MapRef: A Benchmark of Referring Expression Comprehension for Maps

Michael J. Bianco,* Jacob Kovarskiy,* Raphael Tang, Jay Patravali, Cassie Burgess, Heming Liao, Matt Washburn, Qi Zhang, Suvam Bag, Talha Kilic, Xiaojie Guo, Zirui Xu, Fuxun Yu

Microsoft Redmond, WA USA

Abstract

Referring expression comprehension (REC) aims to localize or segment visual regions described by natural language expressions. While extensive REC benchmarks exist in natural image domains, there remains a gap in understanding how vision–language models (VLMs) generalize to structured, geospatial, and map-based imagery. In this work, we present **MapRef**, the first large-scale REC benchmark for geospatial map understanding, spanning a broad spectrum of map modalities and spatial scales: generic cartographic maps, weather visualizations, agricultural and land-cover layers, and other domain-specific map types. Using publicly available raster and vector geospatial data (e.g., NOAA, ESA, OSM), we programmatically generate image–text–mask triplets that cover diverse projections, spatial scales, and reasoning types. Our evaluation across a suite of 3 recent SOTA vision–language models reveals a significant performance gap between natural image domain REC and map-based REC tasks: at 0.10 IoU, GPT-5 achieves only 31.5% accuracy for US-county and 20.7% accuracy for global, country level administrative boundaries. **MapRef** serves as a foundation benchmark for spatial reasoning, cross-modality grounding, and geospatial learning.

Introduction

Despite its central role in geospatial applications, map understanding within vision–language models (VLMs) remains comparatively underexplored. Unlike general-purpose images, maps exhibit structured, symbolic, and projection-dependent properties that present distinct challenges for multimodal reasoning. Although recent VLMs show strong progress on natural scene understanding and object-centric grounding, these successes have not fully translated to cartographic domains.

Existing map benchmarks such as MapQA, MapWise, MapBench, and MapIQ mark significant early attempts but remain narrow in scope. They largely focus on question answering (QA), often with limited geographic diversity, template-based prompts, or simplified visualization styles. More importantly, their QA-only formulation can abstract-away the intermediate spatial reasoning steps required to understand maps. Prior work typically organizes map QA tasks by difficulty ranging from element-level queries to higher order synthesis, but these hierarchies do not explicitly capture the fundamental reasoning operations that underpin geospatial understanding.

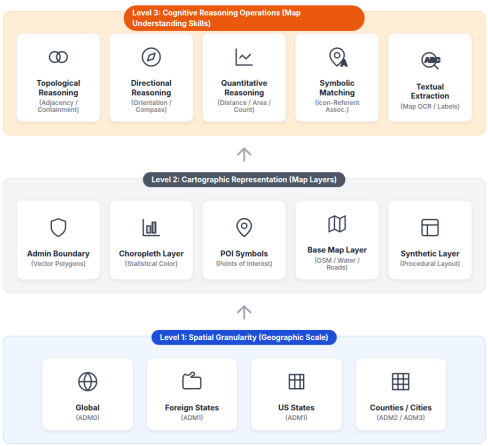

Figure 1: MapRef Taxonomy.

We argue that effective map understanding depends on the ability to **refer, localize, and reason** about spatial entities and their relationships. These include identifying administrative units, interpreting cartographic layers, and linking linguistic expressions to structured spatial representations. QA alone cannot sufficiently evaluate these skills, and chain of thought methods require spatially grounded reasoning chains that these datasets do not provide.

To address these limitations, we introduce MapRef, the first large-scale benchmark for referring expression comprehension (REC) on geospatial maps. The dataset spans diverse domains including geography, weather, infrastructure, and population and is constructed via an inverse generation pipeline that programmatically renders maps and automatically produces aligned image query bbox triplets. In addition, MapRef integrates these dimensions into a unified geospatial REC benchmark comprising the following: synthetic administrative maps across three spatial scales, including administrative boundaries, synthetic spatial fields; choropleth maps derived from both artificial and real statistical attributes; OpenStreetMap-based maps, offering diverse

*Equal contribution.

urban structures paired with vector annotations for POIs and infrastructure; and, lastly, human-annotated maps collected from the wild, which incorporate more open-ended spatial and temporal reasoning prompts. This ensures scalability, semantic consistency, and control over spatial attributes.

We evaluate three leading VLMs on MapRef and analyze their performance across reasoning levels, from single object localization to multi-entity relational grounding. Our results reveal substantial limitations in spatial referencing, topological understanding, and cross-layer association patterns not visible in QA-only evaluations. The VLMs we have evaluated apparently lack geospatial reasoning data in pretraining, or post-training has not enabled these capabilities. Together, MapRef provides a foundation for systematic research in geospatial grounding, spatial reasoning, and cross-modal learning, illuminating both the capabilities and limitations of modern VLMs in understanding cartographic representations.

Related Work

REC in Natural Image Understanding Referring expression comprehension (REC) aims to localize visual regions described by natural language. Foundational datasets such as ReferItGame (Kazemzadeh et al. 2014), RefCOCO, RefCOCO+ (Yu et al. 2016), to name a few, established the modern REC paradigm through short expressions over everyday scenes with bounding-box or mask supervision. Follow-up efforts extended REC to video (e.g., A2D Sentences, Ref-YouTube-VOS) and to cross-modal grounding (audio–visual, multi-sensor variants). Collectively, these benchmarks reveal that REC performance degrades sharply under domain shift, particularly when moving from natural imagery to structured, symbolic, or diagran visual domains.

REC for Remote Sensing In geospatial vision, REC research has focused primarily on overhead object grounding. Datasets such as RSVG and DIOR-RSV(Zhan, Xiong, and Yuan 2023), and NWPU-Refer(Yang et al. 2025) provide expressions for isolated categories like *airplane*, *ship*, or *building*, enabling progress on remote-sensing-oriented REC but remaining narrow in scope. More recent resources such as VRSBench(Wang et al. 2024c) and EarthVQA(Wang et al. 2024a) broaden geospatial reasoning toward captioning and QA, yet still operate on aerial imagery rather than cartographic scenes, and do not evaluate the structured spatial reasoning required for map interpretation.

REC for Maps Compared to natural and remote-sensing imagery, map-based REC remains minimally explored despite maps containing rich symbolic structure, administrative hierarchies, and multi-layered themes (e.g., weather, land use, infrastructure, sociodemographics). Existing map datasets address only limited aspects of map understanding. MapQA(Chang et al. 2022) focuses on simple choropleth QA where supervised models can reach near-saturated accuracy. MapWise(Mukhopadhyay et al. 2024) restricts QA generation to 43 templates across three countries. MapBench(Xing et al. 2025) centers on navigation tasks with fewer than 200 examples. MapIQ(Srivastava et al. 2025) covers choropleths, cartograms, and proportional-symbol maps but is QA-only and omits grounded localization. None of these benchmarks evaluate the core referential, geometric, or cross-layer grounding operations required for full cartographic reasoning.

This gap motivates **MapRef**, the first large-scale REC benchmark designed specifically for maps, spanning diverse cartographic modalities and enabling systematic evaluation of spatial grounding and map-specific reasoning.

MapRef Dataset

This section introduces the design and construction of the MapRef benchmark. We first present an overview of the dataset taxonomy, which organizes map understanding into three structured levels. We then describe the programmatic map synthesis engine that generates large numbers of controlled cartographic scenes. Finally, we introduce the in-the-wild dataset constructed through expert annotation.

Dataset Taxonomy and Hierarchy

As shown in Figure 1, MapRef is organized around a three-level hierarchy that formalizes variability in map structure and reasoning requirements, providing a principled foundation to evaluate core dimensions of map cognition:

- The first level describes **spatial granularity**, which corresponds to the geographic scale of the mapped administrative units. These include global views, foreign country states, United States states, and county- or city-level administrative units. This level controls the density, shape complexity, and spatial resolution of map polygons.
- The second level characterizes **cartographic representation**, which specifies the visual layers present in an image. These include administrative boundary polygons, choropleth layers based on statistical values, point-of-interest symbol layers, basemap layers derived from OpenStreetMap, and synthetic layers created through procedural generation. Differences in these layers influence how spatial objects are visually encoded and how referring expressions must be grounded.
- The third level captures **cognitive reasoning** required for map understanding. These operations include topological reasoning, directional reasoning, quantitative spatial reasoning, symbolic matching, and textual extraction. This level describes the semantic and geometric operations that a vision–language model must perform in order to resolve referring expressions.

This three-level structure underpins the construction of all MapRef data sources and provides a unified framework for evaluating spatial localization performance across diverse map modalities.

Programmatic Map Generation Engine

We build a large-scale, systematically controlled map-generation engine using Cartopy (Met Office 2010 - 2015) and geoBoundaries (geoBoundaries Team 2024). The engine enumerates geographic templates, visual styles, and thematic overlays, exporting raster maps and machine-readable

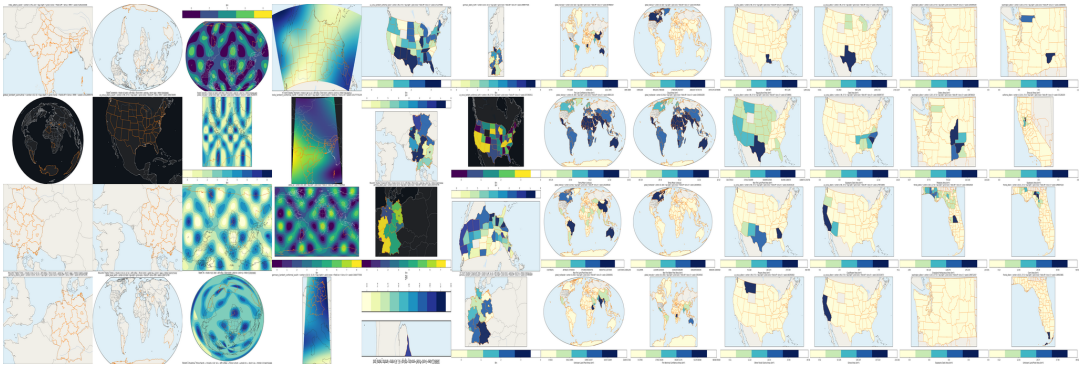

Figure 2: MapRef Cartopy-based Gallery: Diverse and Collected Questions/Answers. Layout similar to level 1 to level 3.

metadata. This enables precise labels for referring expression comprehension while avoiding heavy manual labeling.

Data Sources We produce data-driven choropleths from six real-world geospatial products covering fire activity, temperature, land cover, and elevation. MODIS layers use global 2021 coverage, and land-cover/elevation datasets use 2021 U.S. coverage. Raw rasters are aggregated to administrative units using zonal statistics (mean, max, min for continuous fields; area totals for categorical classes). Data sources include MODIS Thermal Anomalies, MODIS Land Surface Temperature, USDA Cropland Data Layer (CDL), ESA WorldCover, and Copernicus DEM from (Microsoft 2024), and NLCD 2021 from (U.S. Geological Survey 2021).

Geographic Templates and Projections Our library of Cartopy templates spans global projections (Plate Carrée, Robinson, Mercator, Mollweide, Equal Earth, Lambert azimuthal), U.S. regional projections (Albers equal-area, Lambert conformal, Lambert azimuthal), and country-level templates for Germany, Brazil, and India. US state-level templates for California, Florida, and Washington include county overlays. Some regions additionally use offset variants to diversify spatial context.

Visual Factorial Design Each template is rendered under a sweep of visual factors—including background theme, gridline density, orientation (north-up or flipped), presence of synthetic fields, and choropleth quantization settings—to reduce overfitting to stylistic cues. All maps use a fixed DPI for consistent pixel coordinates.

Administrative Overlays and Synthetic Fields We generate three principal map classes: (1) *administrative maps* containing only boundary polygons, isolating shape and topology; (2) *administrative maps with synthetic continuous fields*, mimicking weather or environmental textures while preserving boundaries; and (3) *choropleth maps* with synthetic or real statistical values, supporting controlled experiments and realistic spatial patterns (e.g., land cover, fire counts, cropland, temperature, elevation). Figure 2 shows the gallery of generated maps. All outputs share consistent DPI and figure size.

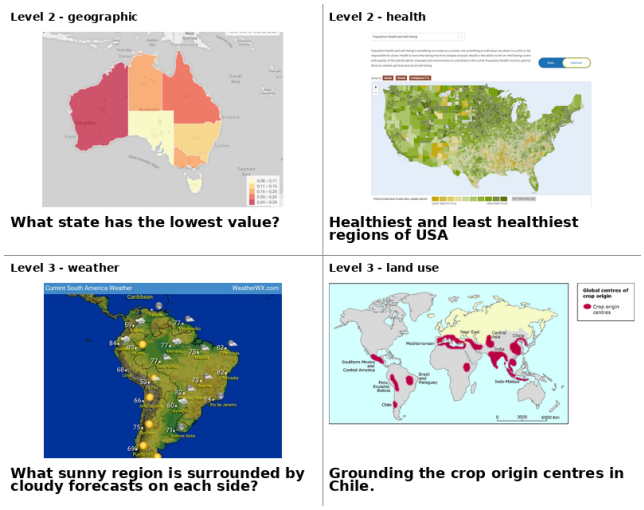

Figure 3: Examples real maps collected from the wild (level 2 and level 3 reasoning shown)

Real Maps Collected "From the Wild"

To capture real-world cartographic variability beyond our synthetic maps, we collected 548 Creative-Commons map images from Google Images using a range of region-specific queries (e.g., "<region> weather map," "<region> topographic map," "GIS map with statistics," "<region> crop map"). Licensing records are included with all images.

A lightweight Streamlit-based annotation UI allowed annotators to view each image, filter invalid examples, write custom questions, and provide answers as either bounding boxes (via a drawing tool) or free-form text. Thirteen geospatial experts produced 1,168 question–answer pairs. Each annotation went through aggregation and a final quality-control pass.

Images were assigned to one of nine categories—weather, land-use, health, disaster, geographic, economic, demographic, environmental, and other—to ensure broad coverage of map domains. The dataset includes 646 single-box answers, 235 multi-box answers, and 198 text-based answers. These examples constitute the Level 3 portion of MapRef, emphasizing open-ended spatial and temporal rea-

soning beyond synthetic settings.

Dataset size

Programmatic Map Generation Our programmatic pipeline produced 1,523 maps with 85,369 total annotations (administrative boundaries and colorbars). The general administrative-boundary subset includes 170 maps with 8,776 annotations (no fields), 420 with 21,472 annotations (with fields), and 332 with 11,016 annotations (random discrete choropleths). Real-data choropleths add 45 country-level maps with 7,965 annotations, 417 county-level maps from CA/FL/WA with 29,468 annotations, and 139 state-level maps with 6,672 annotations.

OpenStreetMaps OSM maps were generated from 500 cities across the US and pulled eight different object types from each city with bounding boxes for each object of interest in the city. City administrative boundaries were pulled from the 500 Cities dataset (Centers for Disease Control and Prevention 2020). Then OSM overpass was queried to obtain locations and bounding boxes for eight different object types of interesting. These objects are: Airports, Energy, Entertainment, Golf Courses, Military, Parks, Stadiums, and Universities. Two instances of each object type was randomly sampled if available in each of the 500 cities. A new map with grounding was generated for each object with the object location randomized in the image.

"Maps in-the-wild" The final dataset resulted in a total of 548 map images and 1168 human labeled quetion-answer pairs. The annotations consist of 646 single bounding box answers, 235 multi-bounding box answers, and 198 open-ended text answers. The dataset consisted of 77 level one, 420 level two, and 668 level three questions respectively. The image category breakdown consists of the following counts: 269 weather, 111 land-use, 36 health, 202 disaster, 286 geographic, 75 economic, 124 demographic, 52 environmental, 11 other.

MapRef Benchmarking

In this section we summarize the evaluation protocol used in MapRef, covering spatial localization, choropleth value prediction, and scale sensitivity.

Grounding Accuracy at IoU Thresholds Spatial grounding accuracy is computed using intersection-over-union (IoU). For each model and map category, we evaluate the proportion of predictions whose bounding box meets or exceeds a specified IoU threshold (e.g., 0.1, 0.2, 0.5). Given N predictions in a category, the accuracy at threshold τ is

$$\text{ground-accuracy}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{IoU}_i \geq \tau],$$

yielding accuracy profiles under both relaxed and strict localization criteria.

Choropleth Value Accuracy For choropleth tasks, we additionally measure whether the predicted discrete value for a region matches the ground truth, conditional on sufficient localization. For each model and category, we filter to examples with $\text{IoU}_i \geq \tau_{\text{choro}}$ (e.g., 0.1 or 0.25) and compute

$$\text{choro-acc}(\tau_{\text{choro}}) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}[\hat{v}_i = v_i],$$

where M is the number of retained features.

Accuracy vs. Object Size and Administrative Class To analyze scale sensitivity and potential pre-training bias, we compute grounding accuracy as a function of ground-truth region size. Each feature's area is derived from its true bounding box, and a binary success indicator (IoU $\geq$ threshold) is recorded. Scatter plots and binned trend curves are generated per model and administrative level to show how localization varies across small and large regions.

Results and Analysis

Benchmark Models

We evaluate one closed-source and two open-source models. The closed-source baseline is GPT-5 (OpenAI 2025). The open-source models are Qwen3-VL and CogVLM-Grounding[1] (Wang et al. 2024b). CogVLM-Grounding recently achieved state-of-the-art performance among open-weight models for general-domain REC (Chen et al. 2025). For Qwen3-VL, we include the 2B and 8B `Instruct` checkpoints to study parameter-scaling effects.

Benchmarking results across reasoning levels appear in Tables 1–4.

Reasoning Levels and Prompting Strategy

Our evaluation spans Levels 1–3 of the MapRef reasoning taxonomy (Fig. 1). Levels 1 and 2 rely on our synthetic map-generation pipeline, and Level 3 uses a human-annotated "maps in the wild" dataset.

Level 1 (Administrative Boundary Grounding). Here we assess the models' ability to localize administrative regions (e.g., country, state, county). The referring expression takes the form `<admin_level> of <admin_name>`, such as "State of California" or "Country of Uganda."

Level 2 (Choropleth and Base-Layer Interpretation). We evaluate two capabilities: (1) grounding an administrative region and (2) extracting its corresponding choropleth value. For choropleth maps, the model must output both the region bbox and the quantitative or categorical value. In addition, for administrative levels 1 and 2, the model must localize the map's colorbar.

We also evaluate basic OSM-style map elements through single-object detection tasks. Referring expressions specify the object to locate, e.g., "The military facility" or "The airport named Albert Whitted Airport."

[1] `THUDM/cogvlm-grounding-generalist-hf` on HuggingFace; 17B parameters.

Table 1: Level 1 REC accuracy across models and administrative levels.

Model	GPT-5			Qwen3-VL-2B			Qwen3-VL-8B			CogVLM		
Admin/*IoU*	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
Country	20.7	13.9	3.0	1.5	0.3	0.1	10.1	5.5	1.1	1.7	0.7	0.1
State	25.7	16.4	3.1	9.1	3.7	0.1	27.7	16.7	3.4	9.3	4.9	1.7
County	31.5	21.6	3.1	3.9	1.1	0.0	10.0	4.4	0.0	7.3	4.0	0.6
Colorbar	99.7	98.5	17.8	63.0	59.9	13.3	94.0	85.8	25.9	100.0	100.0	100.0

Level 1 Prompt

```
You are a vision grounding annotator. Output JSON only.
No prose. Return this JSON exactly:
{"label":"<expression>","bbox":[x0,y0,x1,y1]}
Rules:
- bbox are FRACTIONAL floats in [0,1] relative to the
  image you see.
- x0 < x1 and y0 < y1. Clip to [0,1].
```

Level 2 Prompt

```
You are a vision grounding annotator. Output JSON only.
No prose. Return this JSON exactly:
{"label":"<expression>","feature_bbox":[x0,y0,x1,y1],
 "choropleth_value":"<value>","colorbar_bbox":[x0,y0,x1,y1]}
Rules:
- All bbox coords are FRACTIONAL floats in [0,1].
- feature_bbox: location of "{label}" on the map.
- choropleth_value: the numeric or  categorical value from
  the colorbar.
- colorbar_bbox: location of the colorbar/legend.
- x0 < x1 and y0 < y1. Clip to [0,1]. JSON only.
```

Figure 4: Prompt templates for Level 1 and Level 2 grounding and value prediction.

Level 3 (Human-Annotated Maps in the Wild). Finally, we assess higher-level spatial and temporal reasoning using single-object annotations from our human-curated dataset. Queries include spatial-relational and value-based questions, such as "Which regions located immediately northeast of Palestine have many red or pink legend?" and "What state has the most confirmed cases?" Full multi-object and multi-step reasoning evaluation is left for future work.

Prompting. In Fig. 4, we show examples of prompts used for Qwen3-VL Level 1 and Level 2 tasks (GPT-5 used analogous prompts; CogVLM required simplified versions).

Benchmark Performance

Across all experiments, we evaluate models over three progressively more complex reasoning levels. From the general administrative-boundary and choropleth datasets, we sample 10% of all synthetic map annotations, producing 7605 evaluation examples. From the OSM dataset, we likewise sample 10% of maps, yielding 636 examples. For the human-annotated higher-level reasoning benchmark, we use the full set of 646 single-object QA pairs.

Level 1 Reasoning: Administrative Grounding Table 1 reports Level 1 grounding performance across administrative units and IoU thresholds. Models show consistently strong colorbar grounding—GPT-5, Qwen3-VL-8B, and especially CogVLM achieve near-perfect accuracy at low IoU (CogVLM reaches 100.0% for all thresholds shown). In contrast, all models struggle to localize administrative boundaries. GPT-5 performs best overall (e.g., 31.5% at county-level IoU 0.1), followed by Qwen3-VL-8B, CogVLM, and Qwen3-VL-2B. These results highlight a sharp divergence between grounding simple visual elements (colorbars) and the more demanding geometric and domain-specific reasoning required for administrative boundaries.

Table 2: Level 2 REC accuracy across models and administrative levels.

Model	GPT-5			Qwen3-VL-2B			Qwen3-VL-8B		
Admin/*IoU*	0.1	0.2	0.5	0.1	0.2	0.5	0.1	0.2	0.5
Country	4.4	1.8	0.3	1.2	0.5	0.1	9.2	4.8	1.0
State	17.0	10.0	0.8	4.5	2.5	0.0	29.1	17.8	3.1
County	16.7	10.5	0.9	2.4	1.0	0.0	6.1	2.9	0.4
Colorbar	100.0	99.8	22.1	44.9	40.5	12.1	97.1	89.1	42.8
Choro@REC	6.6	7.2	5.6	3.1	3.0	0.0	11.8	9.9	16.7

Table 3: REC accuracy by category for OSM data.

Model	GPT-5			Qwen3-VL-8B			CogVLM		
Label/*IoU*	0.10	0.20	0.50	0.10	0.20	0.50	0.10	0.20	0.50
Airports	67.3	57.1	16.3	18.4	16.3	8.2	69.4	63.3	38.8
Energy	20.5	12.3	0.0	8.7	4.8	3.8	26.0	18.3	6.7
Entertainment	17.6	10.6	0.0	9.1	9.1	2.7	20.9	16.4	2.7
Golf Courses	46.3	34.1	3.7	12.2	11.1	3.3	29.2	23.6	10.1
Military	45.7	39.1	8.7	18.6	13.6	8.5	29.8	26.3	21.1
Parks	30.1	23.3	4.1	17.2	16.1	10.3	34.5	31.0	23.0
Stadiums	50.0	38.6	10.0	14.5	13.2	10.5	61.8	57.9	34.2
Universities	45.0	30.0	2.5	8.2	4.9	0.0	20.3	11.9	5.1
Overall	38.0	28.6	5.0	12.7	10.7	5.7	34.2	28.8	15.7

Level 2 Reasoning: Choropleth Value Retrieval Table 2 summarizes performance when models must jointly localize a region and retrieve its choropleth value. Colorbar grounding remains strong, yet administrative localization weakens further under the combined task, particularly for GPT-5 and Qwen3-VL-2B. Qwen3-VL-8B improves slightly at the state level but remains limited overall. Value retrieval accuracy is low: at IoU 0.1, GPT-5 attains 6.6% Choro@REC, Qwen3-VL-2B reaches 3.1%, and Qwen3-VL-8B 11.8%. Multiple prompting strategies for CogVLM were unsuccessful or produced malformed outputs, indicating difficulty with multi-field JSON reasoning. Overall, Level 2 exposes a key weakness: models cannot integrate spatial grounding

with discrete value extraction in a composed pipeline.

Level 2 Reasoning: OSM Object-of-Interest Detection OSM-based results (Table 3) show higher performance on feature categories such as airports, stadiums, and parks, reflecting reliance on familiar shapes and text cues from natural-image pretraining. GPT-5 achieves the strongest overall accuracy (38.0% at IoU 0.1), followed by CogVLM (34.2%) and then Qwen3-VL-8B. All models degrade sharply at IoU 0.5, indicating persistent difficulty with fine-grained geometric precision. Relative success on OSM features compared to administrative boundaries suggests stronger priors for naturalistic map symbols than for abstract administrative boundary shapes.

Level 1–3 Reasoning: Human-Annotated High-Level REC Table 4 shows results on human-authored questions spanning reasoning levels 1–3, requiring multi-object reasoning and semantic interpretation. Qwen3-VL-8B performs best (37.4% at IoU 0.1), followed by GPT-5 (25.9%) and CogVLM (20.3%). Despite low absolute accuracy, models appear to generalize better to natural-language phrasing than to synthetic templates. As in all previous tasks, accuracy drops substantially at higher IoU thresholds, underscoring the unresolved challenge of precise geometric grounding even when semantic reasoning succeeds. Taken together, Level 3 results illustrate a persistent gap between coarse object identification and precise geometric grounding, reinforcing conclusions from Levels 1 and 2.

Table 4: Human-annotated map REC accuracy across all labels, all reasoning levels (1-3).

Model/*IoU*	0.10	0.20	0.50
GPT-5	25.9	12.9	3.4
Qwen3-VL-8B	37.4	27.6	6.5
CogVLM	20.3	10.8	5.4

Conclusions

Across all datasets and reasoning levels, the geospatial REC performance of the pretrained VLMs we evaluated remains limited. The models reliably ground simple visual elements such as colorbars, but struggle with administrative boundaries, multi-step value retrieval, and fine-grained geometric localization. OSM-based detection is comparatively stronger—likely reflecting overlap with natural-image pretraining—but still far from sufficient for practical map reasoning. These results highlight the need for targeted training, map-specific embeddings, and RLHF/GRPO-style finetuning aimed at spatial grounding.

Gaps Between Map REC and Natural-Image REC

Our results reveal a substantial gap between map-specific reasoning and natural-image REC benchmarks such as RefCOCO and MDETR. Maps require precise geometric grounding, symbolic interpretation, and multi-level administrative understanding—skills not acquired during general pretraining. MapRef potentially provides a systematic domain for RL-based approaches (e.g., GRPO) to develop spatial reasoning capabilities absent in natural-image settings.

References

Centers for Disease Control and Prevention. 2020. 500 Cities: City Boundaries. https://catalog.data.gov/dataset/500-cities-city-boundaries. Accessed: 2025-11-20.

Chang, S.; Palzer, D.; Li, J.; Fosler-Lussier, E.; and Xiao, N. 2022. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*.

Chen, J.; Wei, F.; Zhao, J.; Song, S.; Wu, B.; Peng, Z.; Chan, S.-H. G.; and Zhang, H. 2025. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 513–524.

geoBoundaries Team. 2024. The geoBoundaries Global Database of Political Administrative Boundaries. https://www.geoboundaries.org.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Met Office. 2010 - 2015. *Cartopy: a cartographic Python library with a Matplotlib interface*. Exeter, Devon.

Microsoft. 2024. Microsoft Planetary Computer. https://planetarycomputer.microsoft.com. Accessed: 2025.

Mukhopadhyay, S.; Rajgaria, A.; Khatiwada, P.; Gupta, V.; and Roth, D. 2024. Mapwise: Evaluating vision-language models for advanced map queries. *arXiv preprint arXiv:2409.00255*.

OpenAI. 2025. GPT-5. https://openai.com. Accessed: 2025-11-18.

Srivastava, V.; Lei, F.; Mukhopadhyay, S.; Gupta, V.; and Maciejewski, R. 2025. MapIQ: Benchmarking Multimodal Large Language Models for Map Question Answering. *arXiv preprint arXiv:2507.11625*.

U.S. Geological Survey. 2021. National Land Cover Database (NLCD) 2021. https://www.usgs.gov/centers/eros/science/national-land-cover-database. Accessed: 2025.

Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; and Zhong, Y. 2024a. EarthVQA: Towards Queryable Earth via Relational Reasoning-Based Remote Sensing Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*. ArXiv preprint arXiv:2312.12222.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2024b. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37: 121475–121499.

Wang, Z.; Yang, X.; Dai, X.; Luo, W.; Cao, Y.; Wang, Y.; and Qiao, Y. 2024c. VRSBench: A Unified Benchmark for Visual Reasoning in Remote Sensing. *arXiv preprint arXiv:2401.00859*.

Xing, S.; Sun, Z.; Xie, S.; Chen, K.; Huang, Y.; Wang, Y.; Li, J.; Song, D.; and Tu, Z. 2025. Can Large Vision Language Models Read Maps Like a Human? *arXiv preprint arXiv:2503.14607*.

Yang, Z.; Yao, H.; Tian, L.; Zhao, X.; Li, Q.; and Wang, Q. 2025. A Large-Scale Referring Remote Sensing Image Segmentation Dataset and Benchmark. *arXiv preprint arXiv:2506.03583*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European conference on computer vision*, 69–85. Springer.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.