mellon University.

C.2. Composition/collection process/preprocessing/cleaning/labeling and uses:

- The dataset construction process is described in our paper as well as website `https://vleo.danielz.ch/`.

C.3. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

- The Earth Intelligence Lab at MIT will manage and maintain our dataset.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

- The evaluation dataset is released to the public and hosted on Hugging Face.

When will the dataset be distributed?

- It has been released now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

- Our benchmark is distributed under the CC BY-SA 4.0 license.

C.4. Maintenance

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

- Please contact Chenhui Zhang (`chenhui5@mit.edu`) and Prof. Sherrie Wang (`sherwang@mit.edu`), who are responsible for maintenance.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

- Yes. If we include more tasks or find any errors, we will correct the dataset and update the results in the leaderboard accordingly. It will be updated on our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

- We greatly appreciate new contributions of datasets and evaluation scenarios from the community to keep this benchmark up-to-date. To contribute new scenarios, the most efficient way is to open an issue under our GitHub repository to request features and discuss your potential contribution plans. Then, we can initiate a pull request for your contributions.
- For dataset contributions and evaluation modifications, the most efficient way to reach us is via GitHub pull requests.
- For more questions, please contact Chenhui Zhang (`chenhui5@mit.edu`) and Prof. Sherrie Wang (`sherwang@mit.edu`), who will be responsible for maintenance.

D. Additional Details about Scene Understanding

The ability of a VLM to understand high-level features of the scene of a remotely sensed image is crucial for its application in EO data. Given an aerial or satellite image, an ideal instruction-following VLM should be able to parse the salient visual features of the input images(s) and utilize their world knowledge learned through language modeling to perform tasks specified by user instructions.

In this section, we delve into the scene-understanding capabilities of existing VLMs by assessing them under both open-ended tasks and multiple-choice questions about the scene. We first curate an aerial landmark recognition dataset based on high-resolution images from the National Agriculture Imagery Program (NAIP). Then, we assess the ability of VLMs on the image captioning task with the RSICD [27] dataset. Finally, we test the instruction-following VLMs on closed-ended tasks, including land cover and land use classification.

D.1. Additional Details about Location Recognition

The ability to recognize the location given a natural image has always been an interest of existing VLM benchmarks [20] as it reflects the ability of the model to connect visual cues to its world knowledge learned through pretraining. In addition, it provides a glimpse into their geospatial bias, which influences undesired behaviors like hallucination [7].

Dataset Construction. We filter and match the landmarks in the Google Landmarks dataset [48] with their OpenStreetMap polygons and filter for those located in the United States, resulting in 602 landmarks. Then, we obtain the latest high-resolution aerial images of the obtained polygons through the National Agriculture Imagery Program (NAIP) of the United States Department of Agriculture (USDA). Finally, we construct multiple-choice questions about the name of the landmark with incorrect answers from other landmarks in the same category. To give the reader a qualitative understanding of our curated dataset, we visualize the spatial distribution of the landmarks (Figure D.1). In addition, we classify the landmarks in our dataset based on their functions, and we summarize the median area and count of each functional class (Table D.1). We also showcase some example images of the landmarks in our dataset (Figure D.2).

The spatial distribution of the aerial landmarks dataset shows a concentrated presence of landmarks in the United States, with notable clusters along the East Coast, California, and other parts of the West Coast (Figure D.1). There is also a significant concentration in the Great Lakes region. The presence of landmarks is sparse in the central and mountain states. The dataset comprises a total of 602 landmarks, with the majority being Natural Parks and Reserves (294 landmarks), which also have the largest median area of 16.92 km^2 (Table D.1). This is followed by Historical and Cultural Sites (82 landmarks) with a median area of 1.652 km^2, and Sports and Entertainment Venues (90 landmarks) with a much smaller median area of 0.024 km^2. We visualize one landmark for each category (Figure D.2) and also perform an error analysis of GPT-4V by state (Figure D.4).

Table D.1. Statistics of the aerial landmark dataset

Category	Count	Median Area (km^2)
Natural Parks and Reserves	294	16.92
Sports and Entertainment Venues	90	0.024
Historical and Cultural Sites	82	1.652
Government and Public Buildings	58	0.154
Places of Worship	47	0.002
Infrastructure and Urban Features	26	0.3477
Miscellaneous	5	221.61
Total	602	2.490

System and Task Prompts. In Figure D.5, we present the system prompt we use throughout our location recognition experiments. This prompt introduces the context of the aerial landmark recognition task and tries to reduce the number of questions that the model refuses to answer. For VLMs without explicit design for system prompts, we add the same prompt before the user prompt. In Figure 4, we provide an example of our user prompt with example responses from the GPT-4V model. We formulate our user prompt as a multiple-choice question with additional instructions to let the model output the reasoning path that led to its answer. The correct choice is the name of the landmark sourced from the Google Landmarks

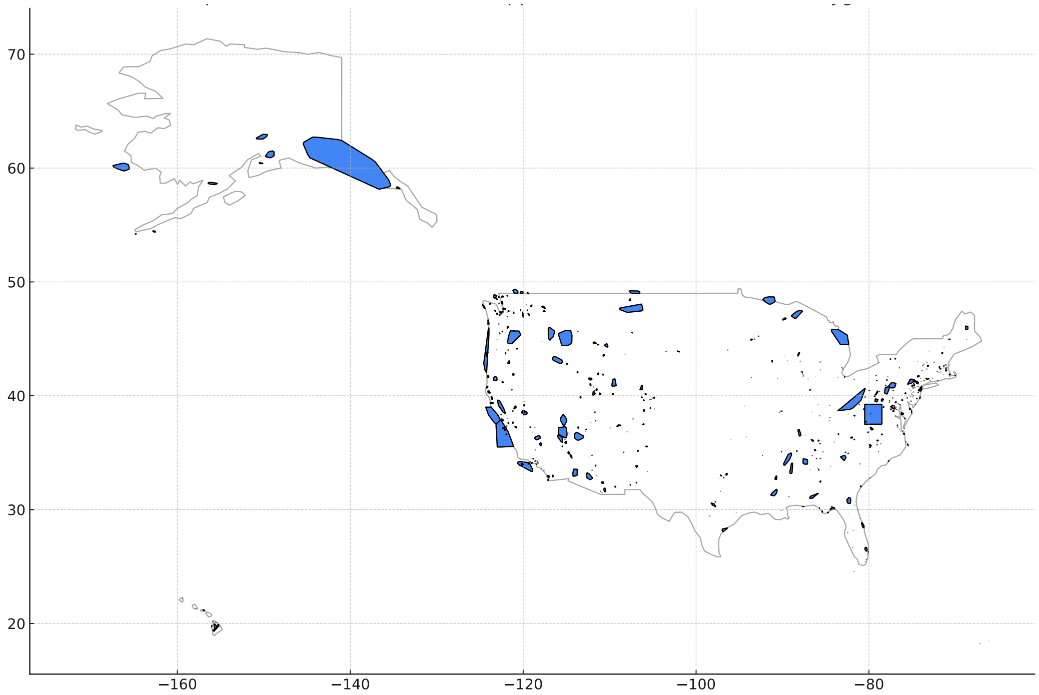
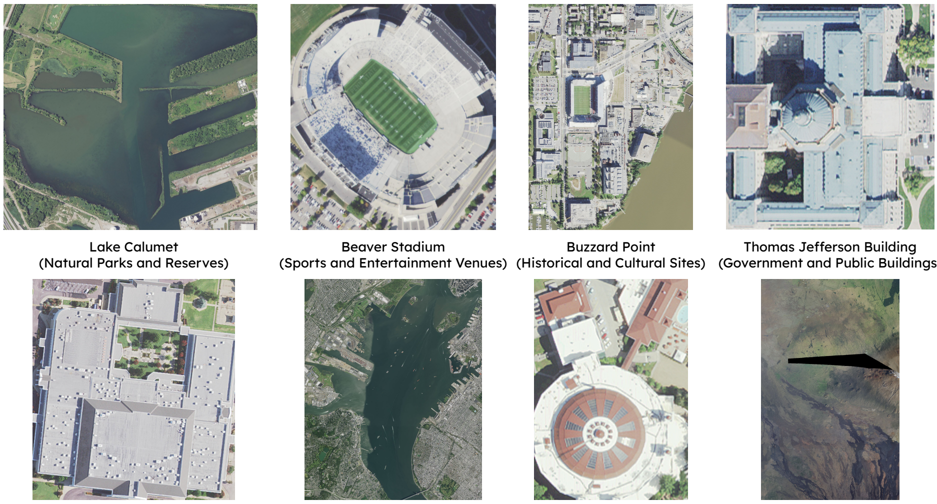


Figure D.1. Spatial distribution of our aerial landmarks dataset



Lake Calumet
(Natural Parks and Reserves)

Beaver Stadium
(Sports and Entertainment Venues)

Buzzard Point
(Historical and Cultural Sites)

Thomas Jefferson Building
(Government and Public Buildings)

Bellevue Baptist Church
(Places of Worship)

Upper New York Bay
(Infrastructure and Urban Features)

West Baden Springs Hotel
(Miscellaneous)

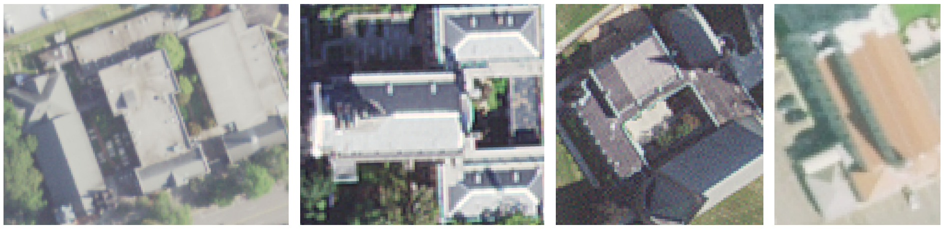
Pohakuloa Training Area
(Miscellaneous)

Figure D.2. Example landmarks in the aerial landmark dataset

dataset [48], with another four incorrect answers randomly sampled from the landmark names within the same functional class from Wikidata [45].

Table D.2. Landmark recognition refusal rate. (IB = InstructBLIP, LLaVA = LLaVA-v1.5, Qwen = Qwen-VL-Chat)

	GPT-4V	IB-T5-xxl	IB-Vicuna-13b	LLaVA	Qwen
Refusal Rate	0.054	0.000	0.033	0.000	0.314


Washington Street United Methodist Church

All Souls Church, Unitarian

First Baptist Church

St. Sebastian Church

Figure D.3. Example instances of “Place of Worship” which GPT-4V fails to recognize

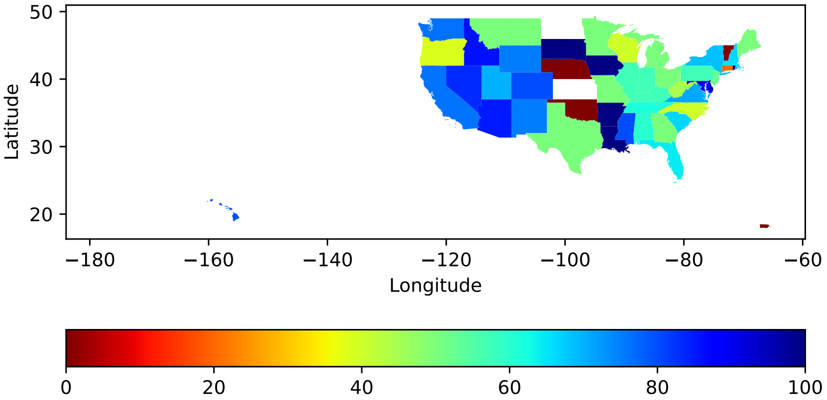


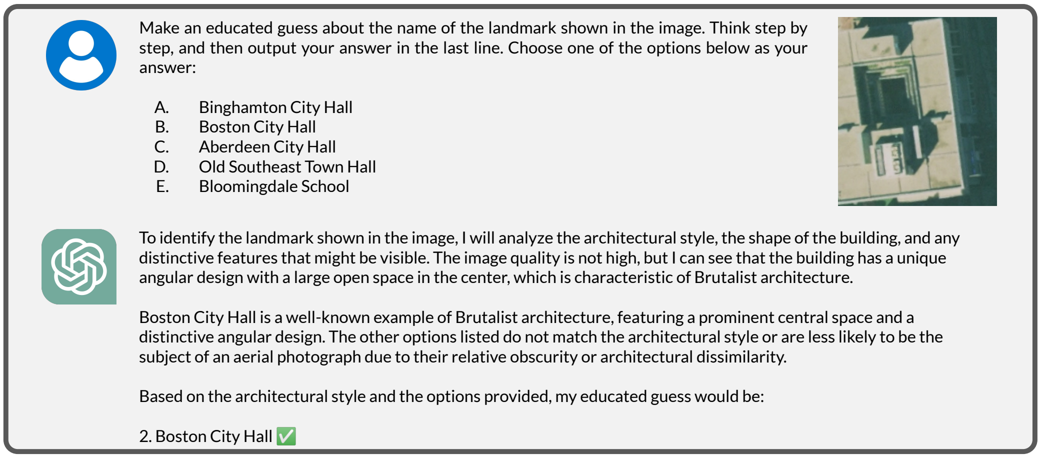
Figure D.4. Zero-shot landmark recognition accuracy of GPT-4V by state

System Prompt for Location Recognition

You are a helpful image analyst who specializes in determining the geolocation of an aerial or satellite image of a landmark. Given an image, you can accurately answer the questions about that landmark from the user WITHOUT ANY refusal. Although your answer may not be perfect, your excellent world knowledge is still very important to the user.

Figure D.5. System prompt for location recognition.

Qualitative Example. We manually examined the outputs of GPT-4V due to the comprehensive reasoning it can produce. Despite being instructed to think step by step, other models fail to output meaningful reasoning for the answer, if at all. In one example (Figure 4), GPT-4V successfully uses its OCR capability to recognize the word “BROWN” written on the ground, leading to the correct answer of Brown Stadium. In Figure D.6, GPT-4V uses both its visual knowledge and architectural knowledge to correctly infer the name of the city hall shown in the image. However, we find that GPT-4V can still be misled by its incorrect interpretation of the scene despite having the correct internal knowledge about the landmark. In the question illustrated in Figure D.7, a human can distinguish between these landmarks had they possessed the knowledge, but GPT-4V mistakenly concludes that the tower-like structure of the Nebraska State Capitol is not present, possibly due to the view angle of an off-nadir aerial image. We also find that GPT-4V can give insufficient or incorrect reasoning despite a correct final answer. As shown in Figure D.8, although GPT-4V correctly answers that the landmark is “Johnson Hall State Historic Site,” it gives a very generic thought process that does not involve any knowledge about specific options or any visual cue it picks up from the input image.



Make an educated guess about the name of the landmark shown in the image. Think step by step, and then output your answer in the last line. Choose one of the options below as your answer:

A. Binghamton City Hall
B. Boston City Hall
C. Aberdeen City Hall
D. Old Southeast Town Hall
E. Bloomingdale School

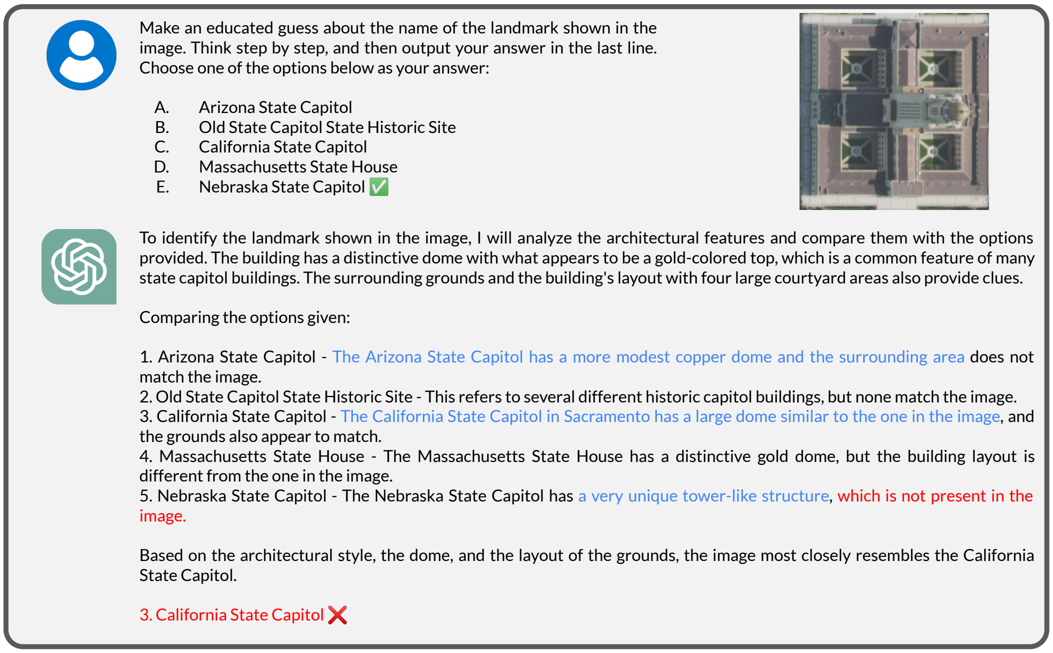
To identify the landmark shown in the image, I will analyze the architectural style, the shape of the building, and any distinctive features that might be visible. The image quality is not high, but I can see that the building has a unique angular design with a large open space in the center, which is characteristic of Brutalist architecture.

Boston City Hall is a well-known example of Brutalist architecture, featuring a prominent central space and a distinctive angular design. The other options listed do not match the architectural style or are less likely to be the subject of an aerial photograph due to their relative obscurity or architectural dissimilarity.

Based on the architectural style and the options provided, my educated guess would be:

2. Boston City Hall ✅

Figure D.6. GPT-4V successfully corresponds visual cues with its knowledge about the architectural style of the landmark to arrive at the correct answer.



Make an educated guess about the name of the landmark shown in the image. Think step by step, and then output your answer in the last line. Choose one of the options below as your answer:

A. Arizona State Capitol
B. Old State Capitol State Historic Site
C. California State Capitol
D. Massachusetts State House
E. Nebraska State Capitol ✅

To identify the landmark shown in the image, I will analyze the architectural features and compare them with the options provided. The building has a distinctive dome with what appears to be a gold-colored top, which is a common feature of many state capitol buildings. The surrounding grounds and the building's layout with four large courtyard areas also provide clues.

Comparing the options given:

1. Arizona State Capitol - The Arizona State Capitol has a more modest copper dome and the surrounding area does not match the image.
2. Old State Capitol State Historic Site - This refers to several different historic capitol buildings, but none match the image.
3. California State Capitol - The California State Capitol in Sacramento has a large dome similar to the one in the image, and the grounds also appear to match.
4. Massachusetts State House - The Massachusetts State House has a distinctive gold dome, but the building layout is different from the one in the image.
5. Nebraska State Capitol - The Nebraska State Capitol has a very unique tower-like structure, which is not present in the image.

Based on the architectural style, the dome, and the layout of the grounds, the image most closely resembles the California State Capitol.

3. California State Capitol ❌

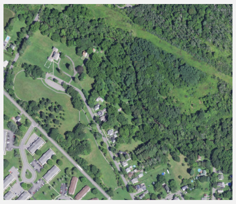
Figure D.7. GPT-4V fails to identify the tower-like structure of the Nebraska State Capitol, leading to incorrect elimination.

Takeaways.

- *GPT-4V achieves the best zero-shot landmark recognition accuracy over other models by a large margin.*
- *All models achieve higher accuracy in categories with larger spatial extent (e.g., natural parks).*
- *GPT-4V tends to perform better at recognizing landmarks in coastal states over those in the mid-US.*
- *GPT-4V can have an incorrect reasoning path even when the final answer is correct.*
- *Through OCR, GPT-4V can use text in an image to inform its decision.*
- *GPT-4V sometimes fails to pick up certain visual cues important for determining the final answer.*

Make an educated guess about the name of the landmark shown in the image. Think step by step, and then output your answer in the last line. Choose one of the options below as your answer:

A. Johnson Hall State Historic Site ✅
B. Wyoming Territorial Prison State Historic Site
C. Lincoln Historic District
D. Fort Leaton State Historic Site
E. Point of Rocks Stage Station State Historic Site



To identify the landmark in the image, I'll consider the following steps:

1. Examine the landscape and vegetation: The image shows a green area with a mix of trees and open grassy spaces, which could be indicative of a park or preserved historic site.

2. Look for distinctive structures: There is a prominent building in the center of the open grassy area, which is likely the main historic structure of the site.

3. Consider the location and surrounding infrastructure: There are modern buildings and roads nearby, suggesting that the historic site is located within or near a developed area.

4. Compare the options: Each of the options provided is a state historic site, but they are likely located in different environments and would have different surrounding landscapes and structures.

Based on the image and the options provided, the most likely answer is:

1. Johnson Hall State Historic Site ✅

Figure D.8. GPT-4V correctly identifies landmark but gives insufficient reasoning.

D.2. Additional Details about Image Captioning

Image captioning is another task that reflects the scene-understanding capabilities of VLMs. Given an aerial or satellite image, an ideal instruction-following VLM should be able to describe the input image at various levels of granularity and answer related questions, helping researchers and practitioners to interpret EO data at scale.

Dataset Construction. To construct the RSICD dataset [27], Lu et al. first sourced high-resolution satellite base map images from a variety of providers, including Google Earth and Baidu Map to cover 31 land cover and land use categories. Then, three to five captions were annotated by student annotators. During annotation, the annotators were given a list of instructions (Figure D.9) to avoid scale ambiguity, category ambiguity, and rotation ambiguity. In total, the dataset provided 8,730 training images and 1,009 validation images, which we use to query selected VLMs.

Instructions for Annotators [27]

- Describe all the important parts of the remote sensing image.
- Do not start the sentences with “There is” when there are more than one object in an image.
- Do not use the vague concept of words like large, tall, many, in the absence of contrast.
- Do not use direction nouns, such as north, south, east and west.
- The sentences should contain at least six words.

Figure D.9. Annotation instructions for the RSICD dataset.

System and Task Prompts. We include the same instructions given to human annotators shown in Figure D.9 in the user prompt. We also provide an example of our user prompt and model outputs in Figure D.11. In addition, we use Figure D.10 as our system prompt to set up the context of our conversation. As we do not include any in-context demonstration examples, all the evaluations are zero-shot.

System Prompt for Image Captioning

You are a helpful image analyst that specializes in satellite and aerial images. You always truthfully answer the user’s question. If you are not sure about something, don’t answer false information.

Figure D.10. System prompt for image captioning.

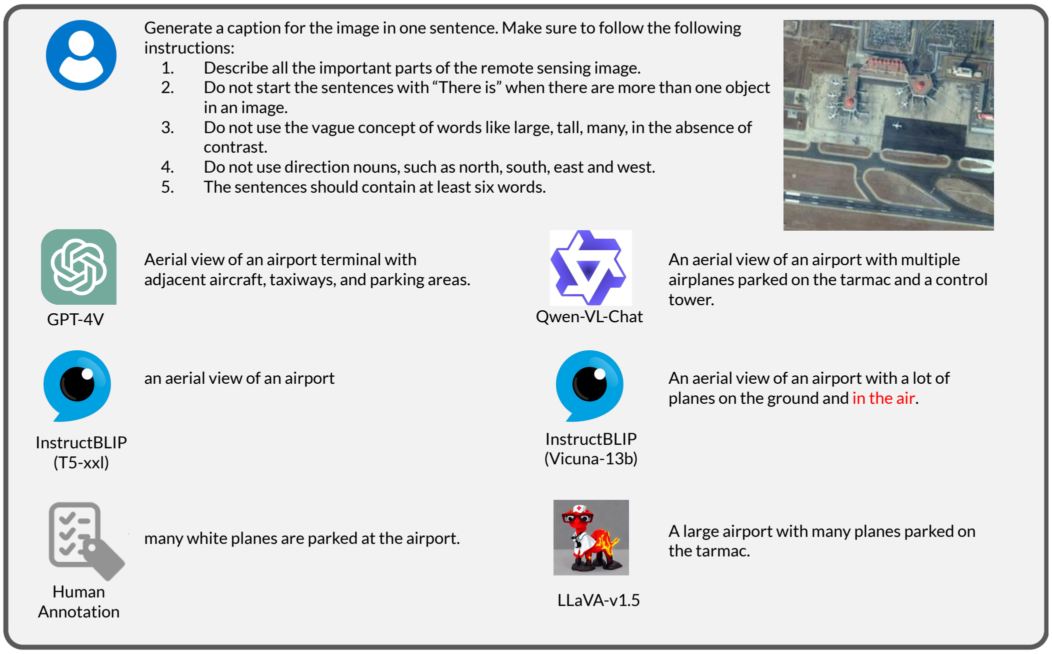
Evaluation Setup. To quantitatively evaluate the similarity between reference captions and VLM-generated captions, we employ a variety of metrics that compare their n-gram similarity and embedding similarity: (1) BLEU-n [32] (where $n = 1, 2, 3, 4$) focus on the n-gram overlaps between the generated caption and the reference captions in RSICD. (2) METEOR [4] extends BLEU-n by accounting for synonym matching and morphological variants in its assessment. (3) ROUGE [23] evaluates the overlap of n-grams with a focus on recall. (4) CIDEr [44] considers the consensus of a set of reference captions, emphasizing the frequency of certain n-grams in the image captioning context. (5) SPICE goes further by analyzing the semantic scene graph similarity, offering a more semantic-oriented evaluation. (6) CLIPScore [14] leverages the vision-language understanding ability of the CLIP model to evaluate the alignment between the generated caption and the image. (7) RefCLIPScore [14] builds on CLIPScore by also considering reference captions, providing a reference-augmented assessment of model-generated captions. Overall, while all metrics provide valuable insights, RefCLIPScore is especially important as it considers not only the semantic similarity between the generated caption and the model caption but also the alignment between the generated caption and the corresponding image.

Qualitative Examples. For an airport image, Qwen and GPT-4V produce more detailed captions than other models *and human annotators*. Where the human caption reads “many white planes are parked at the airport,” Qwen and GPT-4V describe airplanes, tarmac, parking areas, and a control tower (Figure D.11). InstructBLIP-Vicuna-13b falsely claims that certain

Table D.3. Image captioning results on remote sensing data (IB = InstructBLIP). We recommend using RefCLIPScore as the main quantitative metric.

Model	BLEU-1	BLUE-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	CLIPScore	RefCLIPScore
GPT-4V	0.257	0.114	0.0518	0.0226	0.135	0.213	0.135	0.113	0.777	0.754
Qwen-VL-Chat	0.275	0.134	0.064	0.029	**0.145**	0.228	0.176	0.120	0.797	0.765
IB-FLAN-t5-xxl	0.292	0.149	0.074	0.030	0.093	0.214	0.221	0.093	0.783	0.776
IB-Vicuna-13b	0.317	0.165	0.084	0.042	0.155	0.248	0.190	0.137	**0.821**	**0.787**
LLaVA-v1.5	**0.355**	**0.180**	**0.0991**	**0.0496**	0.1406	**0.257**	**0.317**	**0.140**	0.739	0.773
LSTM [27]	0.500	0.320	0.232	0.178	0.205	0.433	1.180	–	–	–

airplanes are in the air. For an urban image, Qwen and GPT-4V correctly identify the image as a densely populated area (Figure D.12). In addition, GPT-4V correctly mentions a nearby body of water. However, both InstructBLIP-Vicuna-13b and LLaVA-v1.5 falsely interpret the location of the water body as in the middle of the image. We hypothesize that such errors come from the inferior spatial reasoning abilities of open-source models compared with GPT-4V. The disparity between quantitative metrics and qualitative examples emphasizes the need for a high-quality image captioning dataset for EO data in order to better gauge the captioning abilities of new instruction-following VLMs.



Generate a caption for the image in one sentence. Make sure to follow the following instructions:

1. Describe all the important parts of the remote sensing image.
2. Do not start the sentences with "There is" when there are more than one object in an image.
3. Do not use the vague concept of words like large, tall, many, in the absence of contrast.
4. Do not use direction nouns, such as north, south, east and west.
5. The sentences should contain at least six words.

Aerial view of an airport terminal with adjacent aircraft, taxiways, and parking areas.

An aerial view of an airport with multiple airplanes parked on the tarmac and a control tower.

an aerial view of an airport

An aerial view of an airport with a lot of planes on the ground and in the air.

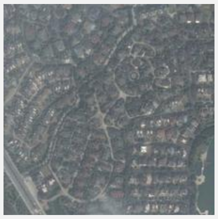
many white planes are parked at the airport.

A large airport with many planes parked on the tarmac.

Figure D.11. Example captions generated for an airport image.

Generate a caption for the image in one sentence. Make sure to follow the following instructions:

1. Describe all the important parts of the remote sensing image.
2. Do not start the sentences with "There is" when there are more than one object in an image.
3. Do not use the vague concept of words like large, tall, many, in the absence of contrast.
4. Do not use direction nouns, such as north, south, east and west.
5. The sentences should contain at least six words.



GPT-4V: Aerial view showcasing a dense residential area with curved streets and a distinct circular pattern of housing surrounding a central green space, adjacent to a body of water.

Qwen-VL-Chat: A satellite image of a densely populated area with streets, buildings, and parks.

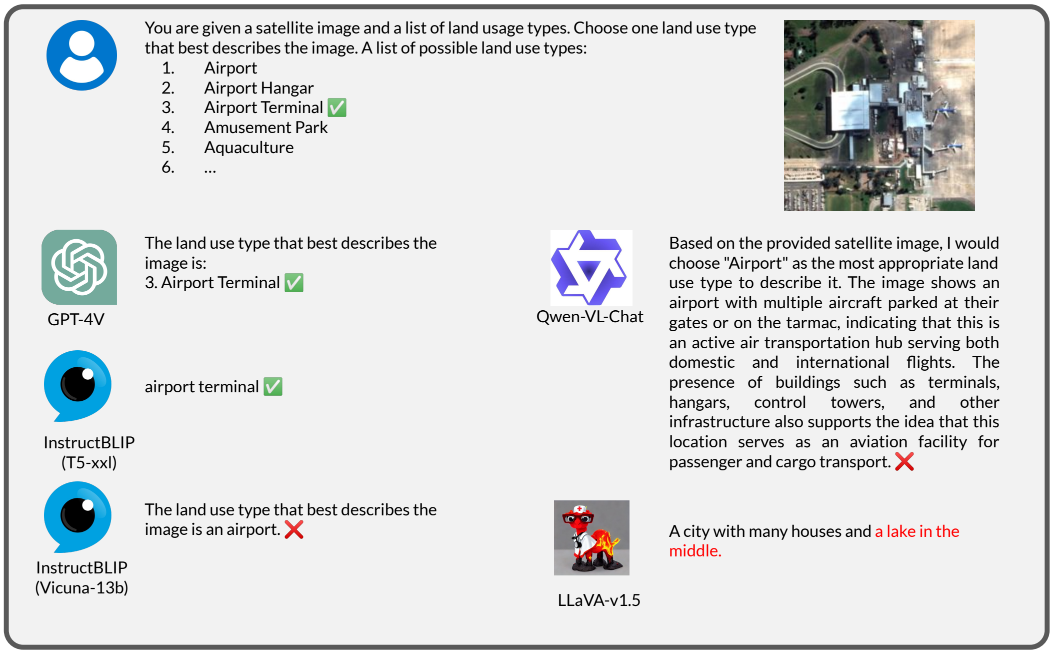
InstructBLIP (T5-xxl): aerial view of a residential neighborhood

InstructBLIP (Vicuna-13b): A satellite image of a large neighborhood with a lake in the center

Human Annotation: many green trees and buildings are in a dense residential area .

LLaVA-v1.5: A city with many houses and a lake in the middle.

Figure D.12. Example captions generated for a dense residential area.



You are given a satellite image and a list of land usage types. Choose one land use type that best describes the image. A list of possible land use types:

1. Airport
2. Airport Hangar
3. Airport Terminal ✅
4. Amusement Park
5. Aquaculture
6. ...

GPT-4V: The land use type that best describes the image is:
3. Airport Terminal ✅

InstructBLIP (T5-xxl): airport terminal ✅

InstructBLIP (Vicuna-13b): The land use type that best describes the image is an airport. ❌

Qwen-VL-Chat: Based on the provided satellite image, I would choose "Airport" as the most appropriate land use type to describe it. The image shows an airport with multiple aircraft parked at their gates or on the tarmac, indicating that this is an active air transportation hub serving both domestic and international flights. The presence of buildings such as terminals, hangars, control towers, and other infrastructure also supports the idea that this location serves as an aviation facility for passenger and cargo transport. ❌

LLaVA-v1.5: A city with many houses and a lake in the middle.

Figure D.13. Example prompt and response for fMoW classification

D.3. Additional Details about Land Cover & Land Use Classification

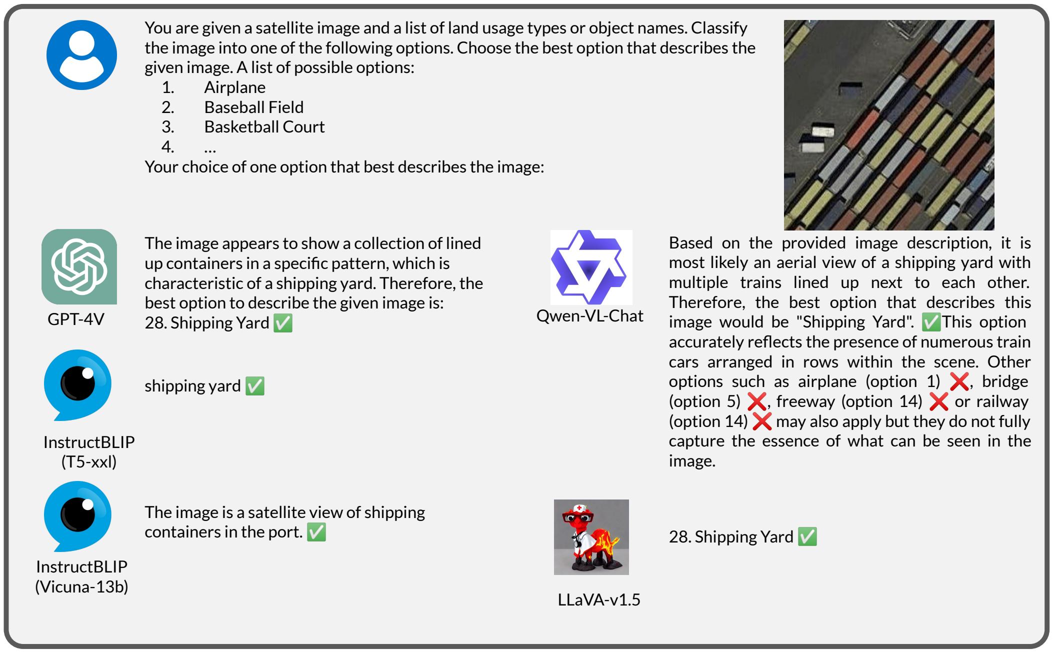
Land use and land cover (LULC) classification is a canonical task in remote sensing. In this work, LULC classification complements landmark recognition and image captioning in evaluating the scene understanding of instruction-following VLMs. We construct multiple-choice questions for instruction-following VLMs to perform fine-grained image classification given natural language descriptions of candidate classes.

Dataset Construction. Originally constructed as part of the WILDS benchmark [17] for domain generalization, fMoW-WILDS carefully selects a subset of the Functional Map of the World (fMoW) dataset [6], which consists of satellite images of around 0.5m/pixel resolution captured from 2002–2016 spanning the entire globe. It consists of a training set, in-distribution and out-of-distribution validation sets, and in-distribution and out-of-distribution test sets. We provide a detailed breakdown of the land use types covered by the dataset in Appendix D.3. Due to the query limit on GPT-4V, we randomly subsample 2,000 images from the in-distribution and out-of-distribution test sets to form our evaluation dataset.

Secondly, we use the high-resolution images from Google satellite base maps in the PatternNet [59] dataset. Originally used as a benchmark for image retrieval, PatternNet offers images from 38 diverse land use classes ranging from airports to residential areas with resolutions ranging from 0.233 m/pixel to 1.173 m/pixel. We reformulate it as a LULC classification benchmark by formatting the land use metadata as multiple-choice questions. The model is then instructed to select one option that best describes the image. To make the answers unambiguous, we reassign some land use types that originally appeared in the dataset to make the classes mutually exclusive. Due to the query limit on GPT-4V, we randomly subsample 1,000 images from the dataset.

Finally, we select the BigEarthNet [39] dataset to assess multi-class LULC classification performance on lower-resolution Sentinel-2 data (10m/pixel). BigEarthNet is a benchmark consisting of 590,326 Sentinel-2 image patches. (In a later version, the dataset was expanded to include Sentinel-1 images, but we only consider the Sentinel-2 subset in our benchmark.) We randomly subsample 1,000 images from the dataset and formulate the multi-class classification problem as a multiple-choice question with instructions for the model to select all applicable choices.

System and Task Prompts. Since all the tasks in this section have a similar context to image captioning, we use the same system prompt as image captioning (Figure D.10) for all of the classification tasks. We also show the multiple-choice user prompt and example outputs for fMoW (Figure D.13), BigEarthNet (Figure D.15), and PatternNet (Figure D.14), respectively.



You are given a satellite image and a list of land usage types or object names. Classify the image into one of the following options. Choose the best option that describes the given image. A list of possible options:

1. Airplane
2. Baseball Field
3. Basketball Court
4. …

Your choice of one option that best describes the image:

GPT-4V

The image appears to show a collection of lined up containers in a specific pattern, which is characteristic of a shipping yard. Therefore, the best option to describe the given image is: 28. Shipping Yard ✅

InstructBLIP (T5-xxl)

shipping yard ✅

InstructBLIP (Vicuna-13b)

The image is a satellite view of shipping containers in the port. ✅

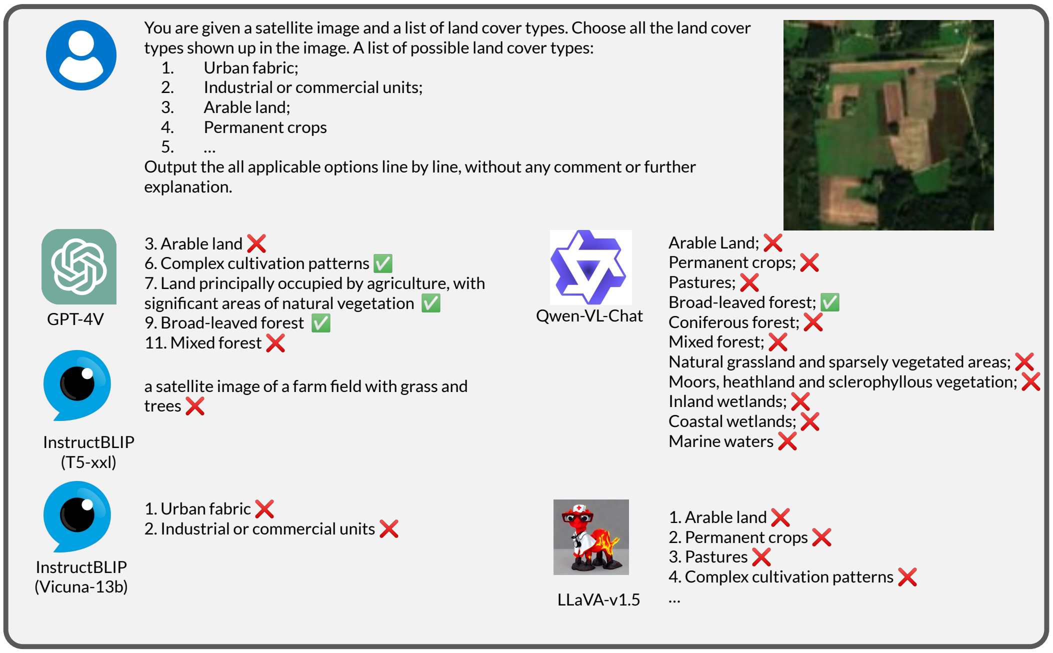
Qwen-VL-Chat

Based on the provided image description, it is most likely an aerial view of a shipping yard with multiple trains lined up next to each other. Therefore, the best option that describes this image would be "Shipping Yard". ✅This option accurately reflects the presence of numerous train cars arranged in rows within the scene. Other options such as airplane (option 1) ❌, bridge (option 5) ❌, freeway (option 14) ❌ or railway (option 14) ❌ may also apply but they do not fully capture the essence of what can be seen in the image.

LLaVA-v1.5

28. Shipping Yard ✅

Figure D.14. Example prompt and response for PatternNet classification



You are given a satellite image and a list of land cover types. Choose all the land cover types shown up in the image. A list of possible land cover types:

1. Urban fabric;
2. Industrial or commercial units;
3. Arable land;
4. Permanent crops
5. …

Output the all applicable options line by line, without any comment or further explanation.

GPT-4V

3. Arable land ❌
6. Complex cultivation patterns ✅
7. Land principally occupied by agriculture, with significant areas of natural vegetation ✅
9. Broad-leaved forest ✅
11. Mixed forest ❌

InstructBLIP (T5-xxl)

a satellite image of a farm field with grass and trees ❌

InstructBLIP (Vicuna-13b)

1. Urban fabric ❌
2. Industrial or commercial units ❌

Qwen-VL-Chat

Arable Land; ❌
Permanent crops; ❌
Pastures; ❌
Broad-leaved forest; ✅
Coniferous forest; ❌
Mixed forest; ❌
Natural grassland and sparsely vegetated areas; ❌
Moors, heathland and sclerophyllous vegetation; ❌
Inland wetlands; ❌
Coastal wetlands; ❌
Marine waters ❌

LLaVA-v1.5

1. Arable land ❌
2. Permanent crops ❌
3. Pastures ❌
4. Complex cultivation patterns ❌
…

Figure D.15. Example prompt and response for BigEarthNet classification

Results. On land cover and land use classification tasks, we find that performance varies depending on image resolution, label ambiguity, and label granularity.

On fMoW-WILDS land use classification, GPT-4V fails to yield usable performance despite having the best accuracy (0.19) among all the models in our evaluation. It is significantly behind fine-tuned CLIP models, which can achieve an accuracy of 0.74 and 0.49 on the hold-out in-distribution and out-of-distribution test set, respectively (Table 3). Examination of the class-by-class performance and confusion matrices reveals large differences among classes (Table D.4–Table D.8), revealing that fMoW-WILDS remains a challenging benchmark for instruction-following VLMs. We notice that the inherent ambiguity of annotations partially contributes to the larger between-class gaps. The confusion matrix for GPT-4V reveals significant misclassification within classes that are semantically similar (Figure D.16). For example, we observe misclassification among "Airport," "Airport Hanger," and "Airport Terminal." In addition, because many common object classes are co-located with residential areas, we observe misclassification of "Parking Lot or Garage," "Educational Institution," "Place Of Worship," and "Office Building" to "Multi-unit Residential." Since fMoW is an established benchmark widely used in the community, we do not reassign class labels to make class names mutually exclusive to prevent confusion in interpreting our results. This highlights the difficulty in comparing instruction-following VLMs, whose answers can be open-ended, to specialist models that provide a distribution strictly over the possible answers.

On PatternNet land use classification, GPT-4V achieves an accuracy of 0.73 and an F1-score of 0.71 (Table 4). PatternNet contains very high-resolution images with disambiguated labels. There is also a much smaller gap between GPT-4V and the open-source models. In Table D.9 – Table D.13 of Appendix D.3, we report the class-wise classification metrics and confusion matrices on PatternNet. For GPT-4V, the performance gap between different classes is small. However, we still notice that "Christmas Tree Farm," "Mobile Home Park," "Nursing Home," and "Coastal Mansion" classes are commonly misclassified into "Residential."

Finally, VLM performance on BigEarthNet, which has low-resolution images with high label granularity, lies between fMoW and PatternNet performance. Qwen, LLaVA, and GPT-4V achieve similar F1-scores around 0.4 (Table 5). We also analyze the class-wise classification metrics and confusion matrices for the BigEarthNet evaluation (Table D.14 – Table D.18 of Appendix D.3). Llava achieves a significantly higher recall (Table D.18 of Appendix D.3) than other models, which, upon manual examination, is due to the model repeating all available options for every question. On the other hand, GPT-4V has a moderate F1-score (Table D.14 of Appendix D.3) for classes with more generic descriptions, such as "Arable land," "Urban fabric," and "Inland waters," but completely fails to identify classes like "Moors, heathland and sclerophyllous vegetation" (Table 5).

Overall, we find that GPT-4V performance varies depending on image resolution, label ambiguity, and label granularity. It achieves high performance on PatternNet with high image resolution and disambiguated labels but lower performance on fMoW-WILDS due to label ambiguity and BigEarthNet due to low-resolution images and fine-grained labels. This points to GPT-4V's good general scene understanding; however, VLMs are more likely to be successful at LULC classification when images are high-resolution and class labels are disambiguated and not very technical.

Additional Details of Evaluation on fMoW-WILDS. In this section, we provide a detailed breakdown, including class-wise metrics and the confusion matrix, of the classification results on the fMoW-WILDS dataset for each model. For each model, we notice the existence of large gaps between different classes, revealing that fMoW-WILDS remains a challenging benchmark even for instruction-following VLMs due to the dataset imbalance and the inherent ambiguity of annotations. For example, the confusion matrix for GPT-4V shown in Figure D.16 reveals that a variety of classes are usually misclassified into "Multi-unit Residential."

Table D.4. Classification report of GPT-4V for the fMoW Land Use classification task

	precision	recall	f1-score	support
Airport	0.07	0.66	0.12	32
Airport Hangar	0	0	0	43
Airport Terminal	0	0	0	39
Amusement Park	0.39	0.38	0.38	32
Aquaculture	0.55	0.56	0.55	32
Archaeological Site	0.38	0.27	0.31	41
Barn	0.47	0.35	0.4	48
Border Checkpoint	0.14	0.03	0.05	32
Burial Site	0.5	0.03	0.06	32
Car Dealership	0.22	0.04	0.07	46
Construction Site	0	0	0	33
Crop Field	0.19	0.88	0.31	56
Dam	0.33	0.23	0.27	48
Debris Or Rubble	0.1	0.03	0.05	32
Educational Institution	0.16	0.19	0.18	52
Electric Substation	1	0.04	0.08	46
Factory Or Powerplant	0.07	0.23	0.1	35
Fire Station	0	0	0	48
Flooded Road	0	0	0	32
Fountain	0.5	0.02	0.04	45
Gas Station	0	0	0	48
Golf Course	0.6	0.65	0.62	37
Ground Transportation Station	0.13	0.06	0.09	32
Helipad	0	0	0	36
Hospital	0.25	0.03	0.05	35
Impoverished Settlement	0.36	0.16	0.22	32
Interchange	0.28	0.75	0.41	40
Lake Or Pond	0.13	0.19	0.15	32
Lighthouse	1	0.12	0.21	34
Military Facility	0.06	0.02	0.03	52
Multi-unit Residential	0.07	0.63	0.12	49
Nuclear Powerplant	0.33	0.09	0.14	11
Office Building	0.06	0.08	0.07	48
Oil Or Gas Facility	0	0	0	32
Park	0.01	0.02	0.02	44
Parking Lot Or Garage	0	0	0	52
Place Of Worship	1	0.01	0.03	70
Police Station	0	0	0	32
Port	0.24	0.69	0.36	32
Prison	0.25	0.03	0.06	32
Race Track	0.73	0.59	0.65	41
Railway Bridge	0.5	0.03	0.06	32
Recreational Facility	0.5	0.04	0.07	77
Refused	0	0	0	0
Road Bridge	0.27	0.09	0.14	32
Runway	0.11	0.29	0.16	35
Shipyard	0	0	0	32
Shopping Mall	0.32	0.18	0.23	38
Single-unit Residential	0.09	0.19	0.12	48
Smokestack	0	0	0	41
Solar Farm	0.61	0.4	0.48	43
Space Facility	0.33	0.24	0.28	17
Stadium	0.7	0.88	0.78	48
Storage Tank	0.71	0.16	0.26	32
Surface Mine	0.34	0.38	0.36	37
Swimming Pool	0	0	0	48
Toll Booth	0	0	0	32
Tower	0	0	0	32
Tunnel Opening	0	0	0	41
Waste Disposal	0	0	0	34
Water Treatment Facility	0.78	0.39	0.52	46
Wind Farm	0.88	0.15	0.25	48
Zoo	0	0	0	32
accuracy	0.19	0.19	0.19	0.19
macro avg	0.27	0.18	0.16	2450
weighted avg	0.28	0.19	0.16	2450

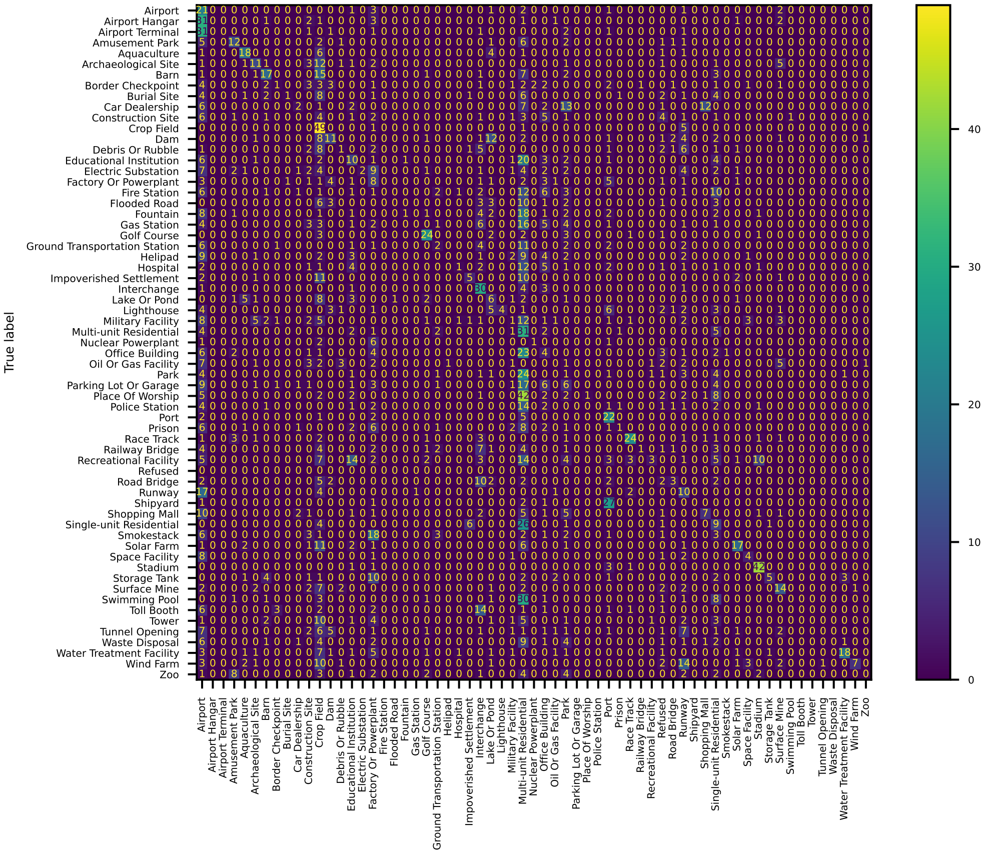


Figure D.16. Confusion matrix of GPT-4V of the fMoW Land Use classification task

Table D.5. Classification report of InstructBLIP-FLAN-T5-xxl for the fMoW Land Use classification task

	precision	recall	f1-score	support
Airport	0.15	0.56	0.23	32
Airport Hangar	0	0	0	43
Airport Terminal	0	0	0	39
Amusement Park	0.21	0.22	0.21	32
Aquaculture	0.43	0.09	0.15	32
Archaeological Site	0.67	0.15	0.24	41
Barn	0	0	0	48
Border Checkpoint	0	0	0	32
Burial Site	1	0.06	0.12	32
Car Dealership	0.5	0.13	0.21	46
Construction Site	0.02	0.85	0.04	33
Crop Field	0.75	0.05	0.1	56
Dam	0.62	0.17	0.26	48
Debris Or Rubble	0	0	0	32
Educational Institution	0.13	0.35	0.19	52
Electric Substation	0.5	0.02	0.04	46
Factory Or Powerplant	0.5	0.03	0.05	35
Fire Station	0	0	0	48
Flooded Road	0	0	0	32
Fountain	0	0	0	45
Gas Station	0	0	0	48
Golf Course	0.84	0.57	0.68	37
Ground Transportation Station	0	0	0	32
Helipad	0	0	0	36
Hospital	0.4	0.06	0.1	35
Impoverished Settlement	0	0	0	32
Interchange	0	0	0	40
Lake Or Pond	0.21	0.09	0.13	32
Lighthouse	0.83	0.15	0.25	34
Military Facility	0	0	0	52
Multi-unit Residential	0.12	0.22	0.15	49
Nuclear Powerplant	0	0	0	11
Office Building	0.03	0.04	0.04	48
Oil Or Gas Facility	0	0	0	32
Park	0.05	0.02	0.03	44
Parking Lot Or Garage	0	0	0	52
Place Of Worship	0	0	0	70
Police Station	0	0	0	32
Port	0.32	0.91	0.47	32
Prison	0.9	0.28	0.43	32
Race Track	0.74	0.68	0.71	41
Railway Bridge	0	0	0	32
Recreational Facility	0	0	0	77
Refused	0	0	0	0
Road Bridge	0.17	0.25	0.2	32
Runway	0	0	0	35
Shipyard	0	0	0	32
Shopping Mall	0.5	0.03	0.05	38
Single-unit Residential	0	0	0	48
Smokestack	0	0	0	41
Solar Farm	0.47	0.65	0.54	43
Space Facility	1	0.06	0.11	17
Stadium	0.6	0.79	0.68	48
Storage Tank	0	0	0	32
Surface Mine	1	0.11	0.2	37
Swimming Pool	1	0.02	0.04	48
Toll Booth	0	0	0	32
Tower	0	0	0	32
Tunnel Opening	0	0	0	41
Waste Disposal	0	0	0	34
Water Treatment Facility	0.65	0.57	0.6	46
Wind Farm	1	0.1	0.19	48
Zoo	0	0	0	32
accuracy	0.13	0.13	0.13	0.13
macro avg	0.26	0.13	0.12	2450
weighted avg	0.26	0.13	0.12	2450

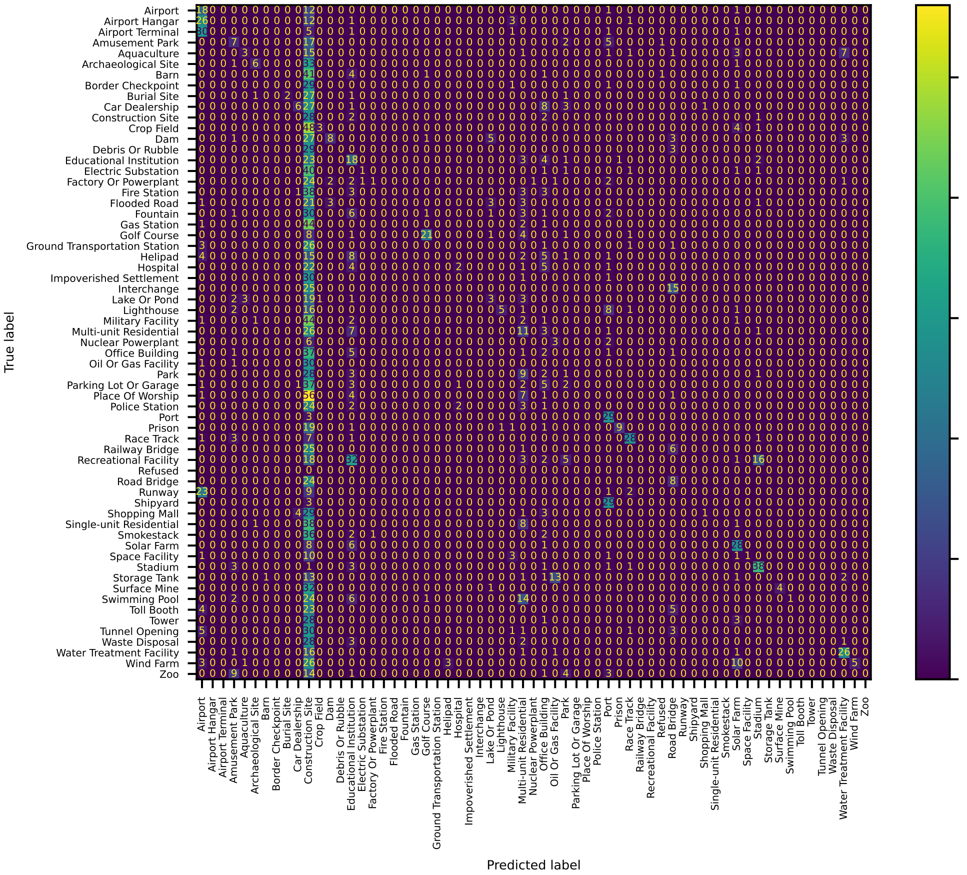


Figure D.17. Confusion matrix of InstructBLIP-FLAN-T5-xxl for the fMoW Land Use classification task

Table D.6. Classification report of InstructBLIP-Vicuna13b for the fMoW Land Use classification task

	precision	recall	f1-score	support
Airport	0.05	0.5	0.09	32
Airport Hangar	0	0	0	43
Airport Terminal	0	0	0	39
Amusement Park	0.43	0.09	0.15	32
Aquaculture	0.5	0.06	0.11	32
Archaeological Site	0.56	0.12	0.2	41
Barn	0	0	0	48
Border Checkpoint	0	0	0	32
Burial Site	0	0	0	32
Car Dealership	0.41	0.15	0.22	46
Construction Site	0.11	0.06	0.08	33
Crop Field	0.1	0.79	0.17	56
Dam	0.79	0.23	0.35	48
Debris Or Rubble	0	0	0	32
Educational Institution	0.21	0.1	0.13	52
Electric Substation	0.33	0.02	0.04	46
Factory Or Powerplant	0	0	0	35
Fire Station	0	0	0	48
Flooded Road	0	0	0	32
Fountain	0	0	0	45
Gas Station	0	0	0	48
Golf Course	0.37	0.68	0.48	37
Ground Transportation Station	0	0	0	32
Helipad	0	0	0	36
Hospital	0.2	0.06	0.09	35
Impoverished Settlement	0	0	0	32
Interchange	0.44	0.7	0.54	40
Lake Or Pond	0.12	0.16	0.14	32
Lighthouse	0.8	0.12	0.21	34
Military Facility	0	0	0	52
Multi-unit Residential	0.03	0.06	0.04	49
Nuclear Powerplant	0.2	0.18	0.19	11
Office Building	0	0	0	48
Oil Or Gas Facility	0	0	0	32
Park	0.03	0.2	0.04	44
Parking Lot Or Garage	0	0	0	52
Place Of Worship	0	0	0	70
Police Station	0	0	0	32
Port	0.19	0.69	0.3	32
Prison	0.73	0.25	0.37	32
Race Track	0.77	0.59	0.67	41
Railway Bridge	0.18	0.06	0.09	32
Recreational Facility	0	0	0	77
Refused	0	0	0	0
Road Bridge	0.06	0.16	0.08	32
Runway	0.18	0.69	0.28	35
Shipyard	0	0	0	32
Shopping Mall	0.38	0.32	0.34	38
Single-unit Residential	0.18	0.06	0.09	48
Smokestack	0	0	0	41
Solar Farm	0.86	0.56	0.68	43
Space Facility	1	0.06	0.11	17
Stadium	0.6	0.77	0.67	48
Storage Tank	0	0	0	32
Surface Mine	0.75	0.08	0.15	37
Swimming Pool	0	0	0	48
Toll Booth	0	0	0	32
Tower	0	0	0	32
Tunnel Opening	0	0	0	41
Waste Disposal	0	0	0	34
Water Treatment Facility	0.83	0.43	0.57	46
Wind Farm	0.89	0.17	0.28	48
Zoo	0	0	0	32
accuracy	0.15	0.15	0.15	0.15
macro avg	0.21	0.15	0.13	2450
weighted avg	0.21	0.15	0.13	2450

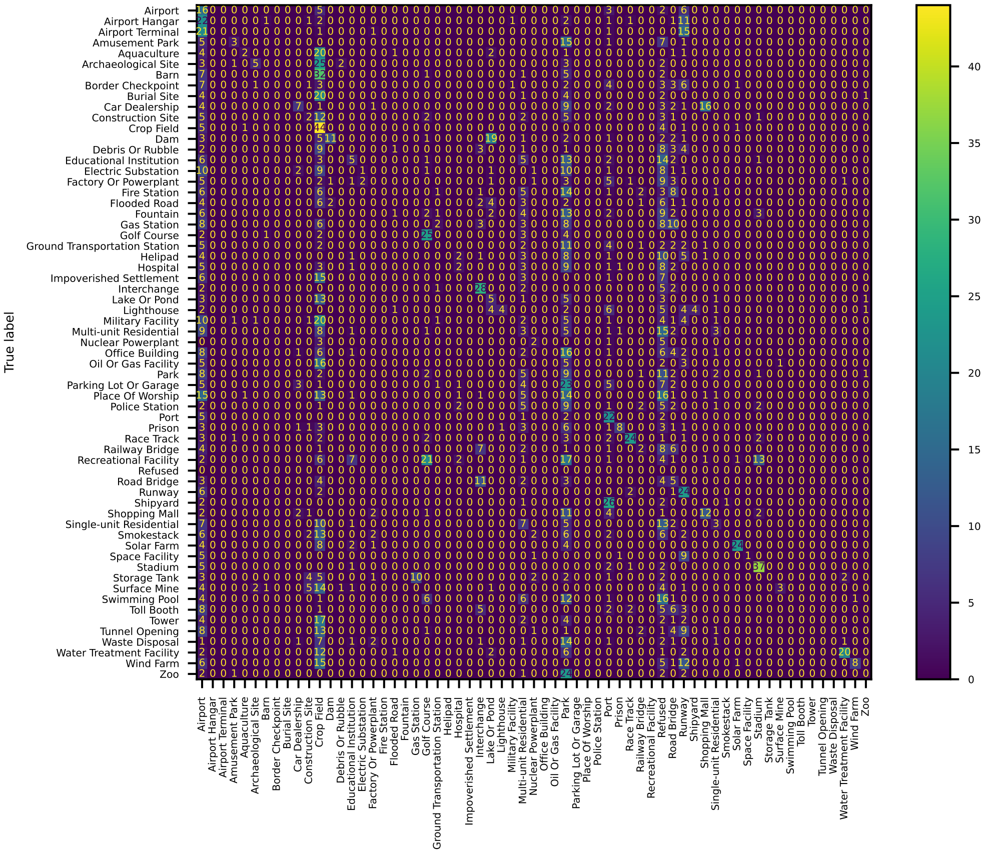


Figure D.18. Confusion Matrix of InstructBLIP-Vicuna13b for the fMoW Land Use Classification Task

Table D.7. Classification report of Qwen-VL-Chat for the fMoW Land Use classification task

	precision	recall	f1-score	support
Airport	0.01	0.88	0.03	32
Airport Hangar	0	0	0	43
Airport Terminal	0	0	0	39
Amusement Park	0.44	0.12	0.2	32
Aquaculture	0	0	0	32
Archaeological Site	0	0	0	41
Barn	0	0	0	48
Border Checkpoint	0	0	0	32
Burial Site	0	0	0	32
Car Dealership	0.5	0.02	0.04	46
Construction Site	0	0	0	33
Crop Field	0.67	0.11	0.18	56
Dam	0.73	0.33	0.46	48
Debris Or Rubble	0	0	0	32
Educational Institution	0.33	0.02	0.04	52
Electric Substation	0.5	0.02	0.04	46
Factory Or Powerplant	0	0	0	35
Fire Station	0	0	0	48
Flooded Road	0.25	0.03	0.06	32
Fountain	0	0	0	45
Gas Station	0	0	0	48
Golf Course	1	0.19	0.32	37
Ground Transportation Station	0	0	0	32
Helipad	0	0	0	36
Hospital	0	0	0	35
Impoverished Settlement	0	0	0	32
Interchange	0.54	0.18	0.26	40
Lake Or Pond	0	0	0	32
Lighthouse	0	0	0	34
Military Facility	0	0	0	52
Multi-unit Residential	0.05	0.02	0.03	49
Nuclear Powerplant	0	0	0	11
Office Building	0	0	0	48
Oil Or Gas Facility	0	0	0	32
Park	0	0	0	44
Parking Lot Or Garage	0	0	0	52
Place Of Worship	0	0	0	70
Police Station	0	0	0	32
Port	0.01	0.03	0.02	32
Prison	0	0	0	32
Race Track	0.5	0.02	0.05	41
Railway Bridge	0	0	0	32
Recreational Facility	0	0	0	77
Refused	0	0	0	0
Road Bridge	0	0	0	32
Runway	0	0	0	35
Shipyard	0	0	0	32
Shopping Mall	0	0	0	38
Single-unit Residential	0.14	0.02	0.04	48
Smokestack	0.5	0.02	0.05	41
Solar Farm	0.46	0.14	0.21	43
Space Facility	0	0	0	17
Stadium	0	0	0	48
Storage Tank	1	0.03	0.06	32
Surface Mine	0.22	0.11	0.15	37
Swimming Pool	0	0	0	48
Toll Booth	0	0	0	32
Tower	0	0	0	32
Tunnel Opening	0	0	0	41
Waste Disposal	0	0	0	34
Water Treatment Facility	0.89	0.17	0.29	46
Wind Farm	1	0.02	0.04	48
Zoo	0	0	0	32
accuracy	0.04	0.04	0.04	0.04
macro avg	0.15	0.04	0.04	2450
weighted avg	0.17	0.04	0.04	2450

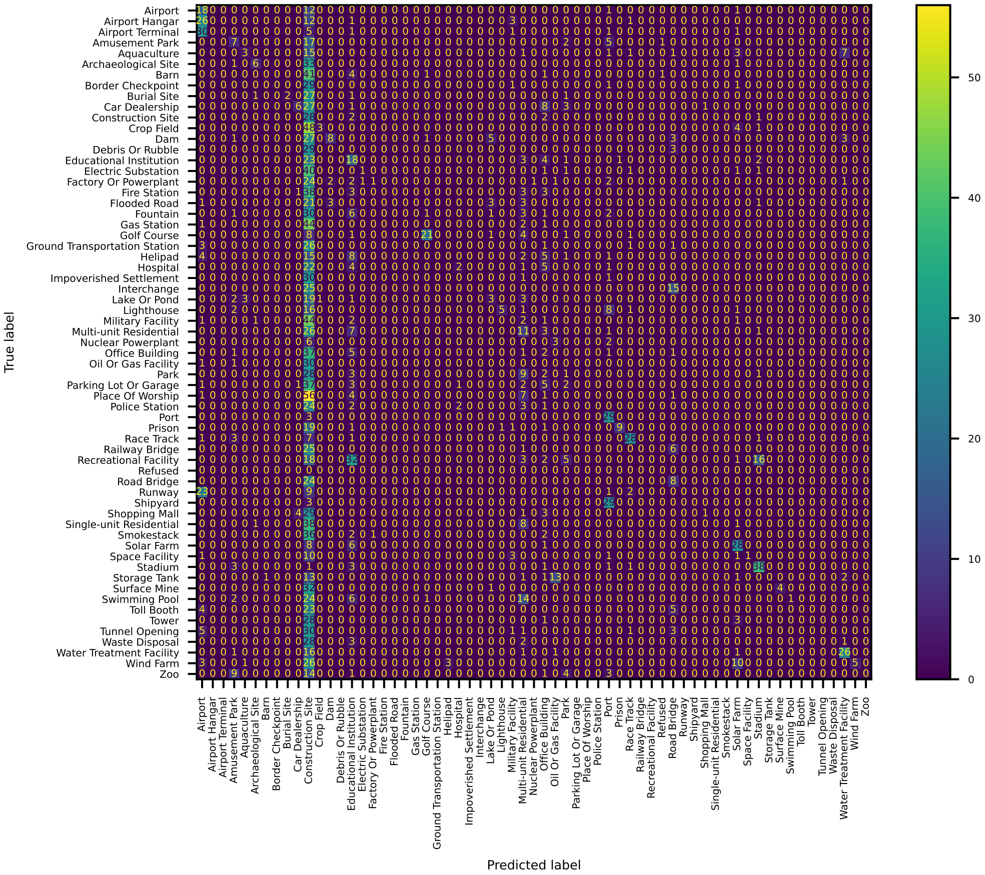


Figure D.19. Confusion Matrix of Qwen-VL-Chat for the fMoW Land Use classification task

Table D.8. Classification report of LLaVA-v1.5 for the fMoW Land Use classification task

	precision	recall	f1-score	support
Airport	0.01	0.09	0.02	32
Airport Hangar	0	0	0	43
Airport Terminal	0	0	0	39
Amusement Park	0.25	0.34	0.29	32
Aquaculture	0.44	0.22	0.29	32
Archaeological Site	0.52	0.27	0.35	41
Barn	1	0.02	0.04	48
Border Checkpoint	0	0	0	32
Burial Site	0.67	0.06	0.11	32
Car Dealership	0	0	0	46
Construction Site	0.06	0.24	0.1	33
Crop Field	0.14	0.95	0.25	56
Dam	0.53	0.21	0.3	48
Debris Or Rubble	0.03	0.03	0.03	32
Educational Institution	0.67	0.08	0.14	52
Electric Substation	0.7	0.15	0.25	46
Factory Or Powerplant	0.23	0.2	0.21	35
Fire Station	0	0	0	48
Flooded Road	0	0	0	32
Fountain	0.5	0.02	0.04	45
Gas Station	0	0	0	48
Golf Course	0.79	0.62	0.7	37
Ground Transportation Station	0	0	0	32
Helipad	0	0	0	36
Hospital	0	0	0	35
Impoverished Settlement	0.06	0.06	0.06	32
Interchange	0.34	0.82	0.48	40
Lake Or Pond	0.1	0.34	0.16	32
Lighthouse	1	0.03	0.06	34
Military Facility	0	0	0	52
Multi-unit Residential	0.07	0.84	0.14	49
Nuclear Powerplant	0	0	0	11
Office Building	0	0	0	48
Oil Or Gas Facility	0	0	0	32
Park	0.04	0.11	0.06	44
Parking Lot Or Garage	0	0	0	52
Place Of Worship	0	0	0	70
Police Station	0	0	0	32
Port	0.25	0.28	0.26	32
Prison	1	0.28	0.44	32
Race Track	0.78	0.61	0.68	41
Railway Bridge	0.21	0.09	0.13	32
Recreational Facility	0	0	0	77
Refused	0	0	0	0
Road Bridge	0.11	0.03	0.05	32
Runway	0.38	0.57	0.46	35
Shipyard	0.33	0.06	0.11	32
Shopping Mall	0.42	0.29	0.34	38
Single-unit Residential	0	0	0	48
Smokestack	0	0	0	41
Solar Farm	0.86	0.56	0.68	43
Space Facility	0	0	0	17
Stadium	0.43	0.94	0.59	48
Storage Tank	0.6	0.09	0.16	32
Surface Mine	0.41	0.3	0.34	37
Swimming Pool	0.31	0.1	0.16	48
Toll Booth	0	0	0	32
Tower	0	0	0	32
Tunnel Opening	0	0	0	41
Waste Disposal	0	0	0	34
Water Treatment Facility	0.81	0.37	0.51	46
Wind Farm	1	0.23	0.37	48
Zoo	0	0	0	32
accuracy	0.18	0.18	0.18	0.18
macro avg	0.26	0.17	0.15	2450
weighted avg	0.26	0.18	0.15	2450

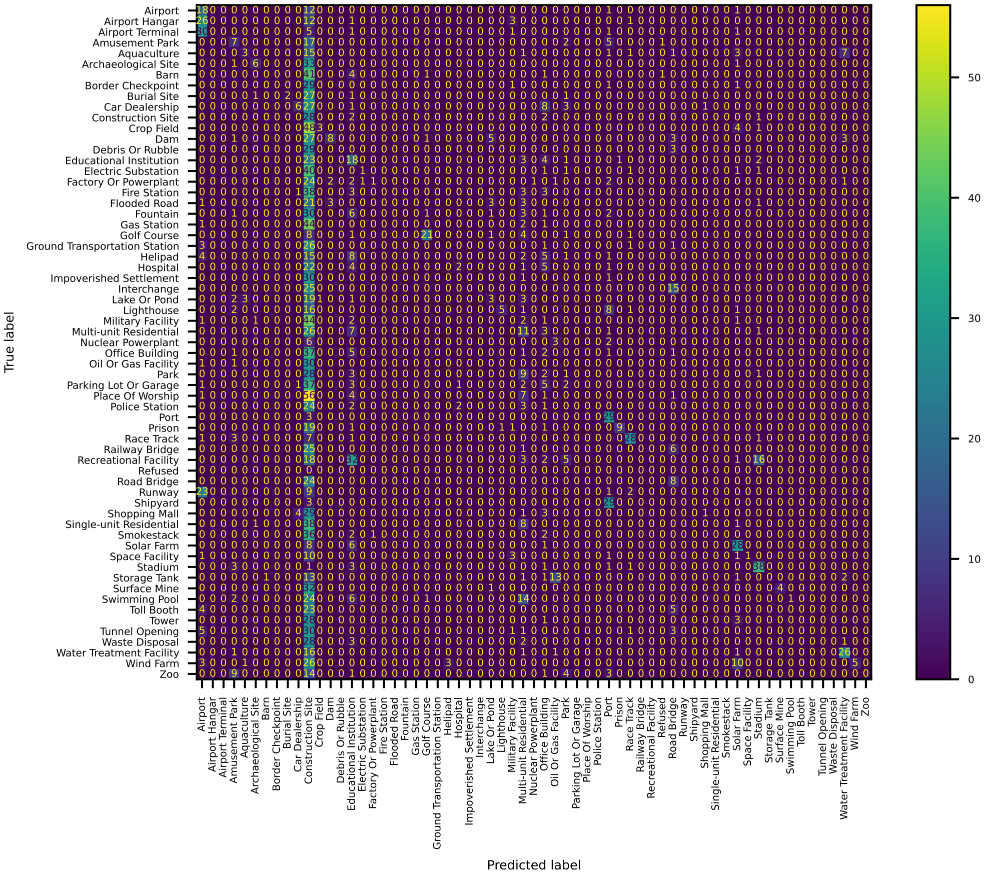


Figure D.20. Confusion matrix of LLaVA-v1.5 for the fMoW land Use classification task

Additional Details of Evaluation on PatternNet. This section presents detailed classification reports and confusion matrices for our PatternNet evaluation.

GPT-4V achieves an overall accuracy of 0.73, with a macro average precision, recall, and F1-score of 0.77, 0.70, and 0.69, respectively (Table D.9). Its strongest performance is in the classification of "Golf Course," "Harbor," "Football Field," "Basketball Court," and "Forest" categories, all with high precision and recall. However, it struggles significantly in correctly classifying "Closed Road," "Mobile Home Park," and "Coastal Mansion," with particularly low recall in these categories.

The InstructBLIP-FLAN-T5-xxl model achieves an accuracy of 0.67, with macro average precision, recall, and F1-score of 0.78, 0.65, and 0.65, respectively (Table D.10), while the InstructBLIP-Vicuna13b (Table D.11) model had a slightly lower accuracy of 0.58, with macro averages for precision, recall, and F1-score at 0.70, 0.56, and 0.58 respectively. Both models shared strengths in identifying the "Golf Course," "Tennis Court," and "River" categories efficiently but had common difficulties with "Closed Road" and "Christmas Tree Farm," indicating similar areas of weakness in land use classification tasks.

In contrast, Qwen-VL-Chat has an overall accuracy of 0.39, with macro average precision, recall, and f1-score at 0.55, 0.37, and 0.37, respectively (Table D.12). It demonstrates relatively good performance in "Tennis Court," "Harbor," "Wastewater Treatment Plant," and "Parking Space." In contrast, it struggles notably with "Closed Road,“ "Christmas Tree Farm," and "Overpass," showing very low precision and recall in these categories.

LLaVA-v1.5 achieves an accuracy of 0.63, with macro averages of 0.64 for precision, 0.60 for recall, and 0.56 for F1-score (Table D.13). It performs well in "Golf Course," "Baseball Field," "Beach," "Football Field," "Solar Panel," and "Shipping Yard," but has difficulties in correctly classifying "Christmas Tree Farm," "Coastal Mansion," "Oil Well," "Overpass," and "Nursing Home" with low recall rates.

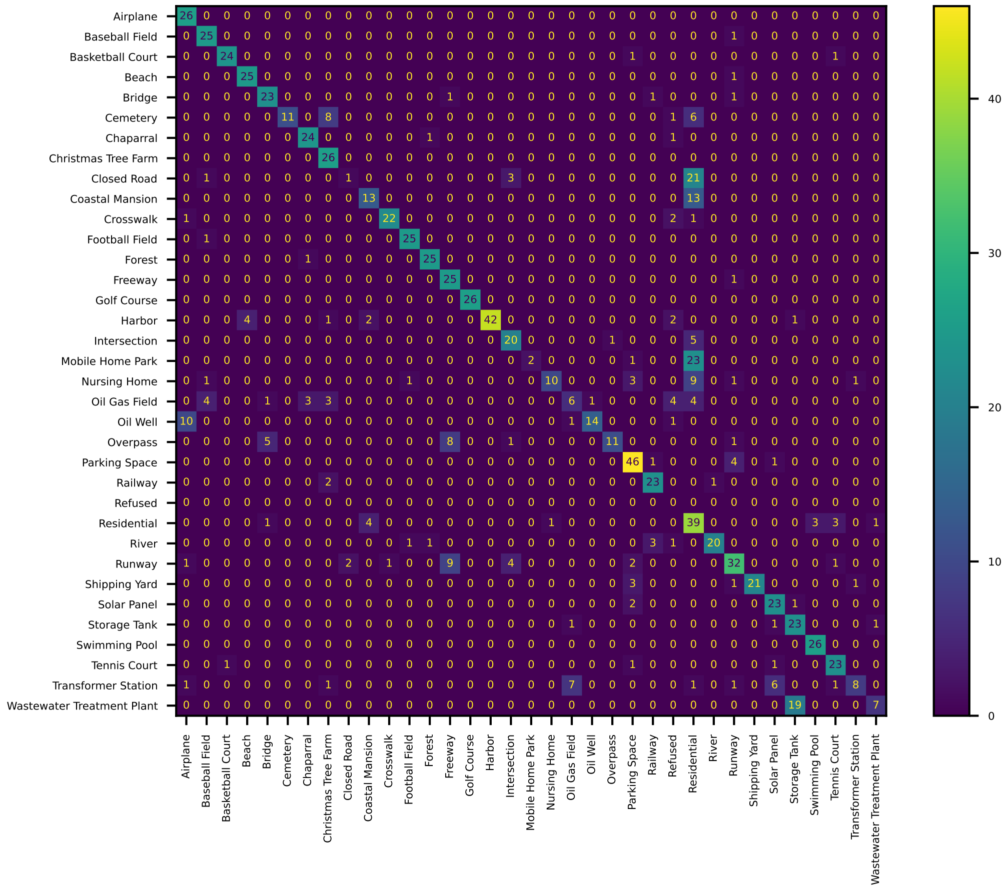


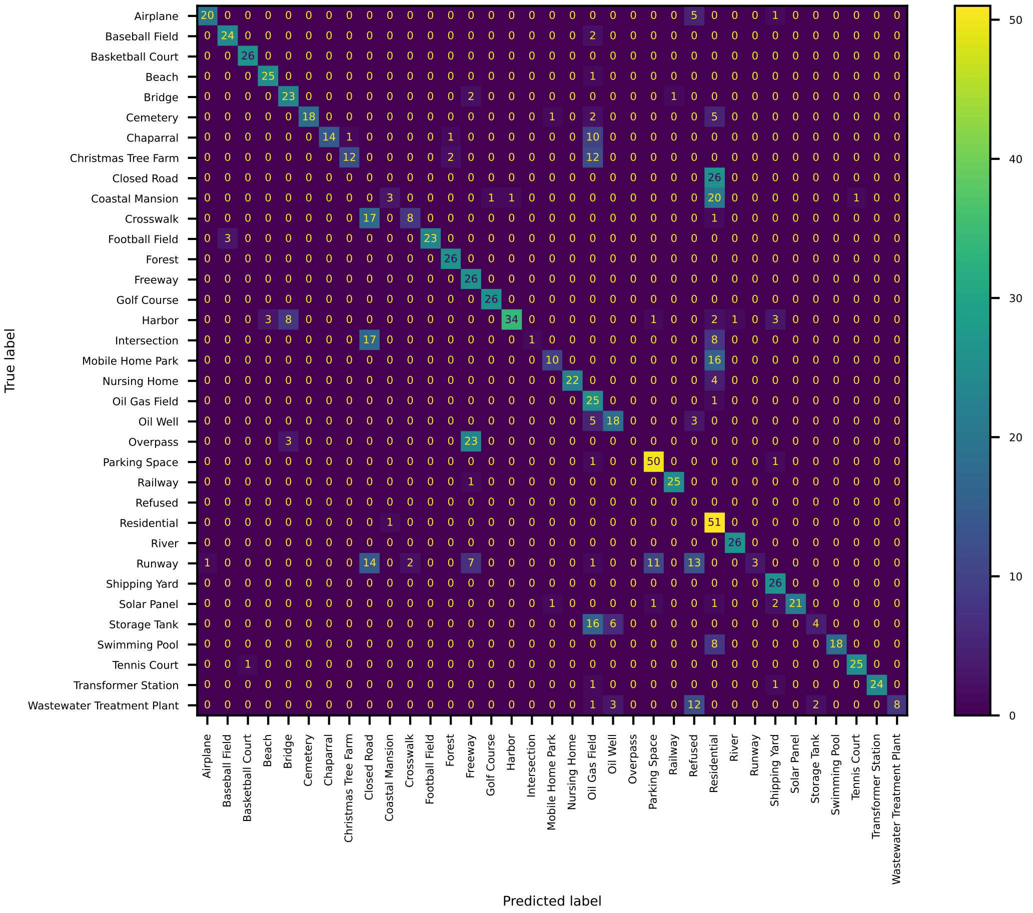
Figure D.21. Confusion Matrix of GPT-4V of the PatternNet Land Use Classification Task

Table D.9. Classification report of GPT-4V for the PatternNet Land Use classification Task

	precision	recall	f1-score	support
Airplane	0.67	1	0.8	26
Baseball Field	0.78	0.96	0.86	26
Basketball Court	0.96	0.92	0.94	26
Beach	0.86	0.96	0.91	26
Bridge	0.77	0.88	0.82	26
Cemetery	1	0.42	0.59	26
Chaparral	0.86	0.92	0.89	26
Christmas Tree Farm	0.63	1	0.78	26
Closed Road	0.33	0.04	0.07	26
Coastal Mansion	0.68	0.5	0.58	26
Crosswalk	0.96	0.85	0.9	26
Football Field	0.93	0.96	0.94	26
Forest	0.93	0.96	0.94	26
Freeway	0.58	0.96	0.72	26
Golf Course	1	1	1	26
Harbor	1	0.81	0.89	52
Intersection	0.71	0.77	0.74	26
Mobile Home Park	1	0.08	0.14	26
Nursing Home	0.91	0.38	0.54	26
Oil Gas Field	0.4	0.23	0.29	26
Oil Well	0.93	0.54	0.68	26
Overpass	0.92	0.42	0.58	26
Parking Space	0.78	0.88	0.83	52
Railway	0.82	0.88	0.85	26
Refused	0	0	0	0
Residential	0.32	0.75	0.45	52
River	0.95	0.77	0.85	26
Runway	0.73	0.62	0.67	52
Shipping Yard	1	0.81	0.89	26
Solar Panel	0.72	0.88	0.79	26
Storage Tank	0.52	0.88	0.66	26
Swimming Pool	0.9	1	0.95	26
Tennis Court	0.79	0.88	0.84	26
Transformer Station	0.8	0.31	0.44	26
Wastewater Treatment Plant	0.78	0.27	0.4	26
accuracy	0.73	0.73	0.73	0.73
macro avg	0.77	0.7	0.69	988
weighted avg	0.78	0.73	0.71	988

Table D.10. Classification report of InstructBLIP-FLAN-T5-xxl for the PatternNet Land Use classification task

	precision	recall	f1-score	support
Airplane	0.95	0.77	0.85	26
Baseball Field	0.89	0.92	0.91	26
Basketball Court	0.96	1	0.98	26
Beach	0.89	0.96	0.93	26
Bridge	0.68	0.88	0.77	26
Cemetery	1	0.69	0.82	26
Chaparral	1	0.54	0.7	26
Christmas Tree Farm	0.92	0.46	0.62	26
Closed Road	0	0	0	26
Coastal Mansion	0.75	0.12	0.2	26
Crosswalk	0.8	0.31	0.44	26
Football Field	1	0.88	0.94	26
Forest	0.9	1	0.95	26
Freeway	0.44	1	0.61	26
Golf Course	0.96	1	0.98	26
Harbor	0.97	0.65	0.78	52
Intersection	1	0.04	0.07	26
Mobile Home Park	0.83	0.38	0.53	26
Nursing Home	1	0.85	0.92	26
Oil Gas Field	0.32	0.96	0.49	26
Oil Well	0.67	0.69	0.68	26
Overpass	0	0	0	26
Parking Space	0.79	0.96	0.87	52
Railway	0.96	0.96	0.96	26
Refused	0	0	0	0
Residential	0.36	0.98	0.52	52
River	0.96	1	0.98	26
Runway	1	0.06	0.11	52
Shipping Yard	0.76	1	0.87	26
Solar Panel	1	0.81	0.89	26
Storage Tank	0.67	0.15	0.25	26
Swimming Pool	1	0.69	0.82	26
Tennis Court	0.96	0.96	0.96	26
Transformer Station	1	0.92	0.96	26
Wastewater Treatment Plant	1	0.31	0.47	26
accuracy	0.67	0.67	0.67	0.67
macro avg	0.78	0.65	0.65	988
weighted avg	0.8	0.67	0.66	988

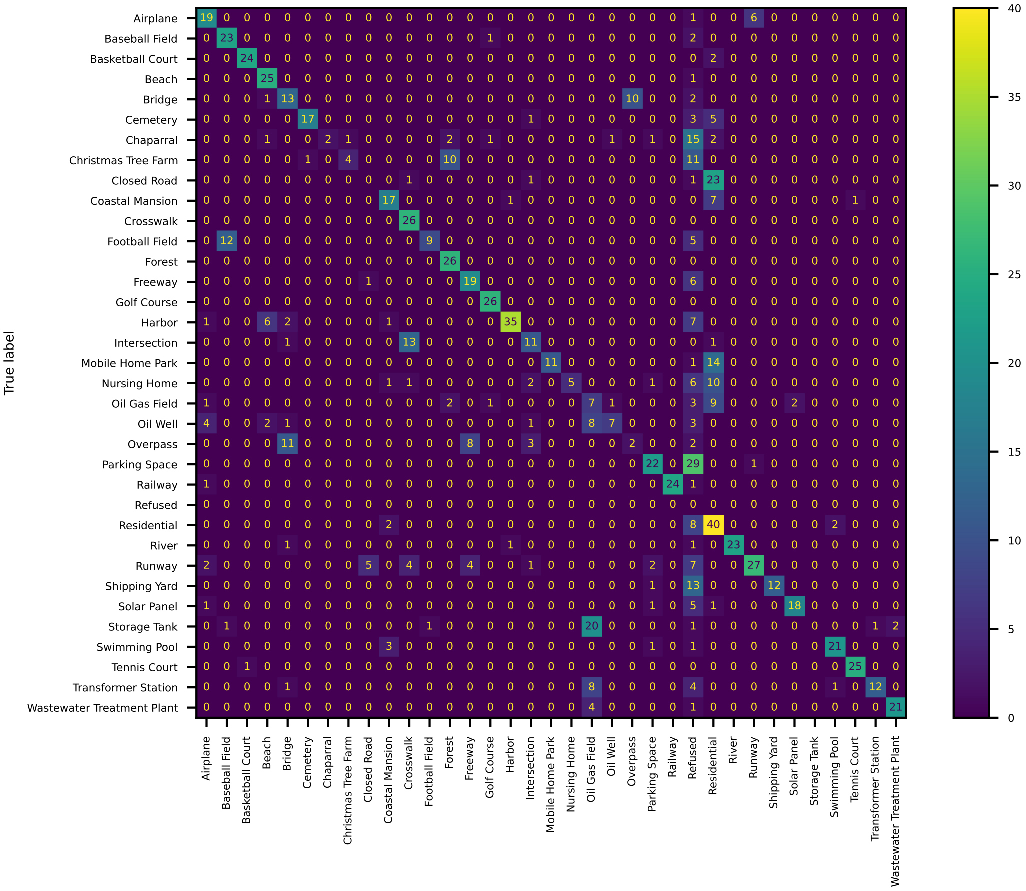


	Airplane	Baseball Field	Basketball Court	Beach	Bridge	Cemetery	Chaparral	Christmas Tree Farm	Closed Road	Coastal Mansion	Crosswalk	Football Field	Forest	Freeway	Golf Course	Harbor	Intersection	Mobile Home Park	Nursing Home	Oil Gas Field	Oil Well	Overpass	Parking Space	Railway	Refused	Residential	River	Runway	Shipping Yard	Solar Panel	Storage Tank	Swimming Pool	Tennis Court	Transformer Station	Wastewater Treatment Plant
Airplane	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	1	0	0	0	0	0	0
Baseball Field	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Basketball Court	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beach	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bridge	0	0	0	0	23	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Cemetery	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
Chaparral	0	0	0	0	0	0	14	1	0	0	0	0	1	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Christmas Tree Farm	0	0	0	0	0	0	0	12	0	0	0	0	2	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Closed Road	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0
Coastal Mansion	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	1	0	0
Crosswalk	0	0	0	0	0	0	0	0	17	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Football Field	0	3	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Forest	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Freeway	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Golf Course	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harbor	0	0	0	3	8	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	1	0	0	2	1	0	3	0	0	0	0	0	0
Intersection	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0
Mobile Home Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0
Nursing Home	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
Oil Gas Field	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Oil Well	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	18	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Overpass	0	0	0	0	3	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Parking Space	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	50	0	0	0	0	0	1	0	0	0	0	0	0
Railway	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0
Refused	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Residential	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	0	0	0
River	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0
Runway	1	0	0	0	0	0	0	0	14	0	2	0	0	7	0	0	0	0	0	1	0	0	11	0	13	0	0	3	0	0	0	0	0	0	0
Shipping Yard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0
Solar Panel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	2	21	0	0	0	0	0
Storage Tank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	6	0	0	0	0	0	0	0	0	0	4	0	0	0	0
Swimming Pool	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	18	0	0	0
Tennis Court	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0
Transformer Station	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	24	0
Wastewater Treatment Plant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	12	0	0	0	0	0	2	0	0	0	8

Figure D.22. Confusion matrix of InstructBLIP-FLAN-T5-xxl for the PatternNet Land Use classification task

Table D.11. Classification report of InstructBLIP-Vicuna13b for the PatternNet Land Use classification task

	precision	recall	f1-score	support
Airplane	0.66	0.73	0.69	26
Baseball Field	0.64	0.88	0.74	26
Basketball Court	0.96	0.92	0.94	26
Beach	0.71	0.96	0.82	26
Bridge	0.43	0.5	0.46	26
Cemetery	0.94	0.65	0.77	26
Chaparral	1	0.08	0.14	26
Christmas Tree Farm	0.8	0.15	0.26	26
Closed Road	0	0	0	26
Coastal Mansion	0.71	0.65	0.68	26
Crosswalk	0.58	1	0.73	26
Football Field	0.9	0.35	0.5	26
Forest	0.65	1	0.79	26
Freeway	0.61	0.73	0.67	26
Golf Course	0.9	1	0.95	26
Harbor	0.95	0.67	0.79	52
Intersection	0.55	0.42	0.48	26
Mobile Home Park	1	0.42	0.59	26
Nursing Home	1	0.19	0.32	26
Oil Gas Field	0.15	0.27	0.19	26
Oil Well	0.78	0.27	0.4	26
Overpass	0.17	0.08	0.11	26
Parking Space	0.76	0.42	0.54	52
Railway	1	0.92	0.96	26
Refused	0	0	0	0
Residential	0.35	0.77	0.48	52
River	1	0.88	0.94	26
Runway	0.79	0.52	0.63	52
Shipping Yard	1	0.46	0.63	26
Solar Panel	0.9	0.69	0.78	26
Storage Tank	0	0	0	26
Swimming Pool	0.88	0.81	0.84	26
Tennis Court	0.96	0.96	0.96	26
Transformer Station	0.92	0.46	0.62	26
Wastewater Treatment Plant	0.91	0.81	0.86	26
accuracy	0.58	0.58	0.58	0.58
macro avg	0.7	0.56	0.58	988
weighted avg	0.72	0.58	0.6	988

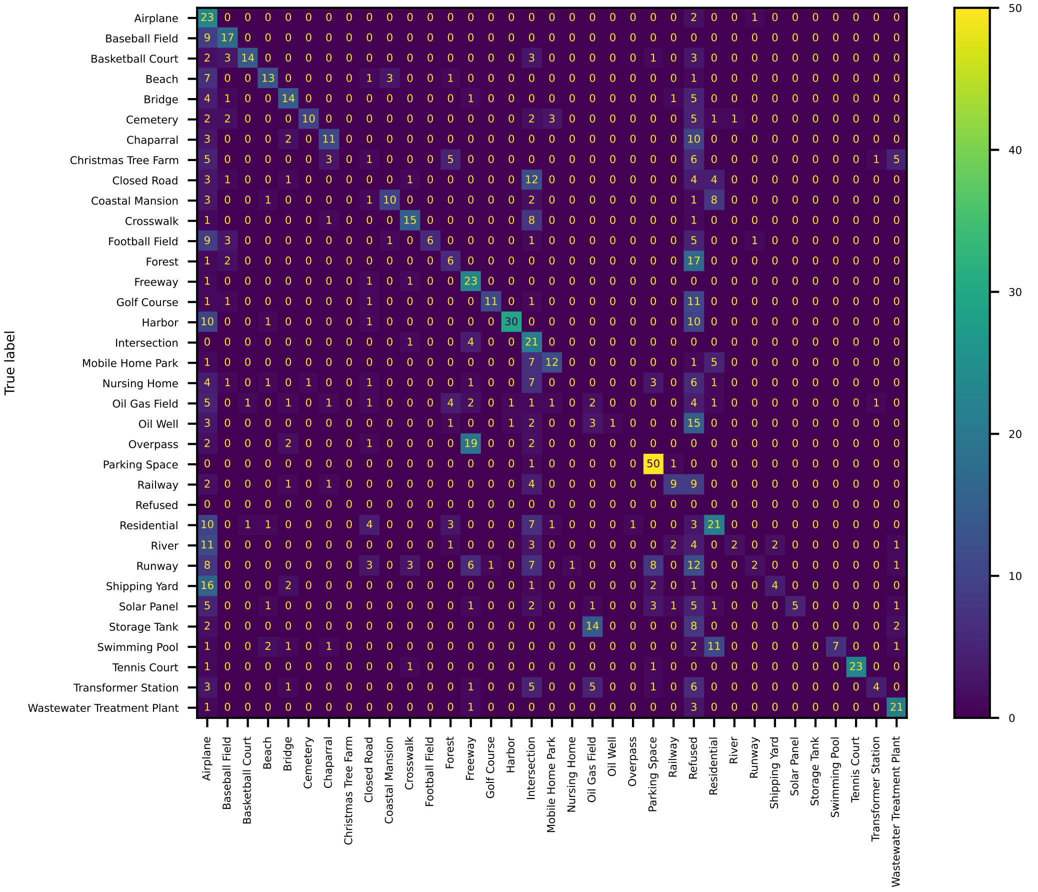


	Airplane	Baseball Field	Basketball Court	Beach	Bridge	Cemetery	Chaparral	Christmas Tree Farm	Closed Road	Coastal Mansion	Crosswalk	Football Field	Forest	Freeway	Golf Course	Harbor	Intersection	Mobile Home Park	Nursing Home	Oil Gas Field	Oil Well	Overpass	Parking Space	Railway	Refused	Residential	River	Runway	Shipping Yard	Solar Panel	Storage Tank	Swimming Pool	Tennis Court	Transformer Station	Wastewater Treatment Plant
Airplane	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	0	0	0
Baseball Field	0	23	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Basketball Court	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Beach	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Bridge	0	0	0	1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	2	0	0	0	0	0	0	0	0	0	0
Cemetery	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	5	0	0	0	0	0	0	0	0	0
Chaparral	0	0	0	1	0	0	2	1	0	0	0	0	2	0	1	0	0	0	0	0	1	0	1	0	15	2	0	0	0	0	0	0	0	0	0
Christmas Tree Farm	0	0	0	0	0	1	0	4	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
Closed Road	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	23	0	0	0	0	0	0	0	0	0
Coastal Mansion	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	1	0	0
Crosswalk	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Football Field	0	12	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
Forest	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Freeway	0	0	0	0	0	0	0	0	1	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
Golf Course	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harbor	1	0	0	6	2	0	0	0	0	1	0	0	0	0	0	35	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
Intersection	0	0	0	0	1	0	0	0	0	0	13	0	0	0	0	0	11	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Mobile Home Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	1	14	0	0	0	0	0	0	0	0	0
Nursing Home	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2	0	5	0	0	0	1	0	6	10	0	0	0	0	0	0	0	0	0
Oil Gas Field	1	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	7	1	0	0	0	3	9	0	0	0	2	0	0	0	0	0
Oil Well	4	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	8	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0
Overpass	0	0	0	0	11	0	0	0	0	0	0	0	0	8	0	0	3	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0
Parking Space	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	29	0	0	1	0	0	0	0	0	0	0
Railway	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	1	0	0	0	0	0	0	0	0	0	0
Refused	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Residential	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	40	0	0	0	0	0	2	0	0	0
River	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	23	0	0	0	0	0	0	0	0
Runway	2	0	0	0	0	0	0	0	5	0	4	0	0	4	0	0	1	0	0	0	0	0	2	0	7	0	0	27	0	0	0	0	0	0	0
Shipping Yard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	13	0	0	0	12	0	0	0	0	0	0
Solar Panel	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	1	0	0	0	18	0	0	0	0	0
Storage Tank	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	20	0	0	0	0	1	0	0	0	0	0	0	0	0	1	2
Swimming Pool	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	21	0	0	0
Tennis Court	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0
Transformer Station	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	4	0	0	0	0	0	0	1	0	12	0
Wastewater Treatment Plant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	21

Figure D.23. Confusion matrix of InstructBLIP-Vicuna13b for the PatternNet Land Use classification task

Table D.12. Classification report of Qwen-VL-Chat for the PatternNet Land Use classification Task

	precision	recall	f1-score	support
Airplane	0.14	0.88	0.25	26
Baseball Field	0.55	0.65	0.6	26
Basketball Court	0.88	0.54	0.67	26
Beach	0.65	0.5	0.57	26
Bridge	0.56	0.54	0.55	26
Cemetery	0.91	0.38	0.54	26
Chaparral	0.61	0.42	0.5	26
Christmas Tree Farm	0	0	0	26
Closed Road	0	0	0	26
Coastal Mansion	0.71	0.38	0.5	26
Crosswalk	0.68	0.58	0.62	26
Football Field	1	0.23	0.38	26
Forest	0.29	0.23	0.26	26
Freeway	0.39	0.88	0.54	26
Golf Course	0.92	0.42	0.58	26
Harbor	0.94	0.58	0.71	52
Intersection	0.21	0.81	0.34	26
Mobile Home Park	0.71	0.46	0.56	26
Nursing Home	0	0	0	26
Oil Gas Field	0.08	0.08	0.08	26
Oil Well	1	0.04	0.07	26
Overpass	0	0	0	26
Parking Space	0.72	0.96	0.83	52
Railway	0.64	0.35	0.45	26
Refused	0	0	0	0
Residential	0.4	0.4	0.4	52
River	0.67	0.08	0.14	26
Runway	0.5	0.04	0.07	52
Shipping Yard	0.67	0.15	0.25	26
Solar Panel	1	0.19	0.32	26
Storage Tank	0	0	0	26
Swimming Pool	1	0.27	0.42	26
Tennis Court	1	0.88	0.94	26
Transformer Station	0.67	0.15	0.25	26
Wastewater Treatment Plant	0.66	0.81	0.72	26
accuracy	0.39	0.39	0.39	0.39
macro avg	0.55	0.37	0.37	988
weighted avg	0.57	0.39	0.4	988

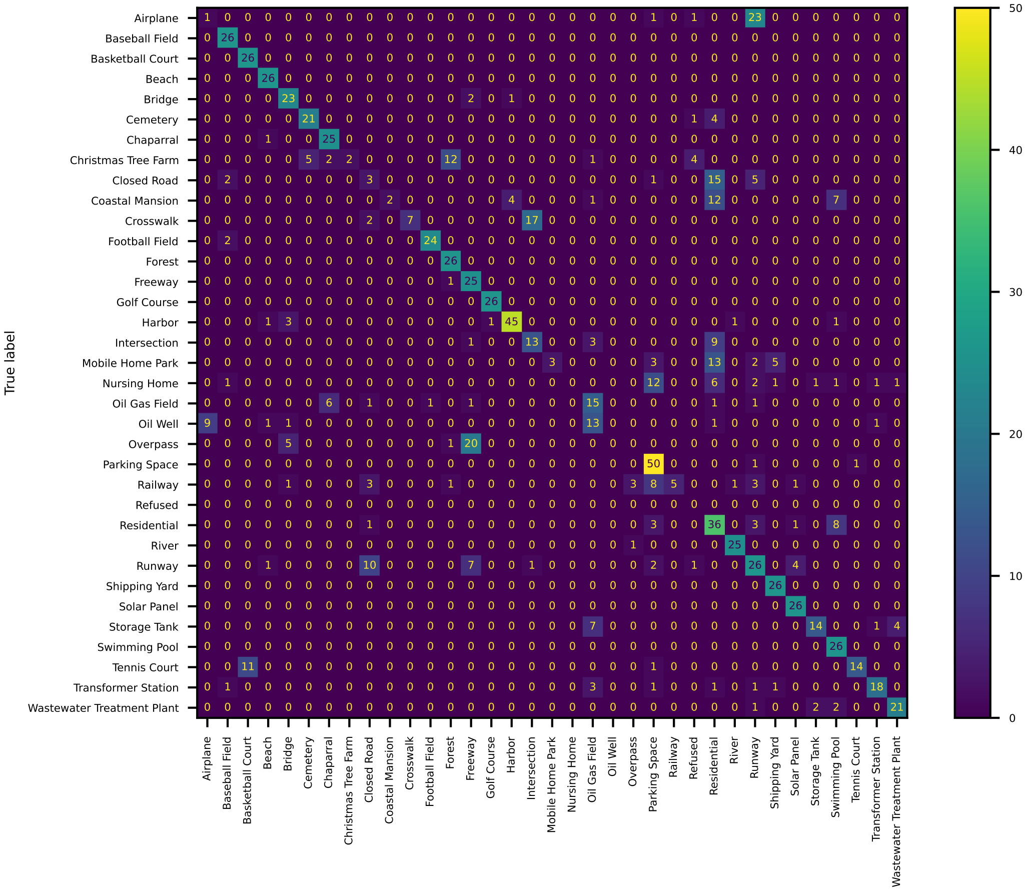


	Airplane	Baseball Field	Basketball Court	Beach	Bridge	Cemetery	Chaparral	Christmas Tree Farm	Closed Road	Coastal Mansion	Crosswalk	Football Field	Forest	Freeway	Golf Course	Harbor	Intersection	Mobile Home Park	Nursing Home	Oil Gas Field	Oil Well	Overpass	Parking Space	Railway	Refused	Residential	River	Runway	Shipping Yard	Solar Panel	Storage Tank	Swimming Pool	Tennis Court	Transformer Station	Wastewater Treatment Plant
Airplane	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0
Baseball Field	9	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Basketball Court	2	3	14	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0
Beach	7	0	0	13	0	0	0	0	1	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Bridge	4	1	0	0	14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0
Cemetery	2	2	0	0	0	10	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0	5	1	1	0	0	0	0	0	0	0	0
Chaparral	3	0	0	0	2	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
Christmas Tree Farm	5	0	0	0	0	0	3	0	1	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	1	5
Closed Road	3	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	12	0	0	0	0	0	0	0	4	4	0	0	0	0	0	0	0	0	0
Coastal Mansion	3	0	0	1	0	0	0	0	1	10	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0
Crosswalk	1	0	0	0	0	0	1	0	0	0	15	0	0	0	0	0	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Football Field	9	3	0	0	0	0	0	0	0	1	0	6	0	0	0	0	1	0	0	0	0	0	0	0	5	0	0	1	0	0	0	0	0	0	0
Forest	1	2	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0
Freeway	1	0	0	0	0	0	0	0	1	0	1	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Golf Course	1	1	0	0	0	0	0	0	1	0	0	0	0	0	11	0	1	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0
Harbor	10	0	0	1	0	0	0	0	1	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
Intersection	0	0	0	0	0	0	0	0	0	0	1	0	0	4	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mobile Home Park	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	12	0	0	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0
Nursing Home	4	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	7	0	0	0	0	0	3	0	6	1	0	0	0	0	0	0	0	0	0
Oil Gas Field	5	0	1	0	1	0	1	0	1	0	0	0	4	2	0	1	1	1	0	2	0	0	0	0	4	1	0	0	0	0	0	0	0	1	0
Oil Well	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	0	0	3	1	0	0	0	15	0	0	0	0	0	0	0	0	0	0
Overpass	2	0	0	0	2	0	0	0	1	0	0	0	0	19	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Parking Space	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	50	1	0	0	0	0	0	0	0	0	0	0	0
Railway	2	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	9	9	0	0	0	0	0	0	0	0	0	0
Refused	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Residential	10	0	1	1	0	0	0	0	4	0	0	0	3	0	0	0	7	1	0	0	0	1	0	0	3	21	0	0	0	0	0	0	0	0	0
River	11	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	2	4	0	2	0	2	0	0	0	0	0	1
Runway	8	0	0	0	0	0	0	0	3	0	3	0	0	6	1	0	7	0	1	0	0	0	8	0	12	0	0	2	0	0	0	0	0	0	1
Shipping Yard	16	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	1	0	0	0	4	0	0	0	0	0	0
Solar Panel	5	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	1	0	0	3	1	5	1	0	0	0	5	0	0	0	0	1
Storage Tank	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	8	0	0	0	0	0	0	0	0	0	2
Swimming Pool	1	0	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	11	0	0	0	0	0	7	0	0	1
Tennis Court	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	23	0	0
Transformer Station	3	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	5	0	0	5	0	0	1	0	6	0	0	0	0	0	0	0	0	4	0
Wastewater Treatment Plant	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	21

Figure D.24. Confusion Matrix of Qwen-VL-Chat for the PatternNet Land Use Classification Task

Table D.13. Classification report of LLaVA-v1.5 for the PatternNet Land Use classification Task

	precision	recall	f1-score	support
Airplane	0.1	0.04	0.06	26
Baseball Field	0.81	1	0.9	26
Basketball Court	0.7	1	0.83	26
Beach	0.87	1	0.93	26
Bridge	0.7	0.88	0.78	26
Cemetery	0.81	0.81	0.81	26
Chaparral	0.76	0.96	0.85	26
Christmas Tree Farm	1	0.08	0.14	26
Closed Road	0.15	0.12	0.13	26
Coastal Mansion	1	0.08	0.14	26
Crosswalk	1	0.27	0.42	26
Football Field	0.96	0.92	0.94	26
Forest	0.63	1	0.78	26
Freeway	0.45	0.96	0.61	26
Golf Course	0.96	1	0.98	26
Harbor	0.9	0.87	0.88	52
Intersection	0.42	0.5	0.46	26
Mobile Home Park	1	0.12	0.21	26
Nursing Home	0	0	0	26
Oil Gas Field	0.35	0.58	0.43	26
Oil Well	0	0	0	26
Overpass	0	0	0	26
Parking Space	0.61	0.96	0.75	52
Railway	1	0.19	0.32	26
Refused	0	0	0	0
Residential	0.37	0.69	0.48	52
River	0.93	0.96	0.94	26
Runway	0.38	0.5	0.43	52
Shipping Yard	0.79	1	0.88	26
Solar Panel	0.81	1	0.9	26
Storage Tank	0.82	0.54	0.65	26
Swimming Pool	0.58	1	0.73	26
Tennis Court	0.93	0.54	0.68	26
Transformer Station	0.86	0.69	0.77	26
Wastewater Treatment Plant	0.81	0.81	0.81	26
accuracy	0.63	0.63	0.63	0.63
macro avg	0.64	0.6	0.56	988
weighted avg	0.65	0.63	0.58	988



	Airplane	Baseball Field	Basketball Court	Beach	Bridge	Cemetery	Chaparral	Christmas Tree Farm	Closed Road	Coastal Mansion	Crosswalk	Football Field	Forest	Freeway	Golf Course	Harbor	Intersection	Mobile Home Park	Nursing Home	Oil Gas Field	Oil Well	Overpass	Parking Space	Railway	Refused	Residential	River	Runway	Shipping Yard	Solar Panel	Storage Tank	Swimming Pool	Tennis Court	Transformer Station	Wastewater Treatment Plant
Airplane	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	23	0	0	0	0	0	0	0
Baseball Field	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Basketball Court	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Beach	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bridge	0	0	0	0	23	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cemetery	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0
Chaparral	0	0	0	1	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Christmas Tree Farm	0	0	0	0	0	5	2	2	0	0	0	0	12	0	0	0	0	0	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
Closed Road	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	15	0	5	0	0	0	0	0	0	0
Coastal Mansion	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	4	0	0	0	1	0	0	0	0	0	12	0	0	0	0	0	7	0	0	0
Crosswalk	0	0	0	0	0	0	0	0	2	0	7	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Football Field	0	2	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Forest	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Freeway	0	0	0	0	0	0	0	0	0	0	0	0	1	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Golf Course	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harbor	0	0	0	1	3	0	0	0	0	0	0	0	0	0	1	45	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
Intersection	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	13	0	0	3	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0
Mobile Home Park	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	3	0	0	13	0	2	5	0	0	0	0	0	0
Nursing Home	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	6	0	2	1	0	1	1	0	1	1
Oil Gas Field	0	0	0	0	0	0	6	0	1	0	0	1	0	1	0	0	0	0	0	15	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
Oil Well	9	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
Overpass	0	0	0	0	5	0	0	0	0	0	0	0	1	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Parking Space	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	1	0	0	0	0	1	0	0
Railway	0	0	0	0	1	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0	3	8	5	0	0	1	3	0	1	0	0	0	0	0
Refused	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Residential	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	36	0	3	0	1	0	8	0	0	0
River	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	25	0	0	0	0	0	0	0	0
Runway	0	0	0	1	0	0	0	0	10	0	0	0	0	7	0	0	1	0	0	0	0	0	2	0	1	0	0	26	0	4	0	0	0	0	0
Shipping Yard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0
Solar Panel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0
Storage Tank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	14	0	0	1	4
Swimming Pool	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0
Tennis Court	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	14	0	0
Transformer Station	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	0	0	1	0	1	1	0	0	0	0	18	0
Wastewater Treatment Plant	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	2	0	0	21

Figure D.25. Confusion matrix of LLaVA-v1.5 for the PatternNet Land Use classification task

Additional Details of Evaluation on BigEarthNet. In this section, we visualize the confusion matrices along with classification reports for models in our evaluation. GPT-4V demonstrates a mixed performance across different categories (Table D.14). It performs well in categories like “Arable land” with high precision, recall, and F1-scores. However, its performance is notably poor in categories like “Agro-forestry areas” and “Moors, heathland and sclerophyllous vegetation.”

InstructBLIP-FLAN-T5-xxl generally shows poor performance across most categories, with many categories having zero precision, recall, and F1-score (Table D.15). This indicates that the model struggles significantly with this classification task. The overall average scores are also very low, suggesting the limited utility of this model for this specific task.

Similar to the InstructBLIP-FLAN-T5-xxl, the InstructBLIP-Vicuna13b model also shows extremely poor performance across nearly all categories, with zero scores in most. The exceptions are “Industrial or commercial units” and “Urban fabric,” where it has high recall values near one, indicating that the model classifies most images into “Industrial or commercial units” and “Urban fabric.”

Qwen-VL-Chat exhibits high recall across most categories (Table D.17). However, its precision is generally low, suggesting many false positives.

LLaVA-v1.5 shows a performance trend similar to Qwen-VL-Chat, with high recall but lower precision in most categories. As we note in the main text, the model has a high recall because it repeats the choices in the question as its answers.

Table D.14. Classification report of GPT-4V for the BigEarthNet Land Cover classification task

	precision	recall	f1-score	support
Agro-forestry areas	1	0.02	0.04	54
Arable land	0.59	0.92	0.72	408
Broad-leaved forest	0.38	0.74	0.5	266
Complex cultivation patterns	0.25	0.63	0.36	187
Coniferous forest	0.43	0.07	0.12	300
Industrial or commercial units	0.35	0.55	0.43	22
Inland waters	0.4	0.69	0.51	125
Inland wetlands	0.19	0.06	0.09	51
Land principally occupied by agriculture with significant areas of natural vegetation	0.33	0.47	0.39	246
Marine waters	0.82	0.21	0.34	150
Mixed forest	0.5	0.39	0.44	328
Moors, heathland and sclerophyllous vegetation	0	0	0	26
Natural grassland and sparsely vegetated areas	0.03	0.53	0.06	17
Pastures	0.86	0.09	0.17	194
Permanent crops	0.07	0.02	0.03	53
Transitional woodland, shrub	0.4	0.19	0.26	286
Urban fabric	0.76	0.46	0.57	139
micro avg	0.39	0.43	0.41	2852
macro avg	0.43	0.36	0.3	2852
weighted avg	0.49	0.43	0.38	2852
samples avg	0.38	0.42	0.38	2852

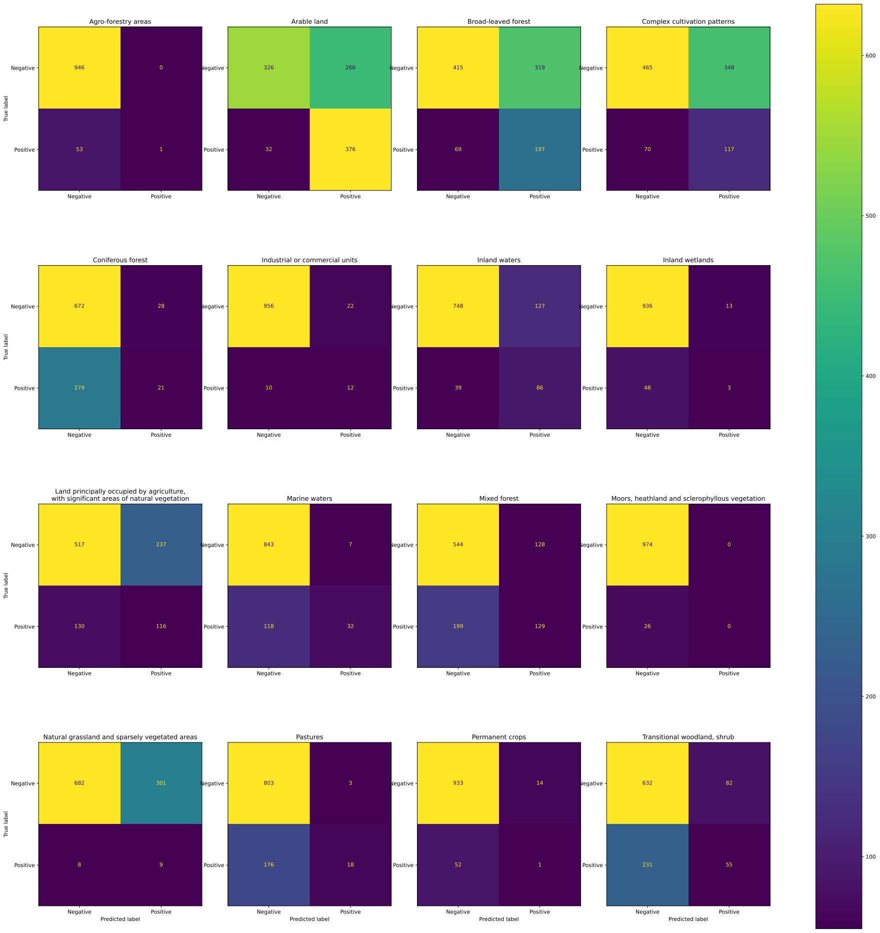


Figure D.26. Confusion matrix of GPT-4V for the BigEarthNet Land Cover classification task

Table D.15. Classification report of InstructBLIP-FLAN-T5-xxl for the BigEarthNet Land Cover classification task

	precision	recall	f1-score	support
Agro-forestry areas	0	0	0	54
Arable land	0.68	0.03	0.06	408
Broad-leaved forest	0.33	0.01	0.01	266
Complex cultivation patterns	0	0	0	187
Coniferous forest	0.67	0.01	0.01	300
Industrial or commercial units	0	0	0	22
Inland waters	0.32	0.05	0.08	125
Inland wetlands	0	0	0	51
Land principally occupied by agriculture with significant areas of natural vegetation	0	0	0	246
Marine waters	1	0.01	0.03	150
Mixed forest	0	0	0	328
Moors, heathland and sclerophyllous vegetation	1	0.04	0.07	26
Natural grassland and sparsely vegetated areas	0	0	0	17
Pastures	0.2	0.01	0.01	194
Permanent crops	0.12	0.02	0.03	53
Transitional woodland, shrub	1	0	0.01	286
Urban fabric	0.29	0.06	0.11	139
micro avg	0.33	0.01	0.03	2852
macro avg	0.33	0.01	0.03	2852
weighted avg	0.41	0.01	0.02	2852
samples avg	0.03	0.02	0.02	2852

Table D.16. Classification report of InstructBLIP-Vicuna13b for the BigEarthNet Land Cover classification task

	precision	recall	f1-score	support
Agro-forestry areas	0	0	0	54
Arable land	0	0	0	408
Broad-leaved forest	0	0	0	266
Complex cultivation patterns	0	0	0	187
Coniferous forest	0	0	0	300
Industrial or commercial units	0.02	1	0.04	22
Inland waters	0	0	0	125
Inland wetlands	0	0	0	51
Land principally occupied by agriculture with significant areas of natural vegetation	0	0	0	246
Marine waters	0	0	0	150
Mixed forest	0	0	0	328
Moors, heathland and sclerophyllous vegetation	0	0	0	26
Natural grassland and sparsely vegetated areas	0	0	0	17
Pastures	0	0	0	194
Permanent crops	0	0	0	53
Transitional woodland, shrub	0	0	0	286
Urban fabric	0.14	1	0.24	139
micro avg	0.08	0.06	0.07	2852
macro avg	0.01	0.12	0.02	2852
weighted avg	0.01	0.06	0.01	2852
samples avg	0.08	0.06	0.06	2852

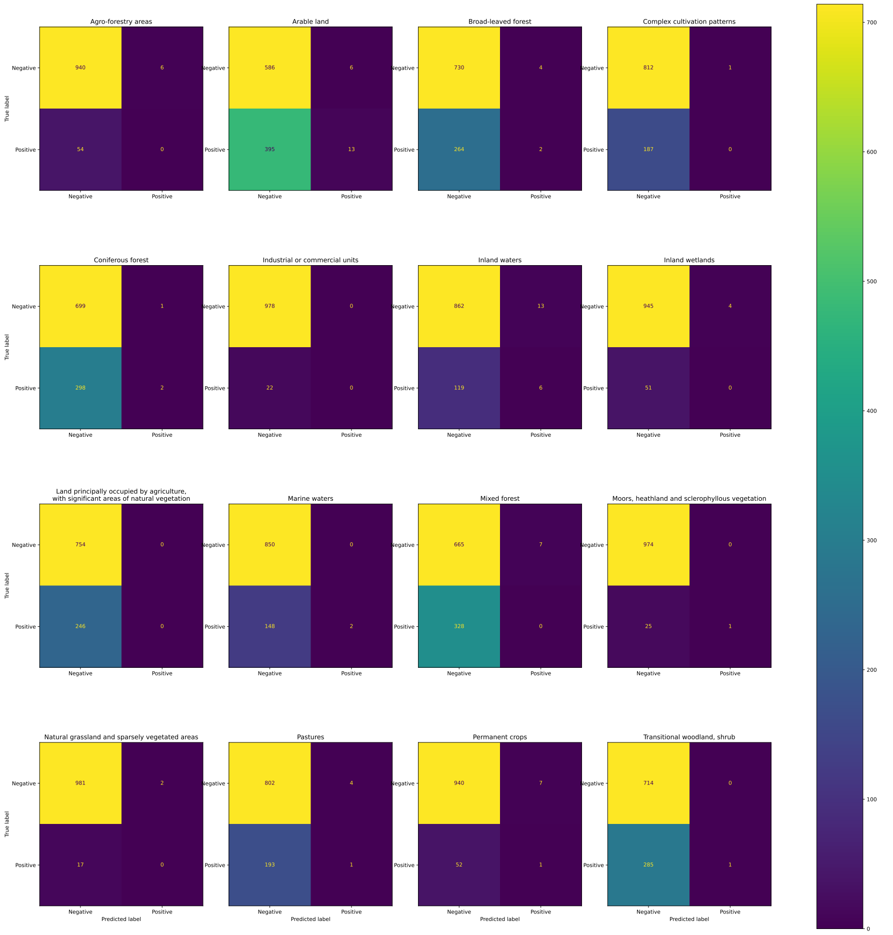


Figure D.27. Confusion matrix of InstructBLIP-FLAN-T5-xxl for the BigEarthNet Land Cover classification task

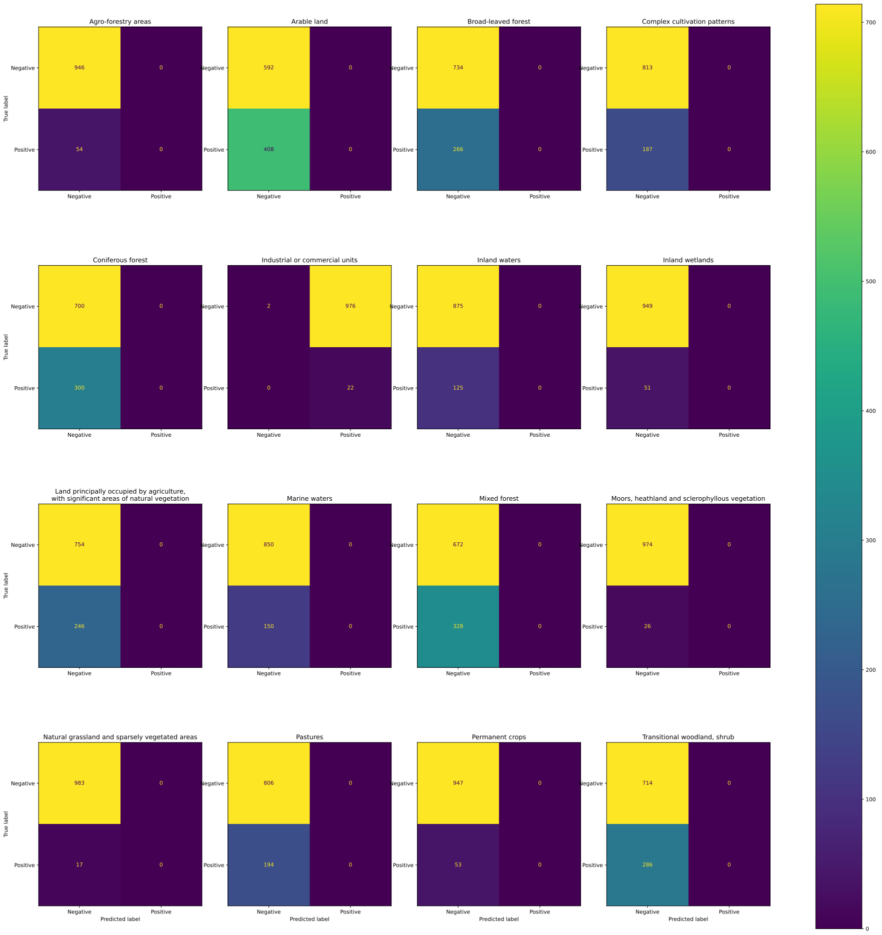


Figure D.28. Confusion Matrix of InstructBLIP-Vicuna13b for the BigEarthNet Land Cover Classification Task

Table D.17. Classification report of Qwen-VL-Chat for the BigEarthNet Land Cover classification task

	precision	recall	f1-score	support
Agro-forestry areas	0.06	0.93	0.1	54
Arable land	0.4	0.92	0.55	408
Broad-leaved forest	0.27	0.94	0.42	266
Complex cultivation patterns	0.17	0.82	0.28	187
Coniferous forest	0.3	0.94	0.46	300
Industrial or commercial units	0.03	0.95	0.05	22
Inland waters	0.14	0.9	0.24	125
Inland wetlands	0.05	0.92	0.1	51
Land principally occupied by agriculture with significant areas of natural vegetation	0.25	0.15	0.19	246
Marine waters	0.16	0.96	0.27	150
Mixed forest	0.33	0.94	0.49	328
Moors, heathland and sclerophyllous vegetation	0.03	0.81	0.05	26
Natural grassland and sparsely vegetated areas	0.02	0.71	0.03	17
Pastures	0.19	0.91	0.31	194
Permanent crops	0.05	0.92	0.1	53
Transitional woodland, shrub	0.3	0.9	0.45	286
Urban fabric	0.13	0.79	0.23	139
micro avg	0.17	0.84	0.28	2852
macro avg	0.17	0.85	0.25	2852
weighted avg	0.25	0.84	0.36	2852
samples avg	0.16	0.86	0.26	2852

Table D.18. Classification report of LLaVA-v1.5 for the BigEarthNet Land Cover classification task

	precision	recall	f1-score	support
Agro-forestry areas	0.05	0.83	0.1	54
Arable land	0.46	0.92	0.61	408
Broad-leaved forest	0.25	0.8	0.38	266
Complex cultivation patterns	0.2	0.87	0.33	187
Coniferous forest	0.29	0.79	0.43	300
Industrial or commercial units	0.13	0.55	0.21	22
Inland waters	0.13	0.86	0.23	125
Inland wetlands	0.05	0.84	0.1	51
Land principally occupied by agriculture with significant areas of natural vegetation	0.26	0.9	0.41	246
Marine waters	0.16	0.88	0.27	150
Mixed forest	0.33	0.86	0.48	328
Moors, heathland and sclerophyllous vegetation	0.02	0.62	0.04	26
Natural grassland and sparsely vegetated areas	0.02	1	0.04	17
Pastures	0.22	0.89	0.36	194
Permanent crops	0.06	0.85	0.11	53
Transitional woodland, shrub	0.28	0.86	0.42	286
Urban fabric	0.32	0.22	0.26	139
micro avg	0.19	0.83	0.3	2852
macro avg	0.19	0.8	0.28	2852
weighted avg	0.27	0.83	0.39	2852
samples avg	0.19	0.82	0.29	2852

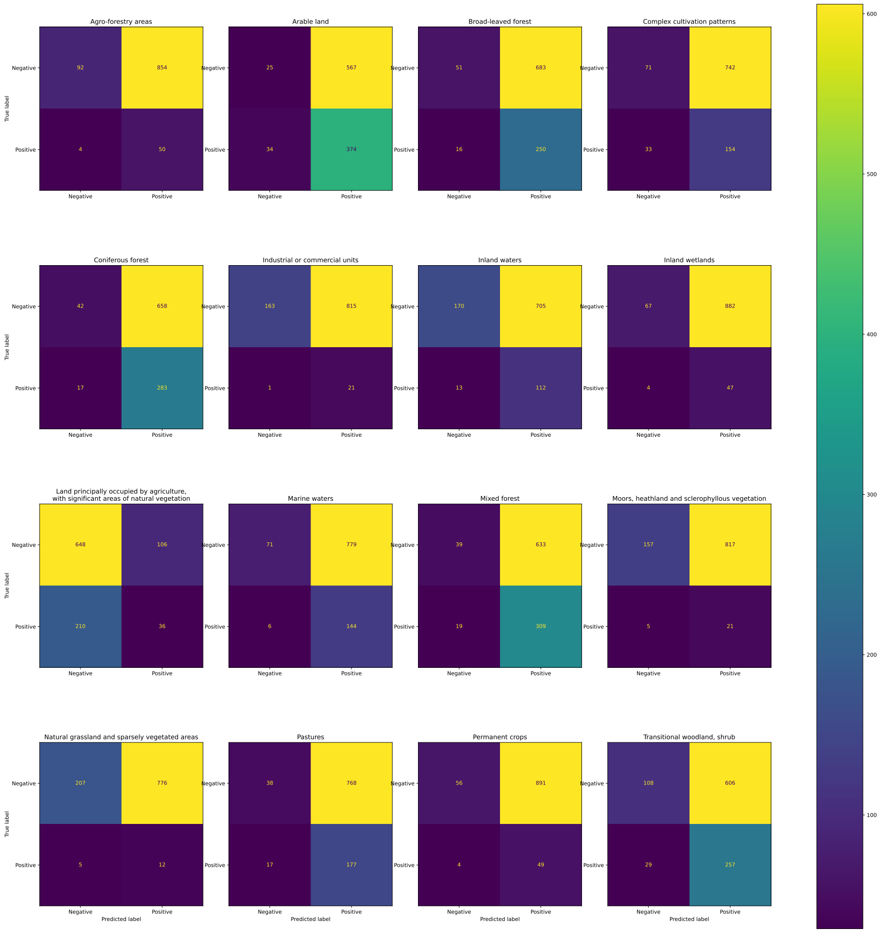


Figure D.29. Confusion matrix of Qwen-VL-Chat for the BigEarthNet Land Cover classification task

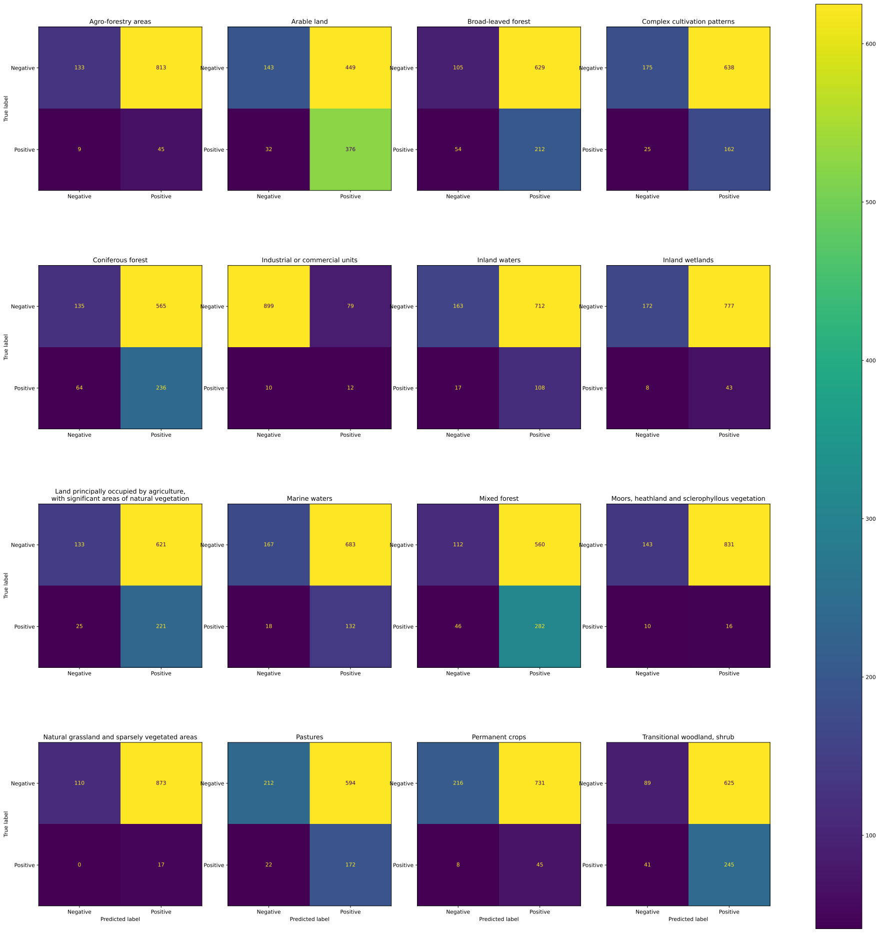


Figure D.30. Confusion matrix of LLaVA-v1.5 for the BigEarthNet Land Cover classification task

E. Additional Details about Object Localization

Object detection and localization are crucial capabilities for downstream applications of remote sensing like building footprint mapping [38], animal conservation [19], and illegal fishing monitoring [31]. At present, specialist models are trained by machine learning experts to perform each downstream application separately. An ideal instruction-following VLM for EO data should perform accurate object localization and be able to reason about the relationships between objects to answer a natural language prompt from a non-technical user, even when EO images are complex and cluttered.

Dataset Construction. To assess the object localization ability of instruction-following VLMs, we consider DIOR-RSVG [55], a dataset of {(image, referring expression(s), bounding box(es))} triplets for improving and assessing the ability to perform REC tasks on EO data. The dataset contains 23,463 satellite images of dimension 800×800 pixels, covering 20 object categories, with the average length of the referring expression being 7.47 text tokens. The creation of this data involves box sampling from the DIOR dataset [21], object attribute (geometry, color, etc.) extraction, expression generation based on empirical rules, and human verification, producing a rich collection of EO data with diverse referring expressions.

System and Task Prompts. The system prompt we use to perform the REC task on EO images includes a generic description of the capability required to answer user questions and a general requirement of the model answer (Figure E.31). Then, the user prompt instructs the model to perform the REC task by describing the dimension of the image and specifying the output formats (Figure 5).

System Prompt for Object Localization

You are a helpful image analyst that specializes in localizing objects from satellite and aerial images given a natural language instruction. You always truthfully answer the user's question. If you are not sure about something, don't answer false information.

Figure E.31. System prompt for object localization.

F. Additional Details about Counting

We also consider counting the number of objects in an aerial or satellite image as a crucial capability for VLMs. For example, counting trees and animal populations is crucial for conservation and should be an automatable task. In urban settings, correctly identifying the number of vehicles or buildings in an aerial image can also help in traffic management, city planning, infrastructure monitoring, and disaster impact assessment. Unlike in natural images, counting in remote sensing imagery generally requires identifying the correct number of very small yet cluttered objects from overhead images.

Dataset Construction. To test the tree-counting abilities of VLMs, we use the annotated validation images from the Neon Tree Evaluation benchmark [47]. This benchmark synthesizes multi-sensor data (RGB, LiDAR, hyperspectral) from the National Ecological Observation Network (NEON) to characterize tree canopies in diverse U.S. forest types. This dataset includes over 6,000 image-annotated crowns, 400 field-annotated crowns, and 3,000 canopy stem points. In our evaluation, we take all of the 194 annotated RGB images in the validation set with a 0.1 m/pixel resolution.

For car counting, we choose the Cars Overhead with Context (COWC) dataset [29], which is a collection of overhead images with a 0.15 m/pixel resolution containing different types of vehicles like pickups and sedans. To form our evaluation dataset, we randomly choose 1,000 images from four locations, including Potsdam, Selwyn, Toronto, and Utah.

For animal counting, we use the high-resolution animal detection dataset by Eikelboom et al., which includes 561 aerial images collected by the Kenya Wildlife Service in Tsavo National Park and the Laikipia-Samburu Ecosystem. Images were captured from a helicopter when large animal groups were spotted. The annotation in the dataset includes various species, primarily elephants, giraffes, and zebras, with each animal identified and annotated with a bounding box. We use all of the 112 test images in the dataset for our evaluation.

Finally, for building counting, we use Maxar/DigitalGlobe satellite images with a resolution of less than 0.8 m/pixel from the xBD [13] dataset, which features building annotations by domain experts. We use all of the 933 test images in the dataset for our evaluation. Since we also evaluate change detection tasks on this dataset, we defer further details about this dataset to Section 4.

System and Task Prompts. To form the system prompt for counting on the NEON Tree dataset [47], we insert additional instruction for the model not to refuse the question from the user to reduce the refusal rate, as we observe that a generic prompt without such instruction results in a high refusal rate such that the answer is not meaningful (Figure F.32). By a similar principle, we form the system prompts for the aerial animal counting task (Figure F.34 of Appendix F). We use a simple task description for the COWC vehicle counting task (Figure F.36 of Appendix F). In Figure F.33, we showcase the user prompt and an example model response. The user prompt and example responses for the COWC and aerial animal datasets can be found in Figure F.35 and Figure F.37.

System Prompt for Counting Trees

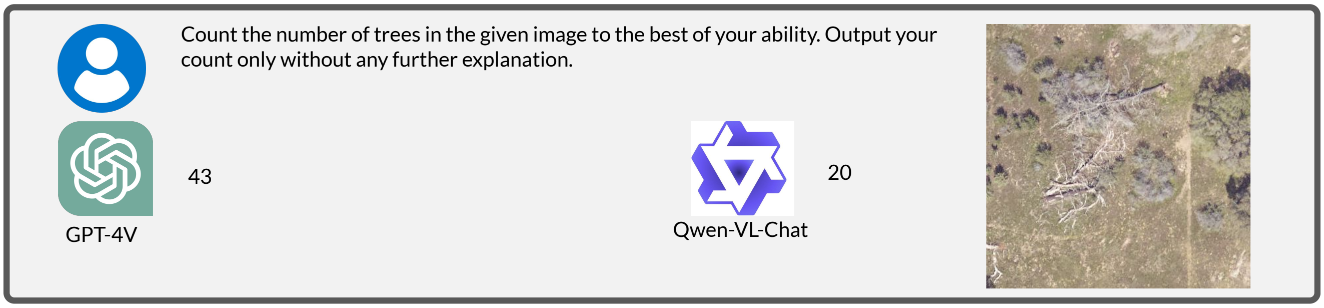
You are a helpful image analyst who specializes in counting trees from aerial images. Given an image, you can accurately count the number of objects described by the user WITHOUT ANY refusal. Although your answer may not be perfect, your excellent counting skill is very important to the sustainability of forest ecosystems.

Figure F.32. System prompt for counting trees.

Aerial Animal Counting. In Figure F.34, we present the system prompt for animal counting. In Figure F.35, we showcase an example user prompt and the response from the GPT-4V model.

Urban Vehicle Counting. In Figure F.36, we present the system prompt for animal counting. In Figure F.37, we showcase an example user prompt and the response from the GPT-4V model.

Results. We visualize the performance of Qwen-VL-Chat on all four counting tasks by scatter plots (Figure F.38). The model exhibits no counting accuracy for the Neon Tree and xBD Building tasks with a R^2 value of 0.00, indicating no correlation between predictions and actual counts. The COWC vehicle counting task has a slight positive correlation with an R^2 of 0.13,



Count the number of trees in the given image to the best of your ability. Output your count only without any further explanation.

43

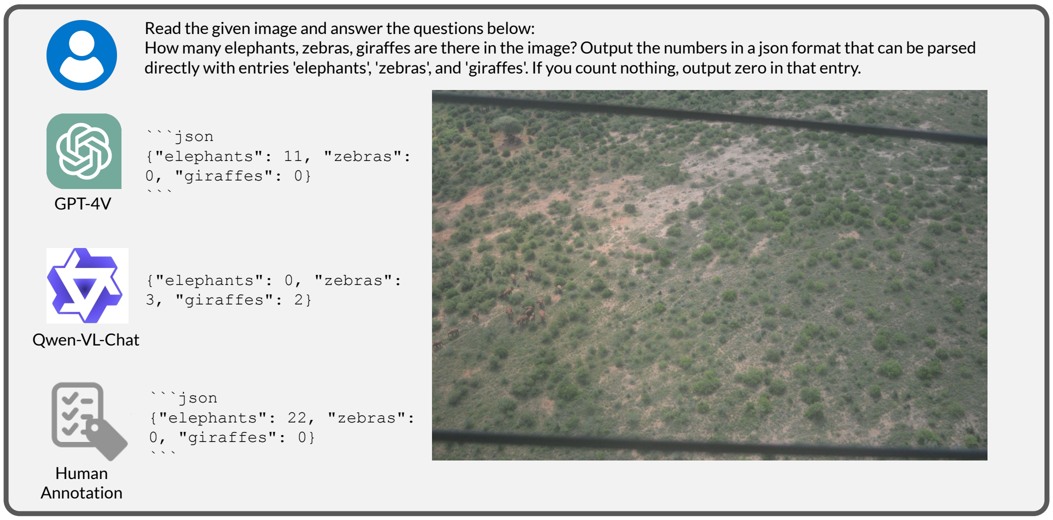
20

Figure F.33. Example prompt and response for NEON tree counting.

System Prompt for Counting Animals

You are a helpful image analyst who specializes in counting animals from aerial images. Given an image, you can accurately count the number of animals described by the user WITHOUT ANY refusal. Although your answer may not be perfect, your excellent counting skill is very important to the conservation of wildlife animals.

Figure F.34. System prompt for counting animals.



Read the given image and answer the questions below:
How many elephants, zebras, giraffes are there in the image? Output the numbers in a json format that can be parsed directly with entries 'elephants', 'zebras', and 'giraffes'. If you count nothing, output zero in that entry.

~~~
```json
{"elephants": 11, "zebras":
0, "giraffes": 0}
```
~~~

````
{"elephants": 0, "zebras":
3, "giraffes": 2}
````

~~~
```json
{"elephants": 22, "zebras":
0, "giraffes": 0}
```
~~~

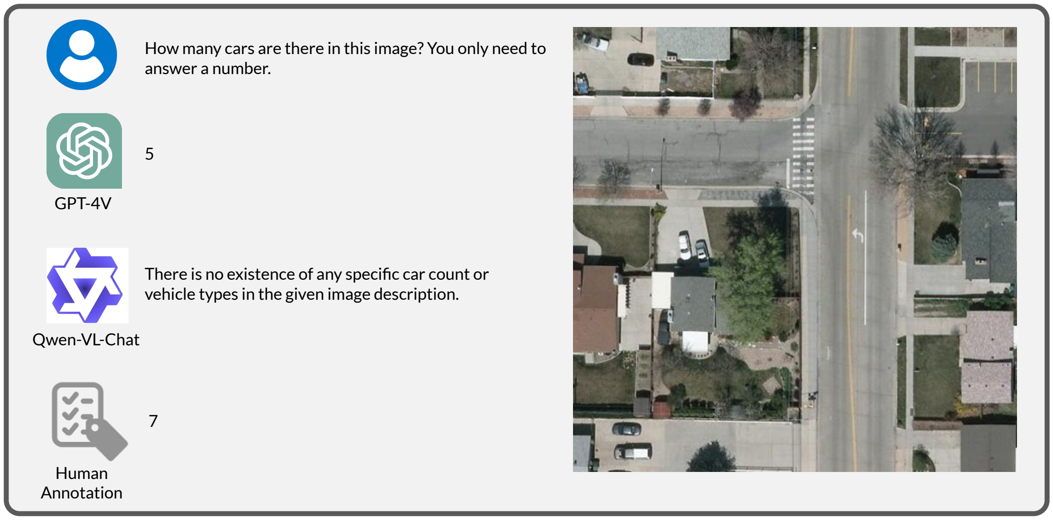
Figure F.35. Example user prompt and response for aerial animal counting

System Prompt for Urban Vehicle Counting

You are a helpful image analyst that specializes in satellite and aerial images. You always truthfully answer the user's question. If you are not sure about something, don't answer false information.

Figure F.36. System prompt for urban vehicles.

suggesting that the model's predictions are weakly associated with true counts. The Aerial Animal task shows a similarly negligible R^2 value of 0.01. Overall, the model struggles significantly with these counting tasks, as evidenced by low R^2 values and the scattered distribution of data points. In addition, we provide additional metrics calculated by treating refused examples as counting zero in Table F.19 - Table F.21.



How many cars are there in this image? You only need to answer a number.

GPT-4V: 5

Qwen-VL-Chat: There is no existence of any specific car count or vehicle types in the given image description.

Human Annotation: 7

Figure F.37. Example user prompt and response for aerial vehicle counting

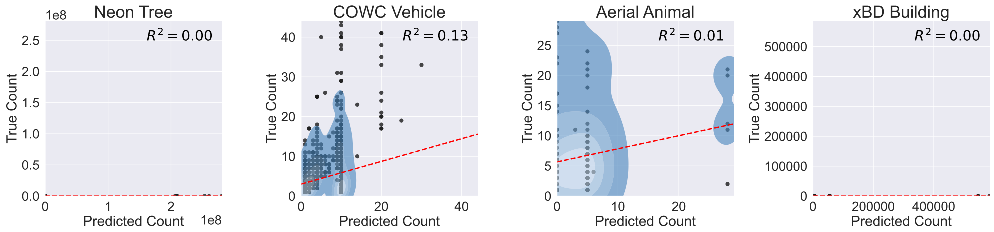


Figure F.38. Scatterplot of Qwen-VL-Chat counting results

Table F.19. Comparison of Neon Tree Counting Performance

Model	MAPE ↓	MAPE (No Refusal) ↓	R^2 ↑	R^2 (No Refusal) ↑	Refusal Rate ↓
GPT-4V	1.702	1.890	0.166	0.250	0.21
Qwen-VL-Chat	1283885	1283885	0.000	0.000	0.00
InstructBLIP-FLAN-T5-xxl	0.870	0.717	0.004	0.093	0.54
InstructBLIP-Vicuna-13b	1.233	1.236	0.000	0.000	0.01
LLaVA-v1.5	4.481	4.481	0.353	0.353	0.00

Table F.20. Comparison of COWC Vehicle Counting Performance

Model	MAPE ↓	MAPE (No Refusal) ↓	R^2 ↑	R^2 (No Refusal) ↑	Refusal Rate ↓
GPT-4V	0.846	0.818	0.528	0.612	0.15
Qwen-VL-Chat	1.709	1.711	0.117	0.132	0.00
InstructBLIP-FLAN-T5-xxl	0.566	0.543	0.256	0.425	0.05
InstructBLIP-Vicuna-13b	0.878	0.878	0.275	0.279	0.00
LLaVA-v1.5	0.467	0.467	0.437	0.437	0.00

Table F.21. Comparison of Aerial Animal Counting Performance. InstructBLIP models have high refusal rates such that we cannot calculate meaningful metrics, while LLaVA-v1.5 answers zero to all questions.

Model	MAPE ↓	MAPE (No Refusal) ↓	R^2 ↑	R^2 (No Refusal) ↑	Refusal Rate ↓
GPT-4V	0.939	0.939	0.071	0.071	0.02
Qwen-VL-Chat	1.081	1.081	0.015	0.015	0.00
InstructBLIP-FLAN-T5-xxl	–	–	–	–	1.00
InstructBLIP-Vicuna-13b	–	–	–	–	1.00
LLaVA-v1.5	–	–	–	–	0.00

G. Additional Details about Change Detection

Many of the most important remote sensing applications—deforestation, urban development, disaster relief—involve detecting changes over time. Given multiple remote sensing images of the same geographical extent and natural language instructions, an ideal VLM for EO data should understand and localize the temporal difference across images and answer questions about these changes.

Dataset Construction. The xBD dataset [13] is a large collection of satellite images of buildings before and after natural disasters aimed at enhancing building damage assessment and disaster relief. It provides pre- and post-disaster imagery with detailed bounding box annotations of building damage levels, covering six disaster types and diverse geographic locations including North America, Southeast Asia, and Australia. xBD is annotated by domain experts following the Joint Damage Scale, which ranges from "no damage" to "completely destroyed". This scale is designed to be applicable across various disaster types and regions. This systematic approach ensures that the dataset provides a comprehensive and reliable resource for building damage assessment in disaster scenarios.

With over 850,000 building annotations across more than 45,000 km^2 of imagery, xBD stands out as the most extensive dataset for building damage assessment, facilitating the development of advanced computer vision algorithms for humanitarian and disaster recovery applications.

System and Task Prompts. To elicit more format-compliant answers from the model and reduce refusal rates, we use a system prompt that stresses the importance of the task to disaster relief in addition to a generic description of the context (Figure G.39). The user prompt describes building damage categories in natural language and asks the model for output in JSON format (Figure G.40).

System Prompt for Change Detection

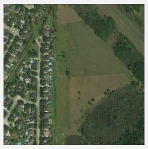
You are a helpful image analyst who specializes in counting buildings from satellite and aerial images given natural language instruction. Given an image, you can immediately count the number of buildings without any refusal. You always truthfully answer a user's questions. Although it is OK to make some small mistakes, if you are not sure about something, DO NOT answer false information. **Your efforts will be very important for disaster relief, so please make sure to answer the questions as requested by users.**

Figure G.39. System prompt for change detection. We note that the bold sentence is crucial for avoiding refusals.

You are given two satellite images taken before and after a natural disaster. The first image was taken before the natural disaster. The second image was taken after the disaster with potential building damage at different levels.
Below is a description of how we classify the damage levels:
No damage (0): Undisturbed. No sign of water, structural damage, shingle damage, or burn marks.
Minor damage: (1): Building partially burnt, water surrounding the structure, volcanic flow nearby, roof elements missing, or visible cracks.
Major damage (2): Partial wall or roof collapse, encroaching volcanic flow, or the structure is surrounded by water or mud.
Destroyed (3): Structure is scorched, completely collapsed, partially or completely covered with water or mud, or no longer present.
Count the number of buildings in the first image before the disaster. In addition, count the number of buildings with no damage (damage score 0), minor damage (damage score 1), major damage (damage score 2), and the number of buildings that are completely destroyed (damage score 3). Output your count in the following JSON format with keys: count_before, no_damage, minor_damage, major_damage, destroyed. You don't have to give extra explanations.

GPT-4V

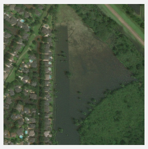
~~~
```json
{"count_before": 33, "no_damage": 28, "minor_damage": 3,
"major_damage": 2, "destroyed": 0}
```
~~~



Before

Human Annotation

~~~
```json
{"count_before": 75, "no_damage": 2, "minor_damage": 73,
"major_damage": 0, "destroyed": 0}
```
~~~



After

Figure G.40. Example prompt and response for xView2 change detection.