AGILE: GIScience Series, 7, 14, 2026. https://doi.org/10.5194/agile-giss-7-14-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Investigating the Generalizability of Segment Anything Model for Large-Scale Geospatial Segmentation

Wejdene Mansour [1], Paul Walther [1], Hao Li [2], and Martin Werner [1]

[1]Department of Aerospace and Geodesy, TUM School of Engineering and Design, Technical University of Munich, Germany
[2]Department of Geography, National University of Singapore, Singapore

Correspondence: Wejdene Mansour (wejdene.mansour@tum.de)

Abstract. Foundation Models (FMs) are promising approaches in multimodal artificial intelligence as they provide foundational task knowledge across computer vision, language understanding, and related domains. Despite their success, the extent to which FMs generalize to domain-specific tasks remains unclear, especially in Earth System Sciences (ESS). In this work, we investigate the geographical and task-level generalizability of Segment Anything Model (SAM) and the vision–language FMs CLIP and Grounding DINO, across two distinct vision tasks: 1) building footprint segmentation from high-quality airborne images at 40cm ground sampling distance (GSD) and 2) surface water segmentation from Sentinel-2 imagery at about 10m GSD. Herein, we explore strategies to improve the zero-shot applicability of the general-purpose SAM by combining it with other pre-trained FMs for detection and classification, and we evaluate the potential performance gains achievable with minimal computational overhead through few-shot adapters on the datasets. Furthermore, we assess whether remote-sensing-specific training in RemoteCLIP and RemoteSAM leads to meaningful improvements over their general-purpose counterparts in large-scale geospatial segmentation. Overall, we conclude that domain-specific FMs can provide performance gains in certain settings, but are neither required nor always useful when compared with lightweight adaptation strategies and mixtures of different general models. This suggests that a more economical pathway might be to increase the remote sensing data used in the training of general FMs instead of training dedicated models specifically for ESS.

Submission Type. Analysis.

BoK Concepts. [AM14] Generalization and aggregation, [IP3] Image understanding, [GC3] Artificial intelligence (AI) in EO and GI

Keywords. Geospatial Artificial Intelligence; Foundation Models; Geospatial Big Data; Remote Sensing

1 Introduction

Nowadays, governmental and commercial satellites, aerial platforms, and street-level data acquisition systems contribute to an ever-growing repository of geospatial observations. However, the value of Earth Observation (EO) data is very much tied to applications that can generate revenue from these observations. On the research side, insights fuel data-driven Earth System Sciences; on the governmental side, they inform political decisions; and on the commercial side, they optimize production in areas such as agriculture, urban planning, and environmental monitoring. In this context, one traditionally defines information products as geospatial information collections *based* on measurements but *interpreted* towards semantics that arise in multiple use cases. Some examples are the Copernicus Land Monitoring Service (European Environment Agency, 2021) and the Global Urban Footprint and its successors (Esch et al., 2017). This data processing towards an information product usable for several applications can thereby involve significant processing requirements, especially if data from various sources is fused (Salcedo-Sanz et al., 2020).

One approach to ease the development of information products is the use of Foundation Models (FMs): neural networks pre-trained on diverse datasets and designed to generalize across a wide range of tasks and domains. FMs are intended to solve complex tasks such as question answering, image segmentation, or image labeling. For all of these tasks, a certain context is required: segmentation might depend on scale or tolerance with respect to color variations. It might be relevant to segment the most prominent object within a bounding box or at a point location. For question answering, the context of previous queries or underlying documents may be important to take into account. Therefore, FMs commonly expect more

input information than raw observations alone, provided in the form of additional knowledge that specifies what the model should focus on, commonly referred to as a prompt.

Various FMs have been proposed for the geospatial domain, and we do not aim to train or fine-tune a new spatial FM. Instead, we assess the applicability of existing FMs from the computer vision domain in geospatial contexts, alongside recent remote-sensing-specific FMs. In this context, we address the following question: *To what extent can advances in FMs trained outside the geospatial domain enable more efficient extraction of information products and thus value from geospatial datasets, compared to domain-specific alternatives?*

In this paper, we investigate general-purpose and geospatial FMs with geospatial imagery without fine-tuning or re-training. Specifically, we analyze how different prompts influence model performance, including prompts derived from other FMs or randomly sampled prompts. In addition, we explore the potential performance enhancement of few-shot adapters on the datasets. *All experiments are conducted under the constraint that no additional energy consumption is incurred through training or fine-tuning large-scale (spatial) FMs.*

Concretely, we use two datasets, orthophotos with building footprint annotations (Werner et al., 2023) and Sentinel-2 imagery with surface water labels (Li et al., 2021), to investigate the zero-shot segmentation FM SAM (Kirillov et al., 2023) in the remote sensing domain. In addition, we assess the impact of domain-specific pre-training with RemoteSAM (Yao et al., 2025), an FM specifically for remote sensing imagery. We evaluate the performance of:

- SAM prompted with ground-truth annotations (points or bounding boxes)
- SAM prompted with prediction targets from a pre-trained object detection model (Grounding DINO)
- SAM's automated prompting mechanism followed by classification of the segments using Vision Language Models (CLIP and RemoteCLIP)
- RemoteSAM with direct text prompting

2 Related Work

This paper aims to investigate the zero- and few-shot adaptability and performance of general-purpose and geospatial FMs in geospatial applications across datasets with differing spatial resolutions and sensing modalities. Towards this, we identify important general-purpose FMs and large models proposed as geospatial FMs.

2.1 General-Purpose Foundation Models

FMs are large models pre-trained in a task-agnostic manner to serve as general solutions adaptable to a wide range of downstream applications (Mai et al., 2023; Bommasani et al., 2022; Hong et al., 2024; Li et al., 2023b). In recent years, such models have been developed for many data domains, including language, graphs, videos, and images (Zhou et al., 2024). They typically foster large neural networks trained on large-scale datasets to learn generic representations. For our use case of evaluating such models in geospatial applications, we analyze two categories of models, Vision FMs and Vision Language FMs, as well as their combinations.

Many works have demonstrated the revolutionary potential of FMs. Vision FMs focus on processing visual data and have been trained to support different tasks, including image segmentation and object detection. Notable examples include SAM (Kirillov et al., 2023) and DINOv2 (Oquab et al., 2024). In contrast, Vision Language FMs are multimodal and jointly process visual and textual information. They allow for more complex tasks like image generation and editing, image captioning, visual question answering, and open-vocabulary segmentation and detection. Examples include CM3leon (Yu et al., 2023), BLIP (Li et al., 2022), and CLIP (Radford et al., 2021).

In this paper, we focus on the three types of general-purpose FMs based on their functionality and input-output characteristics. This selection is motivated by the claim of general applicability, the ability to be combined into a unified processing pipeline, and their relevance. These selected models are the segmentation Vision FM *SAM* (Kirillov et al., 2023), the Vision Language FM for object detection *Grounding DINO* (Liu et al., 2025), and *CLIP* (Radford et al., 2021; Liu et al., 2024), a Vision Language FM for classification.

2.1.1 Segment Anything Model: A Vision Foundation Model for Segmentation

The Segment Anything Model (SAM) by Meta (Kirillov et al., 2023) is a deep learning model that can be applied to two segmentation problems: a prompted approach or an automatic segmentation. Architecturally, SAM is an autoencoder originally featuring an encoder and a decoder (Goodfellow et al., 2016). To solve both previously defined segmentation problems, SAM is extended with a second encoder, as shown in Figure 1.

Images are encoded with a pre-trained Vision Transformer (ViT), while the second prompt encoder processes additional sparse and non-sparse inputs from the user or another algorithm. Possible inputs for the prompt encoder are points, boxes, masks, and text, though the weights of the latter are not public. The decoder is an attention-based transformer architecture (Vaswani et al., 2017). The model was pre-trained on the SA-1B image dataset and is available in three sizes, differing in the number of parameters of the image encoder: ViT-base (91M parameters), ViT-large (308M), and ViT-huge (636M).

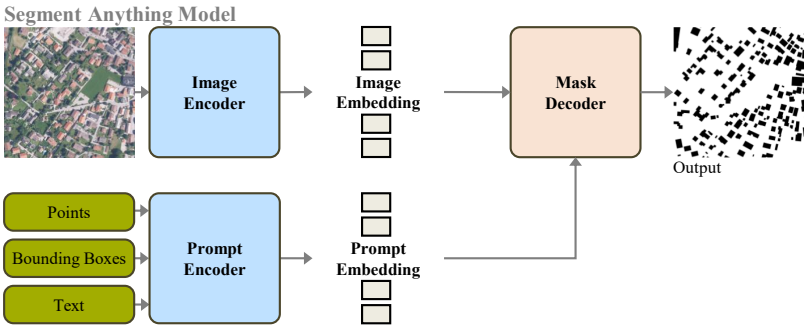

Figure 1. The general SAM architecture as described in (Kirillov et al., 2023).

2.1.2 Grounding DINO: A Vision Language Foundation Model for Object Detection

DINO is a ViT model that features a student-teacher architecture (Caron et al., 2021). It uses self-distillation without labels, training two similar networks with different parameters jointly: a student and a teacher. The learning happens in a self-supervised manner, where the teacher network ensures generalization. Grounding DINO uses this student-teacher architecture to work as an open-set object detection model for images (Liu et al., 2025). Compared to closed-set detectors, language is used as a generalization level for the object detection task: A language model first processes the user query to produce the query to use with the general DINO model (Zhang et al., 2022; Caron et al., 2021). For this, Grounding DINO features a transformer-like architecture for handling both language and image data (Liu et al., 2025). This results in a model capable of producing pairs of bounding boxes and noun phrase labels that fit the textual description given in the prompt. The model comes with different backbones, and in this paper, we focus on the Swin-T (172M parameters) and Swin-B (341M) backbones.

2.1.3 Contrastive Language-Image Pre-Training: A Vision Language Model for Classification

Contrastive Language-Image Pre-Training (CLIP) is a method of training a deep learning model on a large image database with corresponding captions only. Exact image labels are not required for this approach; instead, the model is capable of learning from similar pairs of data (Radford et al., 2021). This is possible with the help of two encoders: one for the textual descriptions and one for the images, to learn a common embedding of text and image. CLIP can be used for two types of tasks – image classification and text generation. At test time, only the image encoding is used to get embeddings that allow to reference back to the textual descriptions of the input image (Radford et al., 2021). The model is particularly tested for object detection within a zero-shot setting and exhibits better performance compared to existing ImageNet-based classification approaches.

2.2 Large Models for Geospatial Applications

Recent literature emphasizes that so-called general-purpose FMs are not well suited for geospatial data (Lacoste et al., 2021, 2024; Radford et al., 2021). Thus, two approaches to attain FMs for this data type have been proposed: (1) Add new pre- and post-processing steps to the existing universal FMs, or (2) train new 'foundation'-like models for the geospatial domain. Thereby, foundational approaches can significantly improve existing methods for geospatial artificial intelligence models in three areas (Mai et al., 2022):

- Make the model task agnostic (basic idea of FMs).
- Improve on sensor agnosticism by allowing the processing of different sensors' data and thus reducing the spatiotemporal sparsity due to the limited number of satellites.
- Enable the spatiotemporal awareness of the model, meaning allow for the additional spatial and temporal information to be fed and processed.

The authors of SAM claim that their model excels in single-shot learning tasks in various domains and often outperforms fully supervised approaches in segmentation. They thereby emphasize its ability to generalize the learned concepts to different domains without any retraining (Kirillov et al., 2023). However, Sultan et al. (2024) noted SAM's poor ability to generalize to mobility infrastructure in aerial imagery. Initial attempts at universal large pre-trained models with geospatial data have already been reported: Osco et al. (2023) investigate SAM with geospatial data and show its promising results in this field, but propose improving its ability through one-shot training. The authors of CLIP state that their model performs poorly on satellite imagery (Radford et al., 2021). Therefore, Liu et al. (2024) adopted approach (2) and trained RemoteCLIP on remote sensing images, thereby adapting CLIP to the geospatial domain. In a similar fashion, RemoteSAM trained a SAM-like approach for the remote sensing domain (Yao et al., 2025).

Additionally, several tailored models were proposed for remote sensing imagery, either in RGB only or for multi-

band: 'Prithvi' (Jakubik et al., 2023) and SatMAE (Cong et al., 2022) use reconstruction-based training paradigms and perform especially well in multi-spectral applications. Others, such as CROMA (Fuller et al., 2023) and AlphaEarth (Brown et al., 2025), are even more tailored to the geospatial domain, e.g., by using additional metadata during training. Foundation-like approaches only for RGB images in the remote sensing field are presented by Cha et al. (2023) and Sun et al. (2023). However, all these models do not serve as a universal FM in the traditional sense, since they are tailored to the geospatial domain only.

Consequently, the diversity of geospatial data types and tasks has motivated the development of dedicated benchmarks. Recent benchmarks for geospatial FMs have been introduced to cover diverse applications (Mai et al., 2023; Lacoste et al., 2024; Marsocci et al., 2025; Wang et al., 2025), with Mai et al. (2023) proposing seven tasks, including geospatial semantics, health and urban geography, and remote sensing.

3 Datasets

Geospatial segmentation can be divided into two categories: the segmentation of relatively small self-contained objects like buildings or cars, and the segmentation of large, heterogeneous objects, e.g., street networks or landscape features. To capture this diversity, we select two datasets that vary in overall appearance and segmentation label structure. While the Bavaria Building Dataset comprises self-contained segments of houses, the Surface Water Dataset involves diverse water bodies: small lakes and ponds, long rivers, and thin streams.

To evaluate the performance of FMs on geospatial tasks, unbiased by data pre-processing, we select only the data dimensions that closely match the natural images in the training sets of FMs. Specifically, we select only the subset of bands corresponding to RGB, since they are widely available across satellite and airborne datasets and can be directly processed by FMs trained on RGB images.

3.1 Bavaria Building Dataset

The Bavaria Building Dataset (BBD) by Werner et al. (2023) consists of RGB orthophotos at 40cm ground sampling distances (GSD) obtained by the Bavarian government. The corresponding building footprint data was obtained from OpenStreetMap (OSM). We process the data from Upper Bavaria in patches of 1024×1024 pixels to align with the input requirements of SAM, retaining only images with segmented objects. Thus, the dataset reduces to 4,650 patches out of the original 15,015. As buildings often belong to contiguous clusters formed by densely packed neighborhoods, this dataset exhibits strong spatial autocorrelation. Additionally, the complexity of the dataset is relatively low, indicated by a low bounding box-to-label ratio.

3.2 Surface Water Dataset

The Surface Water Dataset by Li et al. (2021) consists of Sentinel-2 images at 10m GSD with surface water masks obtained from OSM. For this case study, only the RGB bands and data from Upper Bavaria, Germany were selected, resulting in a 18177×19283 pixel image. The image was tiled into 1024×1024 patches and normalized with a cumulative count cut at 2% to 98%. If a patch was not completely within the base image's boundary, the image was padded with zeros. In total, this dataset consists of 312 images for segmentation. The Surface Water Dataset is characterized by thin, dispersed, and less visible formations. A single river extends along the entire image; however, the resulting average label density equals that of several compact and small buildings from BBD. Because rivers and streams are naturally narrow and the Surface Water Dataset has a significantly higher GSD, these water bodies occupy fewer pixels in width. Further, the bounding box-to-label ratio is comparably high for the dataset, indicating that the bounding boxes are significantly larger than the compact but irregularly shaped objects contained in them.

4 Methods

From a modeling perspective, we mainly investigate the ideas mentioned by Osco et al. (2023): We integrate different FMs and perform segmentation without fine-tuning. In the following, four different approaches are implemented to map two distinct geographical objects in satellite images – buildings and water bodies. An overview of the methods is provided in Figure 2.

4.1 Ground-Truth Prompted SAM

SAM's main appeal is its universal ability to accurately segment any object given a bounding box or a point on the object in question. For our experiments, we use SAM from Huggingface's Transformer library (Wolf et al., 2020). To automate the generation of bounding boxes and point inputs, the ground-truth masks are exploited:

- **Bounding Box**: Each contiguous masked object in the ground-truth is enveloped with an axis-aligned minimum bounding box.
- **Representative Point**: For each building, a single point is selected to identify the object. This is the center-of-mass of the shape if it lies within it, and a cheaply computed point within the shape otherwise.
- **Multiple Randomly Sampled Points**: For each building, multiple uniformly sampled points are selected.
- **Foreground/Background Points**: The same number of points are uniformly sampled and correctly labeled

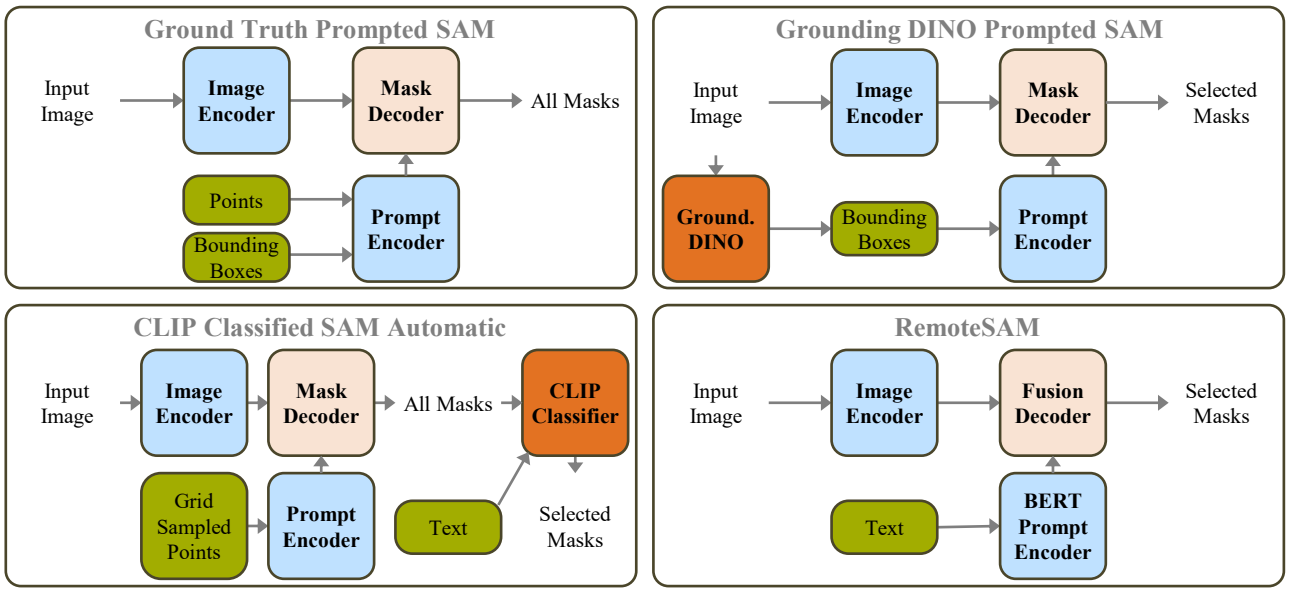

Figure 2. Schemata of the four used methods: Ground-Truth Prompted SAM, Grounding DINO Prompted SAM, CLIP Classified SAM Automatic, and the domain-specific RemoteSAM

from the foreground (buildings or water bodies) and the background.

These evaluations enable a performance analysis of SAM's pure image encoder-decoder architecture and serve as a baseline for the other approaches.

4.2 Grounding DINO Prompted SAM

SAM supports text-based segmentation; however, this prompting mechanism has yet to be released. To circumvent this, we prepend an object detection model to SAM, Grounding DINO (Liu et al., 2025). This pipeline, termed Grounded SAM (Ren et al., 2024), has been applied by many in various domains. Integrating Grounding DINO automates the generation of bounding boxes that prompt SAM based on text inputs. By combining these models, we establish a powerful pipeline that leverages Grounding DINO's ability for object recognition and SAM's capability to convert these boxes into segmentation masks. The quality of results generated by Grounding DINO depends on the classes, with better outcomes for common classes. Therefore, we aim to test the zero-shot generalizability of Grounding DINO for geospatial objects.

4.3 RemoteSAM

RemoteSAM proposes a segmentation-centric FM for EO (Yao et al., 2025). Architecturally, it adopts a Swin-B visual encoder and BERT as text encoder instead of Large Language Models (LLMs). A dedicated cross-modal fusion decoder jointly reasons over visual and textual features to perform referring expression segmentation. The pixel-level masks serve as a unifying task representation for multi-task support through a subsequent mask-to-task conversion for detection, grounding, classification, counting, and captioning. To enable this paradigm, the authors train the model on a newly constructed large-scale image-text-mask dataset RemoteSAM-270K that expands the underrepresented semantic categories of remote sensing imagery. With these adaptations, the authors report state-of-the-art performance across multiple remote sensing benchmarks despite significantly fewer parameters compared to LLM-based remote sensing FMs.

Unlike SAM, RemoteSAM expects images of size 896×896. Since our images are cropped to 1024×1024, they are resized using bilinear interpolation. However, some objects in the Surface Water Dataset are very thin and may be diluted during resizing. To mitigate this, we explored two alternative strategies: The first strategy splits each image into four overlapping tiles, runs RemoteSAM on each tile independently, and then stitches the results by averaging the predictions. The second variant also uses tiling, but considers an object detected if it appears in any of the tiles. We report results obtained with the latter, which performed best.

4.4 CLIP Classified SAM Automatic

In the final experiment, we test SAM Automatic, where the only difference from point-prompted SAM is how these points are selected. Instead of a manual selection, SAM automatically generates several points in a grid-based structure. The user specifies the number of points per side, and SAM Automatic segments the image for each grid point in a point-prompted manner.

Since SAM Automatic cannot classify the predicted masks, an additional classification model is necessary.

Table 1. Evaluation of SAM with varied model sizes on BBD using different prompts generated from ground-truth masks. The highest performance across the model sizes is underlined, while the optimal prompt results within ViT-large are in bold.

	ViT-base			ViT-large			ViT-huge		
	BB	RP	5SP	BB	RP	5SP	BB	RP	5SP
F1	0.833	0.743	0.767	**0.842**	0.752	0.791	0.840	0.750	0.782
IoU	0.713	0.591	0.622	**0.726**	0.602	0.654	0.724	0.600	0.642
Precision	0.816	0.720	0.669	**0.811**	0.750	0.703	0.809	0.754	0.684
Recall	0.850	0.767	0.899	0.874	0.753	**0.903**	0.873	0.746	0.912

For this purpose, we append another FM, CLIP, to decide whether each image segment corresponds to the desired geospatial object. As the primary objective of our benchmarking is to avoid fine-tuning large models for geospatial data, we decided to also evaluate RemoteCLIP (Liu et al., 2024), an FM fine-tuned on geospatial data for remote sensing applications. This is motivated by prior evaluations of CLIP (Radford et al., 2021), which confirm its limitations on satellite imagery. We also evaluate CLIP on the selected datasets and compare it with RemoteCLIP to study the potential performance gains of fine-tuning on geospatial data. In addition, we investigate the impact of adding few-shot adapters to CLIP.

4.5 Data and Software Availability

Both datasets are open source (Werner et al., 2023; Li et al., 2021). The code and preprocessed datasets are available at the following links:

- Code repository: https://github.com/tum-bgd/2026-AGILE-GeoFM
- BBD: https://figshare.com/s/83c2cead0d8fcc65dc6c
- Surface Water Dataset: https://figshare.com/s/a359b71447c18d02b9db

5 Evaluation and Results

We evaluate all four proposed methods based on the end segmentation masks instead of the intermediate outputs, which allows us to compare all methodologies directly. To this end, the F1 score combines precision (how accurate positive predictions are) and recall (how many true positives are correctly identified), while Intersection over Union (IoU) measures the overlap between predicted and ground-truth masks.

5.1 Evaluations on the Bavaria Building Dataset

Table 1 compares the different-sized SAM models on BBD using prompts derived from the ground-truth masks. Among the three model sizes and across all prompt types, bounding box (BB), representative point (RP), and five randomly sampled points (5SP) per building, ViT-large consistently performs the best. Indeed, ViT-huge is known to over-segment images and produce excessively fragmented segments that do not generalize well to these compact objects. Based on these results, ViT-large seems the most effective and is thus selected for the remaining evaluations. Here, bounding box prompts achieve the highest F1 and IoU scores, meaning they produce the most accurate segmentation masks. While point-based prompts demonstrate strong performance, their scores are consistently slightly lower.

Figure 3 depicts an example prediction from ground-truth-generated prompts. Segmentation with bounding box prompts clearly outperforms point-based segmentation. When using multiple points, the segmentation outcomes compared to the single point results are improved, e.g., in the upper left corner, the single point prediction does not fully capture the more complex building. Nevertheless, multi-point prompt predictions produce false positives in the lower right corner with scattered and pixelated predictions. This highlights a trade-off between improved spatial coverage and increased noise with multiple points.

Table 2 examines the trade-off between the number of randomly sampled points per building and segmentation performance. Both the F1 and IoU scores improve as the number of sampled points increases, peaking at 4SP. The lowest performance occurs at 1SP and 20SP, indicating that using too few or too many points adversely affects SAM's overall segmentation accuracy.

Table 2. Evaluation of SAM on BBD using different numbers of randomly sampled points from the ground-truth masks.

	1SP	2SP	4SP	5SP	10SP	20SP
F1	0.722	0.781	**0.793**	0.791	0.768	0.720
IoU	0.565	0.640	**0.657**	0.654	0.623	0.563
Precision	0.731	**0.734**	0.713	0.703	0.668	0.611
Recall	0.712	0.833	0.893	**0.903**	**0.903**	0.877

In the following cases, point-based prompts may cause SAM to segment an object from the background:

- **Dataset Inconsistencies**: Discrepancies between orthophotos and segmentation masks, e.g., missing buildings that are not updated in the OSM data.

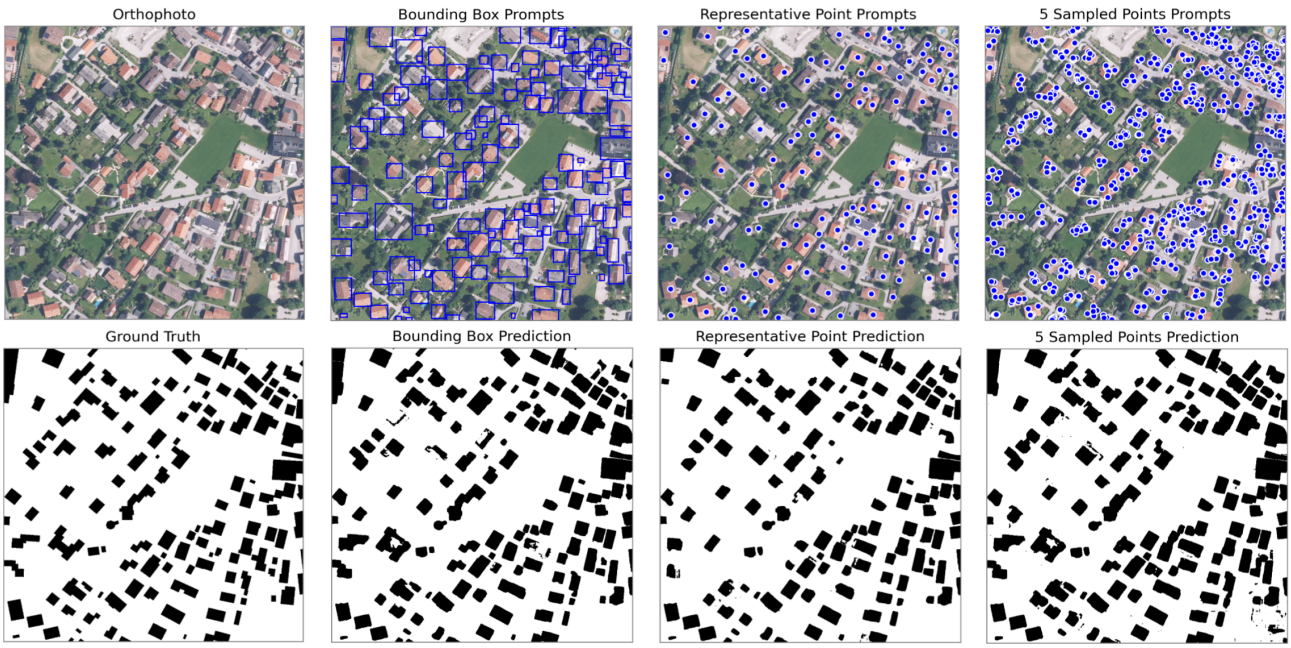

Figure 3. Example predictions on BBD of the ground-truth prompted SAM with different prompts.

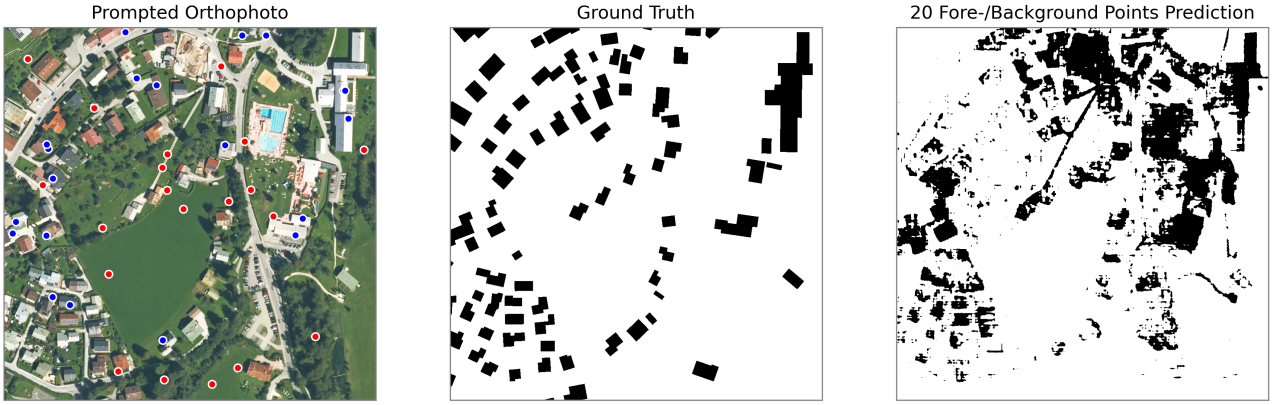

Figure 4. Example prediction from ground-truth prompted SAM with 20 randomly sampled background (red dots) and foreground points (blue dots), respectively.

- **Geospatial Inaccuracies**: The position and shape of polygons in the OSM data are imprecise, which means that some randomly sampled points do not align with the building.
- **Obstructions and Overgrowth**: Overgrown trees or other objects can obscure buildings, which results in points sampled from the greenery, not the building.

To avoid over-penalizing the point-based results, masks that are five times larger than their ground-truth counterparts were excluded. This adjustment increases the F1 scores by 0.3 to the values in Table 2. These erroneous results emphasize the need for data quality management for datasets. By understanding these problems, the segmentation results can be applied as a diagnostic tool to help build a more accurate geospatial data ecosystem that supports strong data quality control practices.

Another SAM prompting strategy uses foreground and background points; however, this yields the worst segmentation performance. Figure 4 shows how SAM fails to generalize to non-prompted buildings and to reliably distinguish foreground from background. While this approach can be successful with simple images with clear background separation, it becomes noisy and ineffective in cluttered aerial imagery. Indeed, aerial views encompass diverse object classes, which limits SAM's ability to generalize when insufficient prompts represent the variety of background structures to be excluded.

For the experiments with Grounded DINO, we use the Swin-T backbone since it outperformed Swin-B. We note that in some cases, Grounding DINO produced bounding boxes that span the entire image; however, based on our domain knowledge from the data, no building is of this size. Most of these undesired outputs occurred in images

Table 3. Evaluation on BBD of Grounded SAM (GS) and the directly text-prompted RemoteSAM (RS). The best performance across GS and RS is underlined, while the optimal prompt results within each model are in bold.

	'building'		'buildings'		'single buildings'		'building and structure'		'building . buildings .'	
	GS	RS	GS	RS	GS	RS	GS	RS	GS	RS
F1	**0.597**	0.533	0.569	0.516	0.520	0.400	0.485	0.498	0.523	**0.537**
IoU	**0.426**	0.364	0.398	0.348	0.351	0.250	0.320	0.332	0.354	**0.367**
Precision	**0.546**	0.656	0.489	**0.659**	0.498	0.306	0.491	0.420	0.511	0.652
Recall	0.659	0.449	**0.680**	0.424	0.544	0.576	0.480	**0.613**	0.536	0.456

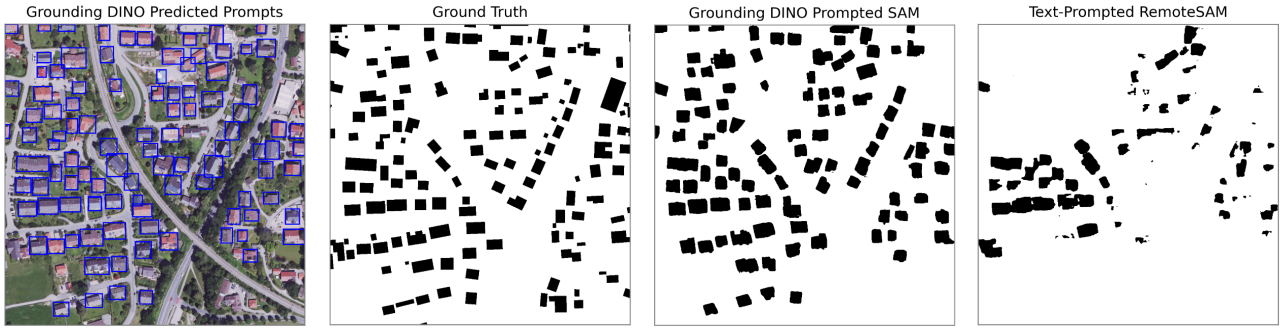

Figure 5. Example bounding box predictions from Grounding DINO and subsequent Grounded SAM and RemoteSAM results, prompted with the term 'building' on BBD.

with only a few structures. Filtering out these bounding boxes increased the F1 and IoU scores from 0.139 and 0.075, respectively, to the values reported in Table 3.

We observed that Grounding DINO's performance was affected by subtle semantic nuances in the prompts. Among the tested expressions, the term 'building' performed best. This highlights the importance of prompt wording for identifying contextually relevant information. Table 3 shows a subset of the evaluated prompts that achieved the highest performance. Grounding DINO supports using phrases as prompts or including a list of text prompts separated with a dot. However, including synonyms or variations in the text prompts, e.g., 'building and structures' or 'building . buildings .' among others, did not lead to performance improvements in this case.

In the same table, we additionally report the results of the third experiment configuration obtained with RemoteSAM using identical text prompts. For completeness, we note that RemoteSAM accepts multiple class names in the form of a list; thus, the prompt 'building . buildings .' is now passed as ['building', 'buildings']. In contrast to Grounded SAM, RemoteSAM does not require a two-step pipeline and can be directly prompted with text. For RemoteSAM, the best-performing prompt is 'building . buildings .'. Despite this, the overall performance of RemoteSAM is mostly lower than that of Grounded SAM, with the exception of the prompts 'building and structure' and 'building . buildings .', which achieve marginally better F1 and IoU scores. These inferior results are unexpected, and the exact underlying cause remains unclear, given that RemoteSAM was specifically trained on geospatial datasets in which the building class is highly prevalent.

To visually illustrate the building detection ability of Grounding DINO, Figure 5 presents a positive example in which Grounded SAM successfully identifies most buildings within the images. Nevertheless, Grounding DINO struggles with smaller buildings and structures with white rooftops. In contrast, Figure 5 also reflects the weaker performance of RemoteSAM, where a considerable number of buildings remain undetected, consistent with the quantitative results reported in Table 3.

For the final evaluation, SAM Automatic was deployed to generate segmentation masks via point-grid prompting. SAM Automatic's mask quality prediction is used to filter out masks that do not satisfy IoU and stability score thresholds. The retained masks are then overlaid on the image to isolate candidate objects. The classification of these as building or background is done with a zero-shot CLIP and RemoteCLIP. The best-performing queries were: 'satellite image of buildings' and 'satellite image of background', and based on the image–text similarity scores, the masked object is classified accordingly.

Figure 6 and Table 4 confirm the promising results of this approach. As expected, RemoteCLIP outperforms the non-domain-specific CLIP at a fraction of the number of parameters. Indeed, compared to the previous pipeline with Grounding DINO, the regular CLIP performs worse. A possible explanation is that CLIP performs classification using only the cropped area from the generated masks, which strips the image of its overall context, potentially affecting the understanding of the object's environment. In contrast, RemoteCLIP is specifically trained on remote sensing images, enabling it to classify objects without the need for the surrounding context of the object.

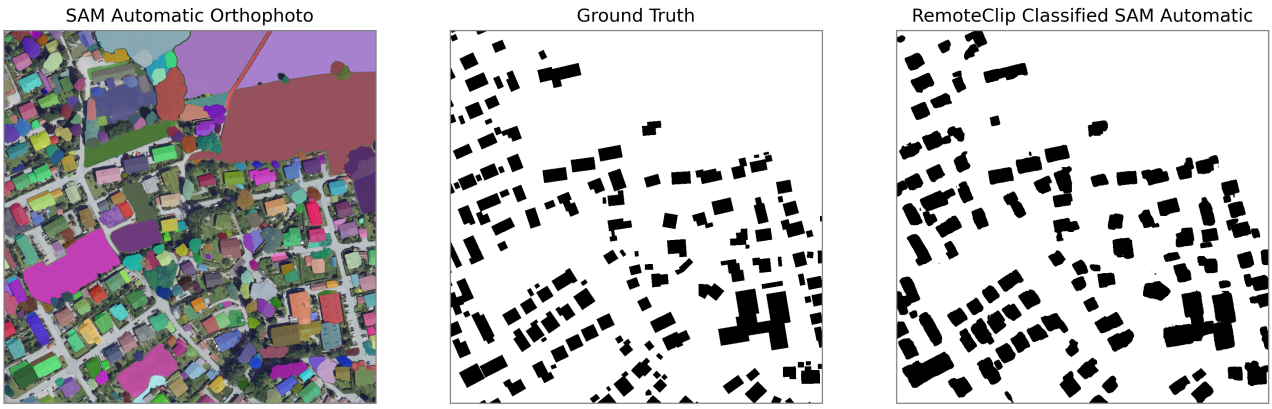

Figure 6. Example prediction on BBD of SAM Automatic classified with RemoteCLIP.

Table 4. Evaluation of SAM Automatic on BBD with the generated masks classified with CLIP or RemoteCLIP as either 'buildings' or 'background' at a threshold of 0.7.

	CLIP ViT-bigG-14			**RemoteCLIP ViT-B-32**			**RemoteCLIP ViT-L-14**		
Grid Points	32x32	64x64	128x128	32x32	64x64	128x128	32x32	64x64	128x128
F1	0.444	0.467	0.485	0.472	0.480	0.482	0.647	0.661	**0.668**
IoU	0.285	0.304	0.320	0.309	0.316	0.317	0.479	0.494	**0.502**
Precision	0.532	0.505	0.504	0.503	0.484	0.472	**0.699**	0.681	0.674
Recall	0.381	0.433	0.467	0.445	0.476	0.492	0.603	0.642	**0.662**

SAM Automatic produces masks at pre-defined point coordinates that have been generated on a grid basis, which means more points lead to a more densely prompted image. We test different numbers of grid points, namely 32x32, 64x64, and 128x128, and observe that a more dense grid results in better performance across all three CLIP models. SAM demonstrates once again its zero-shot segmentation capabilities, and with the assistance of RemoteCLIP ViT-L-14 becomes a potential instance segmentation module for remote sensing.

To minimize extensive training efforts when applying FMs to the geospatial domain, we explore a few-shot learning adapter for CLIP: Tip-Adapter (Zhang et al., 2021). Specifically, we train a small cache model using only eight examples of buildings and backgrounds cropped from the dataset. A classification is performed by combining the adapted features with pre-trained CLIP's features. Applying Sam Automatic to a grid of 128x128 points and using CLIP ViT-bigG-14 and its trained adapter achieves a 0.617 F1 score and 0.446 IoU. Remarkably, these results surpass those of RemoteCLIP ViT-B-32 and approach the performance of RemoteClip ViT-L-14, all achieved with just eight examples and only seconds of training.

5.2 Evaluations on the Surface Water Dataset

The same experiments were evaluated on water body segmentation, with SAM results using ground-truth masks summarized in Table 5. As with the post-processing of BBD predictions, a competitive performance was achieved by filtering out masks that exceed five times the size of the corresponding ground-truth. Indeed, the authors of the Water Surface Dataset report an F1 score greater than 0.8 using a model specifically designed and trained for this dataset and leveraging additional spectral bands beyond RGB. Out of all prompts, two randomly sampled points per masked object achieve the best balance between precision and recall, resulting in an F1 score of 0.797. However, without filtering, SAM struggled to correctly identify water areas, as documented on the left side of Table 5. Notably, when compared to their bounding box counterparts, unfiltered point-based prompts performed significantly worse.

Unlike with BBD, SAM's difficulty in segmenting water bodies can first be attributed to the high GSD of these satellite images. Figure 7 illustrates an example prediction and highlights SAM's lack of robustness on lower-resolution water bodies. Given the inherent differences between buildings and water bodies, these results are to be expected. Most buildings have rectangular-like shapes, and bounding boxes capture most of their structure. On the other hand, the Surface Water Dataset contains a few large lakes and predominantly fine-grained objects, namely small water surfaces and narrow, elongated rivers, streams, and creeks that span much of the image but occupy only a few pixels. Consequently, these irregular shapes produce large bounding boxes that only envelop a very thin-shaped object. However, SAM cannot precisely interpret the shape of the object based on the bounding box. Similarly, point-based prompting for these objects

Table 5. Evaluation of SAM on the Surface Water Dataset using different prompts generated from ground-truth masks before and after filtering out very large predictions.

	Before Filtering					After Filtering				
	BB	RP	2SP	5SP	10SP	BB	RP	2SP	5SP	10SP
F1	**0.432**	0.069	0.079	0.107	0.106	0.777	0.775	**0.797**	0.757	0.738
IoU	**0.276**	0.036	0.041	0.057	0.056	0.636	0.633	**0.663**	0.609	0.584
Precision	**0.285**	0.036	0.041	0.057	0.056	0.736	0.823	**0.826**	0.732	0.761
Recall	0.894	0.919	0.955	0.971	**0.979**	**0.823**	0.733	0.771	0.783	0.713

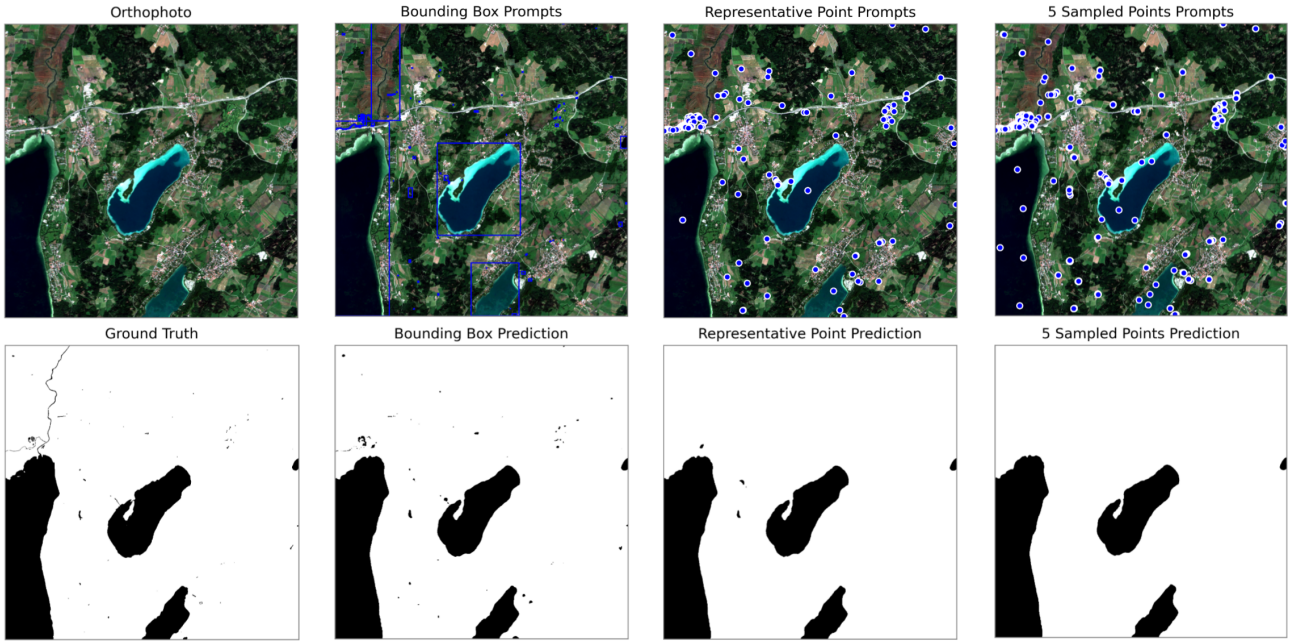

Figure 7. Example predictions on the Surface Water Dataset of the ground-truth prompted SAM with different prompts.

frequently confuses the surrounding land as part of the target. This limitation prevents SAM from only segmenting the streams and instead mistakes them as part of a bigger, unrelated object in the background. This leads to high recall but low precision, which emphasizes the importance of considering topological properties when designing prompts for segmentation tasks of geospatial data and the need for models capable of granular object-level understanding (Li et al., 2023a).

Even worse than the BBD results, most bounding box predictions of Grounding DINO span the whole image. Once again, filtering these improves performance; however, the results remain very low. Grounding DINO's inability to extract water bodies from Sentinel-2 images is confirmed with the results from Table 6 and can be attributed to several factors. First, the surface water class is relatively rare and likely underrepresented by the dataset during DINO's training. Second, the high GSD of Sentinel-2 images means lower spatial resolution. Therefore, the details of the object of interest are less visible compared to BBD. Indeed, most objects are fine structures, either thin streams or small water surfaces, making them more difficult to detect. Lastly, identifying water surfaces in satellite images is a challenging task due to their high visual diversity, particularly color variations that range from blue and green to darker tones depending on water depth and other environmental conditions.

On the Surface Water dataset, RemoteSAM performs slightly better than Grounded SAM for certain prompts. Interestingly, the best-performing prompt is 'water body', which was unexpected because the categories lake, river, and stream present during RemoteSAM's training failed to produce meaningful results. Nevertheless, the highest F1 score achieved by RemoteSAM is 0.265, which is substantially lower than the performance obtained with ground-truth–prompted SAM.

Figure 8 illustrates an example prediction of Grounded SAM and RemoteSAM using the prompt 'water body'. Both models detect only one of the two large lakes in the image. The results further reveal that RemoteSAM is sensitive to color variations. While SAM produces a coherent segmentation mask for the detected lake, RemoteSAM generates a highly fragmented mask. This behavior means that RemoteSAM struggles to capture the full spatial extent of objects with different color ranges.

Table 6. Evaluation on the Surface Water Dataset of Grounded SAM (GS) and the directly text-prompted RemoteSAM (RS).

	'water surface'		'water'		'lake. river. stream.'		'water body'		'blue'		'blue surface'	
	GS	RS	GS	RS	GS	RS	GS	RS	GS	RS	GS	RS
F1	0.042	0.051	0.031	0.022	0.052	0.007	0.052	**0.265**	0.096	0.000	0.099	0.186
IoU	0.021	0.026	0.016	0.011	0.027	0.004	0.027	**0.153**	0.050	0.000	0.052	0.102
Precision	0.025	0.027	0.019	0.014	0.031	0.014	0.031	**0.497**	0.056	0.005	0.057	0.266
Recall	0.136	0.341	0.093	0.055	0.155	0.005	0.175	0.129	0.181	0.000	**0.376**	0.142

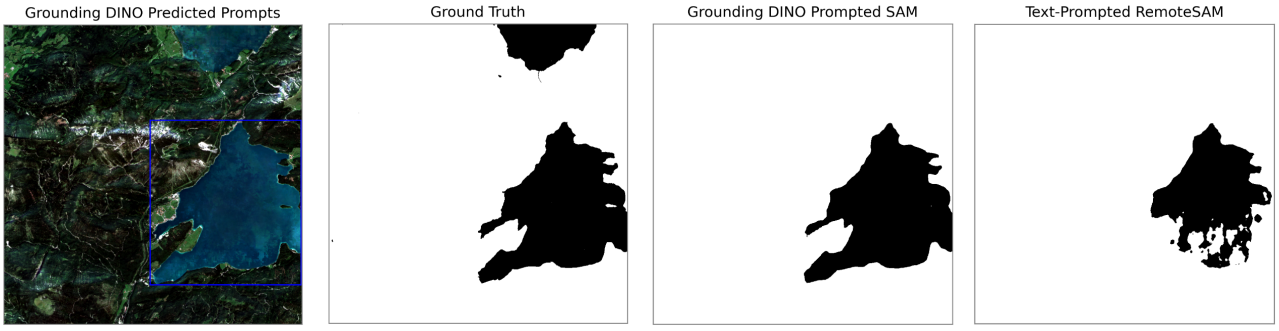

Figure 8. Example bounding box predictions from Grounding DINO and subsequent Grounded SAM and RemoteSAM results, prompted with the term 'water body' on the Surface Water Dataset.

Table 7. Evaluation of SAM Automatic on the Surface Water Dataset with the generated masks classified with CLIP or RemoteCLIP as either 'a surface water' or 'background'.

	CLIP ViT-bigG-14			**RemoteCLIP ViT-B-32**			**RemoteCLIP ViT-L-14**		
Grid Points	32x32	64x64	128x128	32x32	64x64	128x128	32x32	64x64	128x128
F1	0.325	0.328	0.325	**0.614**	0.603	0.602	0.433	0.380	0.377
IoU	0.194	0.196	0.194	**0.443**	0.432	0.430	0.277	0.234	0.233
Precision	0.307	0.298	0.309	**0.807**	0.739	0.738	0.374	0.297	0.288
Recall	0.346	0.365	0.344	0.496	0.510	0.508	0.514	0.527	**0.547**

Similar to the results of the first dataset, CLIP once again performed worse than RemoteClip. In contrast to the previous approach, where Grounding DINO showed its unfamiliarity with the water bodies class, CLIP demonstrated better recognition of water features. As reported in Table 7, RemoteCLIP performed best with ViT-B-32. Among all text prompts we tested, 'a surface water' yielded the highest F1 scores compared to other queries. Unlike BBD, increasing the number of grid points did not improve the F1 score. Nonetheless, the combination of the two FMs still performed suboptimally on zero-shot surface water segmentation and classification when faced with small or sparse water features as shown in Figure 9. In certain cases, SAM Automatic's raster-based approach misses these targets on top of its weaknesses in granular segmentation. In other cases, RemoteCLIP fails to correctly classify the low-resolution pixelated objects.

Finally, Tip-Adapter trained on only eight examples further deteriorates the already poor performance of CLIP ViT-bigG-14. Specifically, when combined with SAM on a 128x128 grid, the pipeline achieves an F1 score of just 0.235 and an IoU of 0.133.

5.3 Discussion

All tested approaches produced promising building segmentation results but struggled with water bodies in Sentinel-2 imagery. This is largely because they are not trained on datasets with varying GSDs and sensor types, and do not support high-dimensional remote sensing data, such as multispectral and multi-sensor inputs.

SAM performed best with bounding box prompts, while point-based prompting introduced a trade-off between coverage and noise. While SAM achieved strong results on building footprint detection, performance degraded notably on Sentinel-2 surface water imagery due to lower spatial resolution and appearance variability. This confirms the persistent challenge with datasets with finer-grained targets, which should be a key objective for future research (Li et al., 2023a).

Considering that Grounding DINO was trained on limited data, it provided reasonable bounding boxes for buildings. In contrast, it failed on surface water detection, which can be partly attributed to a lack of training examples for this class. Surprisingly, RemoteSAM showed low quantitative

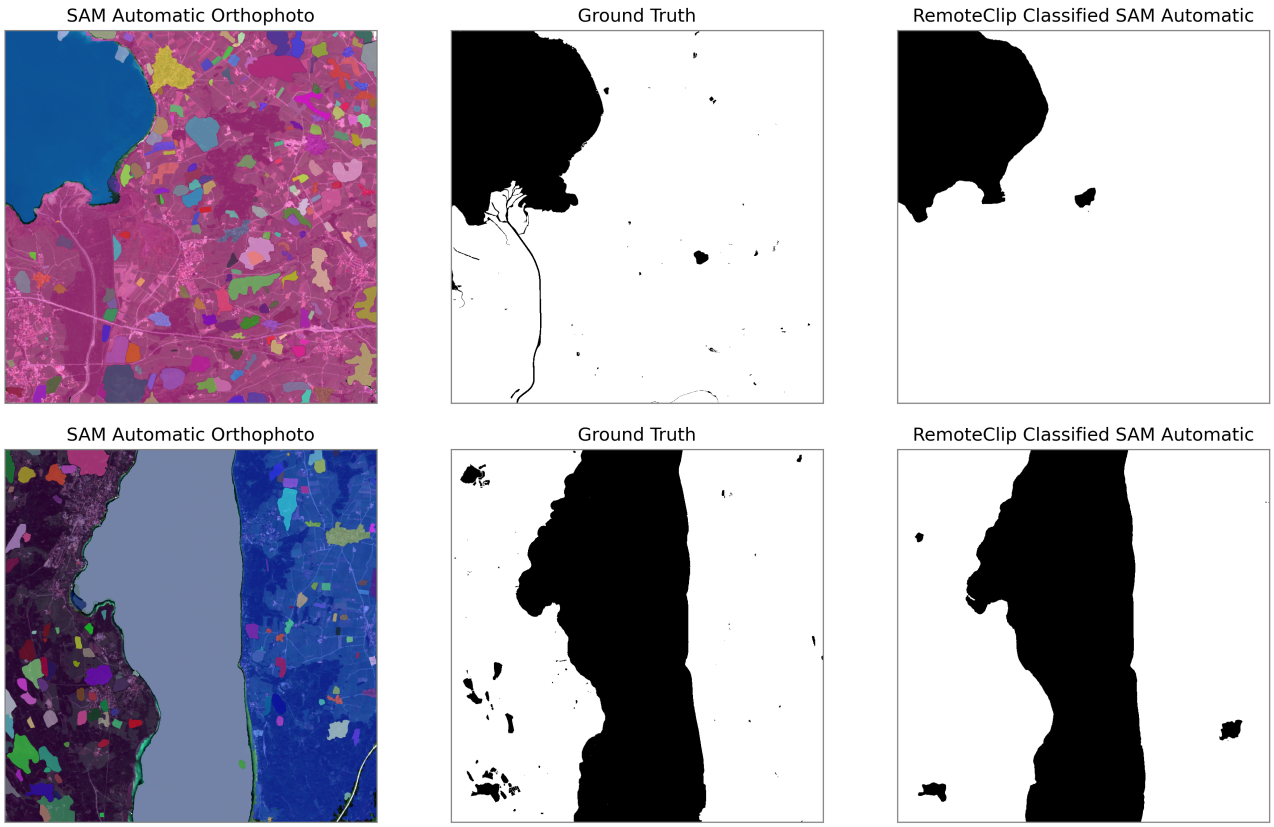

Figure 9. Example predictions on the Surface Water Dataset of SAM Automatic classified with RemoteCLIP.

performance and fragmented masks, despite being trained on geospatial datasets with abundant building and surface water annotations. These findings suggest that domain-specific pre-training alone does not guarantee generalization in zero-shot geospatial segmentation.

Combining CLIP with SAM Automatic creates an effective instance segmentation pipeline for aerial imagery. RemoteCLIP's domain knowledge of geospatial data compared to CLIP showed improvements with both datasets. However, this pipeline was still sensitive to prompt wording, illumination, and object scale. Indeed, SAM Automatic struggles with very small objects due to its grid-based segmentation approach and limited granularity. Finally, adding a few-shot adapter to CLIP improved building detection but did not fully resolve issues with surface water segmentation.

6 Conclusion

In this study, we evaluate three general-purpose FMs, SAM, Grounding DINO, and CLIP, on two geospatial datasets for building and surface water segmentation. We compare their performance with the domain-specific alternatives, RemoteSAM and RemoteCLIP. Our work explores cooperative model pipelines by combining different FMs to evaluate their potential zero- and few-shot segmentation capabilities on geospatial imagery with varying spatial resolutions and appearance characteristics. Our results demonstrate that zero- and few-shot performance is strongly influenced by dataset characteristics, object scale, and prompt semantics. This highlights the limitations of current FMs with visually and semantically complex geospatial data. Indeed, current modeling focuses on conventional benchmarks but neglects their cross-domain capabilities, which limits their potential. Therefore, we reiterate the importance of dataset diversity, prompt design, and cooperative model pipelines for improving generalizability across domains, especially in more complex disciplines such as medical imaging and remote sensing. Instead of relying solely on domain-specific FMs, our findings suggest that careful combinations of general-purpose models and lightweight adaptation strategies to infuse domain knowledge could offer a practical path with geospatial data.

Disclosure Statement

The authors report no competing interests to declare.

Declaration of Generative AI in writing

The authors declare that they have not used Generative AI tools in the preparation of this manuscript.

Acknowledgements

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 507196470.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models, 2022.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., and Kohli, P.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, https://arxiv.org/abs/2507.22291, 2025.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A.: Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

Cha, K., Seo, J., and Lee, T.: A Billion-scale Foundation Model for Remote Sensing Images, 2023.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S.: SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, in: Advances in Neural Information Processing Systems, edited by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., vol. 35, pp. 197–211, Curran Associates, Inc, https://proceedings.neurips.cc/paper_files/paper/2022/file/01c561df365429f33fcd7a7faa44c985-Paper-Conference.pdf, 2022.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., and Strano, E.: Breaking new ground in mapping human settlements from space – The Global Urban Footprint, ISPRS Journal of Photogrammetry and Remote Sensing, 134, 30–42, https://doi.org/10.1016/j.isprsjprs.2017.10.012, 2017.

European Environment Agency: High Resolution Vegetation Phenology and Productivity: Normalized Difference Vegetation Index (raster 10m) - version 1 revision 1, Sep. 2021, https://doi.org/10.2909/5d5f72ce-80bc-4c90-80ed-f135596533e2, 2021.

Fuller, A., Millard, K., and Green, J.: CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders, in: Advances in Neural Information Processing Systems, edited by Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., vol. 36, pp. 5506–5538, Curran Associates, Inc, https://proceedings.neurips.cc/paper_files/paper/2023/file/11822e84689e631615199db3b75cd0e4-Paper-Conference.pdf, 2023.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, http://www.deeplearningbook.org, 2016.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Ghamisi, P., Yokoya, N., Li, H., Jia, X., Plaza, A., Gamba, P., Benediktsson, J. A., and Chanussot, J.: SpectralGPT: Spectral remote sensing foundation model, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 5227–5244, 2024.

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S. K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., and Ramachandran, R.: Foundation Models for Generalist Geospatial Artificial Intelligence, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.: Segment Anything, pp. 4015–4026, 2023.

Lacoste, A., Sherwin, E. D., Kerner, H., Alemohammad, H., Lütjens, B., Irvin, J., Dao, D., Chang, A., Gunturkun, M., Drouin, A., Rodriguez, P., and Vazquez, D.: Toward Foundation Models for Earth Monitoring: Proposal for a Climate Change Benchmark, 2021.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al.: Geo-bench: Toward foundation models for earth monitoring, Advances in Neural Information Processing Systems, 36, 2024.

Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., and Gao, J.: Semantic-SAM: Segment and Recognize Anything at Any Granularity, 2023a.

Li, H., Zech, J., Ludwig, C., Fendrich, S., Shapiro, A., Schultz, M., and Zipf, A.: Automatic mapping of national surface water with OpenStreetMap and Sentinel-2 MSI data using deep learning, International Journal of Applied Earth Observation and Geoinformation, 104, 102 571, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, pp. 12 888–12 900, PMLR, 2022.

Li, W., Lee, H., Wang, S., Hsu, C.-Y., and Arundel, S. T.: Assessment of a New GeoAI Foundation Model for Flood Inundation Mapping, in: Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI '23, p. 102–109, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3615886.3627747, 2023b.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., and Zhou, J.: Remoteclip: A vision language foundation model for remote sensing, IEEE Transactions on Geoscience and Remote Sensing, 62, 1–16, 2024.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection, in: European Conference on Computer Vision, pp. 38–55, Springer, 2025.

Mai, G., Cundy, C., Choi, K., Hu, Y., Lao, N., and Ermon, S.: Towards a Foundation Model for Geospatial Artificial

Intelligence (Vision Paper), in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '22, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3557915.3561043, 2022.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., et al.: On the opportunities and challenges of foundation models for geospatial artificial intelligence, 2023.

Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., and Nascetti, A.: PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models, https://arxiv.org/abs/2412.04204, 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision, 2024.

Osco, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., and Marcato, J.: The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot, International Journal of Applied Earth Observation and Geoinformation, 124, 103 540, https://doi.org/10.1016/j.jag.2023.103540, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I.: Learning transferable visual models from natural language supervision, in: International conference on machine learning, pp. 8748–8763, PMLR, 2021.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, 2024.

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A., and Camps-Valls, G.: Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources, Information Fusion, 63, 256–272, 2020.

Sultan, R. I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., and Zhu, D.: GeoSAM: Fine-tuning SAM with Sparse and Dense Visual Prompting for Automated Segmentation of Mobility Infrastructure, 2024.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., and Fu, K.: RingMo: A Remote Sensing Foundation Model With Masked Image Modeling, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–22, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, p. 6000–6010, Curran Associates Inc., Red Hook, NY, USA, 2017.

Wang, Y., Xiong, Z., Liu, C., Stewart, A. J., Dujardin, T., Bountos, N. I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., and Zhu, X. X.: Towards a Unified Copernicus Foundation Model for Earth Vision, https://arxiv.org/abs/2503.11849, 2025.

Werner, M., Li, H., Zollner, J. M., Teuscher, B., and Deuser., F.: Bavaria Buildings - A Novel Dataset for Building Footprint Extraction, Instance Segmentation, and Data Quality Estimation (Data and Resources Paper), in: Proceedings of the 31st International Conference on Advances in Geographic Information Systems (ACM SIGSPATAIL GIS'23), https://doi.org/10.1145/3589132.3625658, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing, 2020.

Yao, L., Liu, F., Chen, D., Zhang, C., Wang, Y., Chen, Z., Xu, W., Di, S., and Zheng, Y.: RemoteSAM: Towards Segment Anything for Earth Observation, in: Proceedings of the 33rd ACM International Conference on Multimedia, MM '25, p. 3027–3036, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3746027.3754950, 2025.

Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., Ross, C., Polyak, A., Howes, R., Sharma, V., Xu, P., Tamoyan, H., Ashual, O., Singer, U., Li, S.-W., Zhang, S., James, R., Ghosh, G., Taigman, Y., Fazel-Zarandi, M., Celikyilmaz, A., Zettlemoyer, L., and Aghajanyan, A.: Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning, 2023.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y.: DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, 2022.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H.: Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling, 2021.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt, International Journal of Machine Learning and Cybernetics, pp. 1–65, 2024.