GroundSet: A Cadastral-Grounded Dataset for Spatial Understanding with Vector Data

Roger Ferrod[1], Maël Lecene[1], Krishna Sapkota[2], George Leifman[2], Vered Silverman[2], Genady Beryozkin[2], and Sylvain Lobry[1]

[1] Université Paris Cité, LIPADE, F-75006 Paris, France
sylvain.lobry@u-paris.fr
[2] Google Research

Abstract. Precise spatial understanding in Earth Observation is essential for translating raw aerial imagery into actionable insights for critical applications like urban planning, environmental monitoring and disaster management. However, Multimodal Large Language Models exhibit critical deficiencies in fine-grained spatial understanding within Remote Sensing, primarily due to a reliance on limited or repurposed legacy datasets. To bridge this gap, we introduce a large-scale dataset grounded in verifiable cadastral vector data, comprising 3.8 million annotated objects across 510k high-resolution images with 135 granular semantic categories. We validate this resource through a comprehensive instruction-tuning benchmark spanning seven spatial grounding tasks. Our evaluation establishes a robust baseline using a standard LLaVA architecture. We show that while current RS-specialized and commercial models (e.g., Gemini) struggle in zero-shot settings, high-fidelity supervision effectively bridges this gap, enabling standard architectures to master fine-grained spatial grounding without complex architectural modifications. Data, pretrained model and code are available at: `https://huggingface.co/datasets/RogerFerrod/GroundSet`

Keywords: Remote Sensing · Multimodal LLMs · Spatial Understanding · Geospatial Vector Data

1 Introduction

Multimodal models have become ubiquitous, embedding themselves into any sort of modern application. In particular, the advent of Multimodal Large Language Models (MLLMs) [11,19,26] has fundamentally reshaped the AI landscape. Driven by unprecedented scaling, MLLMs have demonstrated remarkable zero-shot capabilities and broad world knowledge, enabling them to solve a diverse array of tasks with minimal supervision. Furthermore, when integrated into agentic frameworks, these models can orchestrate external tools to address complex, multi-step problems [10,15]. Despite these advancements, however, a critical limitation persists: current MLLMs exhibit a severe deficiency in fine-grained spatial understanding [1, 5]. They significantly struggle with fundamental vision tasks –

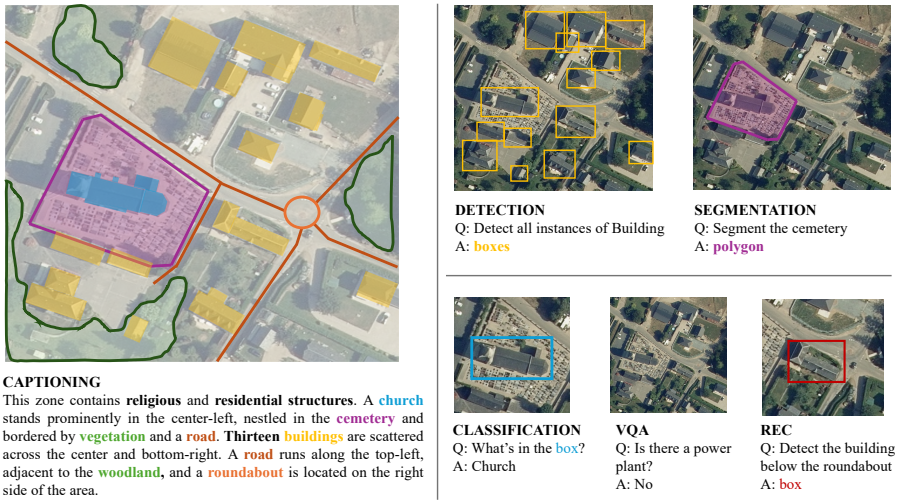


Fig. 1: Overview of our dataset. We release both the raw vector data and a derived instruction-tuning dataset. The cadastral vector layers (e.g., buildings, vegetation, roads) serve as the ground truth and are automatically transformed into instructions for downstream tasks including scene captioning, object detection, semantic segmentation, localized classification, Visual Question Answering (VQA) and Referring Expression Comprehension (REC).

such as precise object detection and segmentation – which are prerequisites for deployment in high-stakes domains like Remote Sensing (RS), where sub-meter precision is critical and current MLLM inaccuracies actively prevent operational use.

The adaptation of MLLMs to Earth Observation (EO) is compelling yet practically stalled by the scarcity of high-quality spatial annotations. While unlabelled satellite imagery is abundant, dense and accurate annotations remain a bottleneck: human labeling is cost-prohibitive and lacks the scalability required to support modern machine learning research. Consequently, recent literature (e.g., [14, 17, 20, 22]) has increasingly relied on repurposing legacy datasets originally designed for closed-set object detection (e.g., DOTA [24], DIOR [16]). These works typically transform older data by synthetically generating question-answer pairs to fit modern instruction-tuning formats. However, this approach is inherently bounded by the quality, limited semantic scope and nature of the underlying source data, failing to address the fundamental issues of scalability and the acquisition of novel domain knowledge.

In this work, we propose a novel, large-scale dataset comprising 3.8M annotated objects across 510k high-resolution images, specifically tailored for spatial understanding in EO. The primary innovation lies in our data curation methodology: rather than relying on legacy sets or noisy crowd-sourcing, we leverage cadastral vector data as our source of ground truth. This approach ensures the

acquisition of high-fidelity geometric data that is manually curated and legally verified by national agencies, effectively solving the scalability-quality trade-off. Moreover, this rigorous sourcing allows us to introduce novel imagery while offering a semantic depth previously unavailable in the field. Whereas existing datasets are typically restricted to a few dozen generic classes (e.g., [14, 17, 20]), we provide 135 highly specific semantic labels. These include fine-grained categories such as power plants, heritage sites, specific crop types and diverse civil and industrial infrastructures. By exposing models to this level of administrative detail, we provide a challenging new benchmark for fine-grained visual recognition (Figure 1). Furthermore, we posit that this rigorous geometric foundation can support a broad spectrum of research applications beyond the tasks proposed in this paper (e.g., [2–4])

To empirically demonstrate the utility of this vector data, we establish a comprehensive instruction fine-tuning benchmark comprising seven spatial grounding tasks as a primary use case. Our experiments highlight a critical gap in the field: existing RS-specialized models (e.g., GeoChat [14], SkySenseGPT [20]) exhibit severe degradation in zero-shot generalization, sometimes underperforming relative to their base architectures. In contrast, we show that high-fidelity data enables standard architectures to master these tasks without complex architectural modifications. Our fine-tuned baseline significantly improves upon the zero-shot performance of specialized spatial models (e.g., Ferret [27]) and commercial baselines. This performance gap serves as quantitative proof of the unique domain knowledge contained in our cadastral annotations—knowledge that is currently absent from web-scale pre-training.

The main contributions of this work are:

1. **The Data:** A novel large-scale EO dataset distinguished by its verified cadastral origin, offering unprecedented semantic richness (135 classes) and density (3.8M objects).
2. **The Benchmark:** A comprehensive instruction fine-tuning suite comprising 7 spatial understanding tasks, derived from the vector data, to demonstrate its applicability to modern VLM training.
3. **The Baseline:** A fine-tuned model serving as a robust reference for future research, demonstrating that standard VLMs can master complex grounding tasks when trained on high-density verified vector data.

2 Related Work

General-Purpose Spatial VLMs Recent advances in spatial understanding rely on modular architectures like LLaVA [19] and PaliGemma [7]. To enable grounding, pioneering works such as Shikra [9], Kosmos-2 [21] and MiniGPT-v2 [8] treat spatial coordinates as textual tokens, effectively bootstrapping spatial skills from pre-trained LLMs. Ferret [27] further refined this by introducing a hybrid sampler for fine-grained feature extraction. These findings demonstrate the feasibility of embedding visual grounding directly into a single unified model, leveraging LLM's established reasoning and knowledge retrieval capabilities.

Remote Sensing MLLMs The Remote Sensing community adapted these paradigms via "continue pre-training" strategies. Early adopters GeoChat [14], SkySenseGPT [20], EarthGPT [29] and RSGPT [13], followed by recent works like GeoGround [30] and GeoPixel [22], fine-tune generic architectures on domain corpora. While validating RS VQA feasibility, these models remain bottlenecked by the quality and scale of training data.

Datasets Data remains the primary constraint. Most benchmarks (e.g., VRS-Bench [17], GeoChat [14], GeoPixel [22], RSVG [28]) rely on legacy closed-set datasets like DOTA [24] and DIOR [16], which are limited to generic categories. While large-scale initiatives like SkyScript [23] and SatLas [6] rely on OpenStreetMap (OSM) for global coverage, they often trade off resolution and precision. In contrast, our work leverages high-resolution (20cm) cadastral data to provide the dense, verified vector annotations required for deep spatial understanding.

3 Data

To ensure precise ground truth, we leverage cadastral records maintained by state administrations, which offer superior reliability compared to crowd-sourced alternatives. As a proof-of-concept, we utilize France's open-data infrastructure provided by the national mapping agency (IGN), combining high-resolution (20 cm) aerial imagery with legally verified vector data. Our pipeline is provider-agnostic, allowing seamless adaptation to other national standards (e.g., USGS, Ordnance Survey or GSI).

This design choice prioritizes annotation fidelity over immediate global diversity. While platforms like OSM provide expansive coverage, they often suffer from inconsistent taxonomy and geometric inconsistency. In contrast, we argue that for fine-grained spatial instruction tuning, the professional-grade cadastral data is the decisive factor for model performance.

Our dataset spans $85,864$ km^2 across 20 administrative departments, prioritizing urban density while ensuring diverse environments (alpine, maritime, rural). The optical aerial orthophotos, featuring a 20cm spatial resolution, are aligned with a map database containing vector models of a wide spectrum of stationary objects. These include buildings, transport infrastructure, land use zones and water bodies. Entities are modeled with 1-meter precision as (multi)polygons, lines or points. Crucially, annotations go beyond generic categories to include granular usage subtypes (e.g., distinguishing Catholic from Orthodox churches), enabling highly specific semantic recognition.

3.1 Pipeline

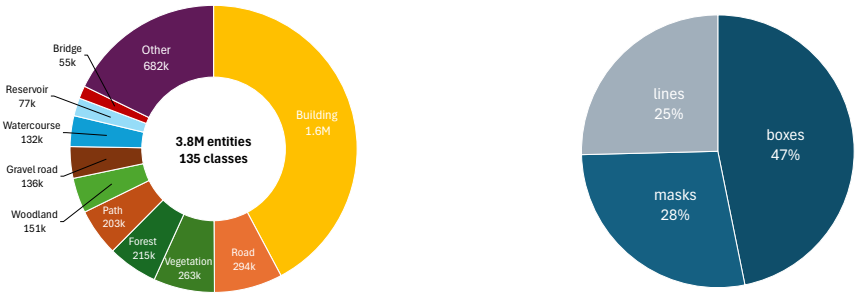
From the selected departments, we extracted over 4 million patches. Each patch has a size of 672×672 pixels, corresponding to a spatial extent of approximately 134×134 meters. These patches underwent a rigorous processing pipeline designed to eliminate redundancy, filter categories, reduce noise and generate instruction sets.

Preprocessing Initially, we filter entities that are not visually discernible in orthophotos, such as tunnels, subterranean conduits or underground parking structures. This ensures strict alignment between visual features and vector data, preventing downstream hallucinations. The 735 unique semantic terms in the database were automatically translated into English and manually verified to serve as ground-truth labels. See Figure 3 for some qualitative examples.

Given the sparsity of rural territories, a significant portion of raw scenes contain minimal information or repetitive patterns. The scene composition follows a long-tail distribution, dominated by single-element images (e.g., uniform forests) or highly frequent combinations. To mitigate this semantic redundancy, we applied a resampling strategy: we group images based on their unique set of present entities (e.g., {'Forest', 'Path'}) and capped each unique combination at a maximum of 64 samples. This strategy reduced the dataset size by 85% while effectively balancing the distribution of scene compositions.

The raw cadastral data often exhibits administrative granularity misaligned with visual semantics. For instance, a single functional structure (e.g., a residence and its detached garage) or a dense urban block may be legally partitioned into multiple units despite appearing as a continuous entity. To solve this issue, we merge contiguous polygons of the same class to ensure the vector ground truth reflects the visual footprint rather than legal partitions, which are often indiscernible.

Furthermore, padding is applied to exclude small or partial objects cropped at image borders which lack sufficient context for recognition. Finally, geographical coordinates are normalized into relative pixel coordinates in the range $[0, 1]$. After the preprocessing phase, the dataset comprises 510,483 patches containing 3,829,755 objects across 135 unique categories (Figure 2).



(a) Occurrences of the top-10 classes. **(b)** Geometric annotations in the dataset.

Fig. 2: The data covers 135 unique semantic categories: a) beyond the top-10, common classes include Canals (44k), Railroads (40k), Parking lots (16k), Tennis courts (15k), Cemeteries (13k) and Stadiums (10k). In terms of geometry, b) the dataset comprises 3.8M entities including linear features (e.g., roads, railroads, canals), bounding boxes (derived from original GIS polygons) and segmentation masks.

Annotations While the original vector data provides polygons, lines and points, we augment these with Horizontal Bounding Boxes (HBB) and Oriented Bounding Boxes (OBB) to support diverse downstream tasks. Segmentation masks are represented as coordinate sequences $[x_1, y_1, \ldots, x_N, y_N]$. HBBs are defined as $[x_{min}, y_{min}, x_{max}, y_{max}]$, while OBBs are represented as $[x_{center}, y_{center}, h, w, \theta]$. Elements with expansive or irregular geometries (e.g., airports, rivers, extensive vegetation) are assigned to a "segmentation-only" category to avoid ambiguous or too wide bounding boxes.

Our dataset maintains a high annotation density (median: 8 objects/image), significantly outperforming standard benchmarks like VRSBench (median: 2). While GeoPixel exhibits a comparable median (9), its density is largely driven by transient objects (e.g., vehicles), whereas our density is derived exclusively from stationary cadastral elements. Furthermore, unlike prior baselines that approximate polygons from raster masks, we utilize native vector data. This ensures precise, human-verified geometric primitives that are topologically accurate and inherently more efficient for sequence-based modeling.

In addition to geometric annotations, we generate descriptive text to synthesize scene information. To ensure scalability, this process is fully automated using state-of-the-art VLMs. However, unlike prior benchmarks (e.g., [17,20,22]) that task VLMs with interpreting visual attributes or background context – effectively distilling possible biases into the dataset – we strictly limit the VLM to elaborate on pre-defined raw captions. We ground the generation of raw captions on rigid geometric rules and templates, constraining the model to merely rephrase existing facts rather than perceive new ones. This prevents the hallucination of non-existent objects or attributes while allowing for the inclusion of novel semantic categories.

Visual-Semantic Discrepancy While cadastral data provides high-fidelity ground truth, the alignment between abstract vector records and optical imagery might contain some discrepancies. We estimate a residual divergence rate comprising: $\sim$3% errors (i.e., annotations without visual evidence, due for example to temporal misalignment); $\sim$5% incomplete records (i.e., visible elements lacking geometry in the cadastre); and $\sim$7% occluded instances (i.e., objects partially or fully covered by vegetation/shadows). Crucially, we argue that the latter category represents a realistic operational challenge rather than noise. In real-world scenarios, detecting an object requires inferring its existence from context (e.g., a driveway disappearing into a forest patch) even when not visible.

To systematically characterize these visual-semantic discrepancies, we implement an automated profiling pipeline using *Gemini-2.5 Flash*. Each annotation is visually highlighted (via Set-of-Marks prompting [25]) and submitted for verification, where the model is tasked solely to confirm the visual presence of the element. Crucially, instances deemed "invisible" are flagged rather than discarded. This approach preserves the integrity of the original cadastral record while providing downstream users with the flexibility to select between "clean" (visually confirmed) and "hard" (occluded/inferred) training splits.

Postprocessing The high granularity of cadastral metadata required semantic simplification for an effective learning. Subtypes lacking distinctive visual features (e.g., *primary* vs. *high school*, *shrubby* vs. *herbaceous heath*) were aggregated into their super-categories (e.g., *School*, *Vegetation*). Conversely, visually distinct functional categories (e.g., *hospitals*, *greenhouses*) were retained.

The resulting label space is hierarchical and not mutually exclusive. A query for a generic parent class (e.g., *Building*) must accept all valid subtypes (e.g., *church*, *train station*) as positive instances. To address this, we release a comprehensive taxonomy to guide training and evaluation.

3.2 Dataset

We provide two distinct data releases: a large-scale pre-training dataset (3.8M objects, 510k images) and a curated VQA dataset tailored for instruction fine-tuning (880k objects, 60k images, 1.8M questions). While this work focuses on the latter to benchmark models and fine-tune our baseline, the full-scale release is intended to support broader research beyond the MLLM paradigm. A comparison with existing benchmarks is provided in Table 1.

Dataset	Tasks	Annotations	Classes	Objs.	Imgs.	Instructions
VRSBench [17]	3/7	HBB	26	52k	29k	205k
GeoPixel [22]	2/7	Poly	15	662k	19k	53k
GeoChat [14]	4/7	OBB	45	–	140k	318k
SkySenseGPT [20]	5/7	OBB	48	210k	82k	**1.8M**
GroundSet (full)	4/7	**HBB, OBB, Poly**	**135**	**3.8M**	**510k**	–
GroundSet (sft)	**7/7**	HBB, Poly	126	880k	60k	**1.8M**

Table 1: Comparison with existing benchmarks across seven core tasks: captioning [14, 17, 20, 22], classification [14, 20], object detection [20] and multi-class detection, segmentation [22], referring expression comprehension [14,17,20] and VQA [14,17,20]. Our contribution includes a large full release with detailed geometry and a smaller supervised fine-tuning subset focused on instructions.

Tasks In our benchmark, we define seven tasks to assess spatial understanding capabilities (Figure 1).

- **Captioning:** generate a coherent paragraph describing the scene.
- **Localized Classification:** classify a given region (could be a box or complex polygon).
- **Object Detection:** localize specific classes using Horizontal Bounding Boxes (HBB).
- **Multi-Class Detection:** localize instances from a list of multiple target classes.

- **Referring Expression Comprehension (REC):** localize an object based on a textual description.
- **Segmentation:** localize target classes using polygonal masks.
- **Visual Question Answering (VQA):** binary verification of object presence.

Instructions We partition the finetuning dataset into 14,000 images for testing and 45,988 images for training. While we provide a comprehensive instruction-tuning dataset generated via templates to train our baseline, we emphasize that the release of the underlying raw vector data enables significantly broader methodological exploration. We encourage the research community to leverage these native geometric primitives to design novel training paradigms that extend beyond the specific instruction formats presented here. Conversely, to ensure consistent and comparable benchmarking, the test set questions are fixed.

Test Set: Evaluation instructions are generated by integrating annotations into diverse templates (10 linguistic variations per question) to ensure generalization. To evaluate robustness against hallucination, we incorporate negative samples (21% probability) into detection and segmentation tasks, prompting the model to locate absent objects. The test set comprises 72,597 instructions distributed across the tasks.

Training Set: For our baseline, we extend the standard task instructions with auxiliary objectives to augment training data. These include exhaustive segmentation queries for every annotated object and text-only geometric reasoning tasks. The latter utilize synthetic random polygons, asking the model to: 1) simplify shapes by removing vertices while preserving topology, and 2) predict bounding boxes for the polygon inputs. These auxiliary tasks are designed to acclimatize the model to the coordinate system (discretized in the interval $[0 - 1000]$). The training set contains in total 1,845,076 instructions.

4 Evaluation

Given the newly acquired imagery and novel semantic categories introduced by our dataset, we benchmark state-of-the-art models in a strictly zero-shot setting to assess generalization. We evaluate a diverse suite of models:

1. Open-Source Baselines: *LLaVA-1.5* [19] and *LLaVA-1.6* [18].
2. Remote Sensing Specialists: *GeoChat* [14], *SkySenseGPT* [20] and a *VRS-Bench* baseline[3].
3. Spatial Grounding Specialists: *Ferret* [27], *MiniGPT-v2* [8] and *PaliGemma-2* [7].
4. Commercial Baseline: *Gemini-2.5 Flash*, representing native multimodal capabilities at scale.

[3] We fine-tune LLaVA-1.5 on VRSBench [17] with the same hyper-parameters reported in their paper.

Fig. 3: Qualitative samples from the proposed dataset, illustrating the semantic granularity and geometric precision of the cadastral annotations across diverse environments.

To ensure a fair comparison, we select open-source models with comparable parameter counts (approx. 7B), with the exception of PaliGemma (10B). While architectural nuances exist, they all share a modular design philosophy that contrasts with the native multimodal architecture of Gemini. Crucially, all Remote Sensing specialists – including our own fine-tuned baseline – are based on the LLaVA architecture. By fixing the model architecture, we effectively isolate the training data as the primary variable, allowing us to rigorously assess the impact of our proposed dataset against existing alternatives.

4.1 Metrics

We adapt our evaluation protocols to accommodate the diverse requirements of heterogeneous models in a zero-shot setting. On the output side, we strictly do not penalize formatting deviations; predictions are geometrically recovered (e.g., converting OBB to HBB or rescaling coordinate norms) prior to scoring. On the input side, standard conversational instructions often yields poor zero-shot performance. Consequently, we tailor the input prompts to match the specific template expected by each architecture[4]. Furthermore, some models are exclusively trained on OBB thus interpreting the HBB as oriented boxes with 90° angle. Thus, questions are systematically adapted per-model to ensure a fair assessment of their true spatial grounding capabilities.

Captioning. We report standard Natural Language Generation metrics: BLEU@4, METEOR and CIDEr. To prevent score distortion, we filter bounding box tokens and unwanted artifacts generated by specialized models prior to evaluation.

Classification. Alongside exact string matching, we employ a Semantic Similarity metric for this open-vocabulary task. Predictions are normalized and, if embedded in verbose sentences[5], extracted using an LLM (*Gemma-3*) to isolate the named entity. A prediction is deemed correct if its embedding (computed via `all-mpnet-base-v2`) exceeds a similarity threshold k with the ground truth (Acc@k), robustly handling synonyms (e.g., *"tennis court"* vs. *"tennis field"*).

Grounding. For Object Detection, Multi-Class Detection, Referring Expression Comprehension and Segmentation, we compute mIoU regardless of the output modality (HBB, OBB or mask). Crucially, evaluation is taxonomy-aware: predicting a valid child class (e.g., *Church*) when queried for a parent class (e.g., *Building*) is considered correct. We select the candidate with maximum IoU per target and report F1, Precision and Recall based on:

- **TP:** IoU ≥ 0.5 with an unclaimed ground truth.
- **FP:** IoU < 0.5 or duplicate prediction for the same ground truth.
- **FN:** Unmatched ground truth object.
- **TN:** Correct rejection (no object predicted when none exists).

[4] For instance, PaliGemma expects specific task prefixes (e.g., `detect` followed by the object name) rather than conversational instructions.

[5] (e.g., "The label that applies is 'bridge'" or "This image likely contains a parking lot")

Visual Question Answering (VQA). Treated as binary classification (Presence/Absence). We report Precision, Recall, F1 and Accuracy after standard text normalization.

4.2 Finetuning

To establish a robust baseline for this benchmark, we fine-tune LLaVA-1.6 7B on the proposed instruction set. We select this architecture as the canonical representative of modular MLLMs, a design paradigm adopted by prominent domain-specific models such as GeoChat and SkySenseGPT[6]. Crucially, LLaVA-1.6 is distinguished by its dynamic resolution capability (AnyRes), which is particularly suitable for processing our high-resolution aerial imagery.

Concerning the implementation, we adhere to a parameter-efficient fine-tuning strategy. The model is trained for a single epoch using LoRA (Low-Rank Adaptation) [12] without the introduction of auxiliary spatial modules or architectural modifications. Training was conducted on $8 \times$ A100 GPUs over approximately 72 hours. By deploying the vanilla architecture, we aim to isolate the contribution of the data from architectural priors, thereby providing a clean and reproducible reference point for future research.

4.3 Results

We benchmark the models on the test set described in Section 3.2. To ensure computational efficiency, all models were evaluated using 4-bit quantization, following preliminary experiments confirming negligible performance impact. We summarize the results in Table 2, where the *RS-ft* column distinguishes between models fine-tuned on Remote Sensing corpora and general-purpose baselines. Each RS-specialist is trained on its own dataset and evaluated zero-shot on our test set. This comparison isolates the impact of training data.

Captioning. Given the strict structural constraints of our dataset's ground truth, only our fine-tuned baseline achieves fully satisfactory alignment. SkySenseGPT marginally outperforms other models, which cluster around similar baseline values.

Classification. In the classification task, a distinct trend emerges: general-purpose models significantly outperform specialized RS models. This suggests the strategy employed by models like SkySenseGPT and GeoChat induces catastrophic forgetting, eroding zero-shot generalization. Conversely, larger commercial models benefit from extensive pre-training on diverse data. However, a critical gap remains: our fine-tuned model surpasses the strongest commercial baseline (Gemini) by a mean margin of 44 points. This substantial gap underscores the unique value of our dataset, which introduces semantic categories and visual distributions absent from web-scale training corpora.

[6] We do not fine-tune RS-specialists because their shared LLaVA architecture would conflate legacy pre-training with GroundSet, preventing a strict data ablation.

	RS ft	Cap. (CIDER)	VQA (F1)	Class. (acc@0.8)	Det. (F1@0.5)	Mult. (F1@0.5)	REC (F1@0.5)	Seg. (F1@0.5)
Gemini-2.5	×	16.68	83.56	49.84	3.79	9.98	4.70	2.79
LLaVA-1.5	×	18.74	87.01	28.62	1.48	2.96	2.71	0.20
LLaVA-1.6	×	17.93	86.36	29.20	0.92	3.96	4.10	0.07
MiniGPT-v2	×	0.24	76.41	6.97	11.69	18.48	17.47	-
PaliGemma-2	×	18.04	85.34	-	17.85	20.42	-	15.12
Ferret	×	18.73	68.20	15.34	18.08	32.83	13.87	19.77
GeoChat	✓	3.99	88.96	7.59	5.52	5.84	6.13	-
SkySenseGPT	✓	24.03	89.69	8.15	3.13	0.14	1.63	-
VRSBench	✓	2.65	87.22	-	2.33	7.26	3.25	-
GroundSet	✓	**104.48**	**97.09**	**94.18**	**49.47**	**72.14**	**39.45**	**44.65**

Table 2: Zero-shot evaluation of state-of-the-art models compared to Remote Sensing fine-tuned baselines (RS-ft) and our new baseline which is trained on the proposed dataset. We observe that: (i) General spatial models (Ferret, PaliGemma) surpass RS-specialists in their own domain; (ii) Previous RS models might underperform their foundational baselines and do not generalize well on our benchmark; and (iii) Commercial-grade solutions (Gemini) show strong general competence but ultimately lack the specific semantic knowledge to close the gap with our fine-tuned model.

Visual Question Answering (VQA). The VQA performance landscape is more nuanced. Unlike classification, this task requires less precise localization, reducing the penalty for models with weak spatial priors. Consequently, RS-specialized models demonstrate a slight advantage over generic baselines due to domain adaptation.

Grounding Tasks. For tasks demanding high-fidelity spatial grounding – such as Object Detection, Multi-Class Detection and Segmentation –architectural priors provide some benefits. Specialized architectures like Ferret (with its hybrid region sampler) and PaliGemma significantly outperform both RS-specialized models and commercial APIs in zero-shot settings. Notably, Ferret achieves the highest zero-shot detection scores, suggesting that specific solutions for general purpose localization are more impactful than domain-specific pre-training alone.

The Data Factor. We utilize a standard LLaVA-1.6 architecture to strictly isolate the impact of training data from architectural priors. Since leading RS-specialists (e.g., GeoChat, SkySenseGPT, VRSBench) are built upon the same LLaVA meta-architecture, they effectively serve as control baselines for legacy training data. The substantial performance disparity observed in Table 2 is therefore not a result of superior model design, but is directly attributable to the semantic richness and geometric precision of our cadastral annotations. This result exposes a critical "knowledge gap" in prior datasets and demonstrates that standard VLMs can master fine-grained spatial understanding when grounded in high-fidelity data, without requiring complex architectural modifications.

Training mix. We validate our training strategy with ablation studies on a 100k-instruction subset, as detailed in Table 3. The results demonstrate that augmenting the training mixture with auxiliary geometric tasks – specifically text-only polygon reasoning and exhaustive segmentation queries described in Section 3.2 – improves spatial grounding capabilities. While these augmentations induce a negligible drop in classification accuracy, they drive meaningful gains in detection and segmentation scores. In contrast, unfreezing the vision encoder yields inconsistent results, notably degrading segmentation performance. Finally, we observe that filtering occluded 'invisible' instances improves the segmentation scores, albeit at a slight cost in other tasks. Based on these findings, we maintain a frozen vision encoder and incorporate both geometric augmentation strategies in our final training recipe.

Cross-Dataset Generalization. To assess robustness beyond our specific data distribution, we evaluate our GroundSet-trained baseline on the VRSBench benchmark in a zero-shot setting. Since the original release is limited to captioning, REC and VQA, we expanded the benchmark by generating new questions for Object Detection directly from the underlying annotations. This evaluation represents a severe out-of-distribution test: VRSBench is heavily tilted toward mobile objects (e.g., vehicles, ships, airplanes) categories strictly absent from our stationary cadastral training set.

Consequently, we observe an uneven transfer across different task types. As shown in Table 4, our model trails RS-specialists in tasks such as Captioning and VQA. We attribute this to the fact that baseline models are trained with differing task definitions; consequently, performance on these tasks is highly sensitive to domain-specific vocabulary and rigid prompt templates. In zero-shot settings, even minor deviations from the expected prompt structure can degrade performance on these language-dependent tasks.

Despite this domain shift, a remarkable trend emerges in the core spatial grounding tasks. Our model outperforms leading RS-specialists in both Detection and Referring Expression Comprehension. This highlights a striking performance asymmetry: while existing RS-specialists collapse when evaluated zero-shot on our dataset (Table 2), our model preserves robust grounding capabilities when transferred to VRSBench. This achievement is particularly notable considering that prior RS-specialists were pre-trained on DOTA and DIOR imagery – the source domains of VRSBench – giving them an in-domain visual advantage, whereas our baseline operates in a strictly out-of-distribution setting. Ultimately, these results clarify that while out-of-the-box cross-domain robustness on language-heavy tasks requires downstream fine-tuning, GroundSet is highly effective at instilling foundational visual grounding. While semantic vocabulary is domain-dependent, the core geometric understanding acquired from our dense vector data successfully transfers to new domains.

	Class. (acc@0.8)	Det. (F1@0.5)	Seg. (F1@0.5)
Baseline	**90.47**	12.77	12.62
Text Aug.	<u>90.06</u>	**13.39**	14.52
Seg. Aug.	89.95	<u>13.28</u>	<u>14.54</u>
ViT Unfrozen	89.93	13.12	11.52
Filtered	89.99	11.95	**15.31**

Table 3: Ablation results (100k subset).Integrating text-only geometric reasoning (*Text Aug.*) and dense segmentation queries (*Seg. Aug.*) boosts spatial task performance over the baseline, while unfreezing the visual encoder proves detrimental. Filtering invisible objects (*Filtered*) significantly improves segmentation at the cost of detection.

	Cap. (CIDEr)	VQA (F1)	Det. (F1@0.5)	REC (F1@0.5)
LLaVA-1.5	21.16	91.54	3.45	9.96
LLaVA-1.6	6.75	81.85	2.84	11.39
GeoChat	17.18	87.97	9.20	12.07
SkySenseGPT	24.36	90.35	12.81	1.67
VRSBench	**45.55**	**92.36**	4.38	9.46
GroundSet	12.97	68.03	**15.47**	**21.49**

Table 4: Cross-Dataset Evaluation (Zero-Shot on VRSBench). Our GroundSet-trained model outperforms RS-specialists in core grounding tasks (Det. and REC). Despite operating strictly out-of-distribution (missing vehicles/planes) against competitors with in-domain pre-training advantages on the source imagery, our baseline demonstrates superior cross-domain spatial understanding.

5 Conclusion

In this work, we addressed the critical scarcity of high-fidelity spatial annotations in Remote Sensing by introducing a large-scale dataset grounded in verifiable cadastral records. Comprising 3.8 million objects across 510k high-resolution images, this release provides a level of semantic granularity (135 classes) and geometric precision previously unavailable in the field. Unlike synthetic or crowdsourced alternatives, our data offers legally verified ground truth, establishing a rigorous foundation for spatial AI research.

To validate this resource, we benchmarked state-of-the-art models on seven spatial tasks. Our experiments demonstrate that fine-tuning a standard, unmodified VLM on this data is sufficient to acquire robust spatial grounding capabilities that are currently beyond the reach of specialized architectures and massive commercial baselines. This result highlights a critical "knowledge gap" in current web-scale models and confirms that our dataset encapsulates new specific and granular domain knowledge.

Future Work. We release both the instruction dataset and the raw vector data to foster further methodological innovation. While our autoregressive baseline establishes a new state-of-the-art, the release of native vector primitives uniquely positions the community to explore entirely new modeling paradigms. We encourage future research to leverage these geometric structures to investigate non-autoregressive architectures – such as GNNs, JEPA-inspired architectures or diffusion LLMs – to go beyond the standard strategies of current VLMs.

Acknowledgements

This work was supported by Google under a research collaboration agreement with Université Paris Cité. We also gratefully acknowledge the support of the Google Cloud Platform (GCP) and HPC resources from GENCI–IDRIS (Grant 2025-AD011016789).

References

1. Adejumo, A., Yeganli, F., Broni-bediako, C., Xiao, A., Yokoya, N., Siam, M.: A vision centric remote sensing benchmark (2025)
2. Audebert, N., Saux, B.L., Lefèvre, S.: Joint learning from earth observation and openstreetmap data to get faster better semantic maps. CVPRW (2017)
3. Bai, L., Zhang, X., Zhang, S., Zhang, Z., Wang, H., Qin, W., Du, S.: GeoLink: Empowering remote sensing foundation model with openstreetmap data. arXiv (2025)
4. Bai, L., Huang, W., Zhang, X., Du, S., Cong, G., Wang, H., Liu, B.: Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. ISPRS Journal of Photogrammetry and Remote Sensing **201**, 193–208 (2023)
5. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. ArXiv (2024)
6. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: SatlasPretrain: A large-scale dataset for remote sensing image understanding. In: ICCV (2023)
7. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D.M., Neumann, M., Alabdulmohsin, I.M., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A.A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J.M., Kabra, R., Bauer, M., Bovsnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Hénaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai, X.Q.: PaliGemma: A versatile 3b vlm for transfer. ArXiv (2024)
8. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. arXiv (2023)
9. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv (2023)
10. Chen, Y., Wang, W., Kurtz, C., Lobry, S.: Automating geospatial vision tasks with large language model agent. In: ECML PKDD (2025)

11. Gemini Team: Gemini: A family of highly capable multimodal models (2024)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
13. Hu, Y., Yuan, J., Wen, C., Lu, X., Liu, Y., Li, X.: RSGPT: A remote sensing vision language model and benchmark. ISPRS Journal of Photogrammetry and Remote Sensing **224**, 272–286 (2025)
14. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S.H., Khan, F.S.: GeoChat:grounded large vision-language model for remote sensing. In: CVPR (2024)
15. Lahouel, L., Lopata, L., Gruening, S., Meoni, G., Petit, G., Lobry, S.: IC-EO: Interpretable code-based assistant for earth observation (2026)
16. Li, K., Wan, G., Cheng, G.: DIOR (2025). `https://doi.org/10.21227/tq1e-nx82`, `https://dx.doi.org/10.21227/tq1e-nx82`
17. Li, X., Ding, J., Elhoseiny, M.: VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In: NeurIPS (2024)
18. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
20. Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., Li, Y.: SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. ArXiv (2024)
21. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding multimodal large language models to the world. In: ICLR (2024)
22. Shabbir, A., Zumri, M., Bennamoun, M., Khan, F.S., Khan, S.: GeoPixel: Pixel grounding large multimodal model in remote sensing. In: ICML (2025)
23. Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R.: SkyScript: a large and semantically diverse vision-language dataset for remote sensing. In: Proc. AAAI (2024). AAAI'24/IAAI'24/EAAI'24 (2024)
24. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: CVPR (June 2018)
25. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv (2023)
26. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision). arXiv (2023)
27. You, H., Zhang, H.A., Cao, L., Gan, Z., Zhang, B., Wang, Z., Du, X., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. In: ICLR (2023)
28. Zhan, Y., Xiong, Z., Yuan, Y.: RSVG: Exploring data and models for visual grounding on remote sensing data. IEEE TGRS **61**, 1–13 (2023)
29. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X.: EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE TGRS **62**, 1–20 (2024)
30. Zhou, Y., Lan, M., Li, X., Ke, Y., Jiang, X., Feng, L., Zhang, W.: GeoGround: A unified large vision-language model. for remote sensing visual grounding. ArXiv (2024)