MapRef: A Referring Expression Benchmark for Maps

Michael J. Bianco[*], **Jacob Kovarskiy**[*], **Raphael Tang, Qi Zhang, Xiajie Guo, Jay Patravali, Cassandra Burgess, Heming Liao, Matthew S. Washburn, Suvam Bag, Talha Kilic, Zirui Xu, Nikolaos Karianakis, Fuxun Yu**
Microsoft
v-mikebianco@microsoft.com

Abstract

Referring expression comprehension (REC) localizes or segments image regions described by natural language. Despite substantial progress on natural-image benchmarks, it remains unclear how well vision–language models (VLMs) transfer to structured, geospatial, and map-based imagery. In this work, we introduce **MapRef**, a large-scale REC benchmark for map understanding covering diverse map modalities and spatial scales, including general geographic maps, weather visualizations, land-cover layers, and other domain-specific map types. Evaluating three state-of-the-art VLMs, we find a pronounced performance drop from natural-image REC to map-based REC, highlighting significant challenges and opportunities in spatial reasoning and cross-modality grounding for maps.

1 Introduction

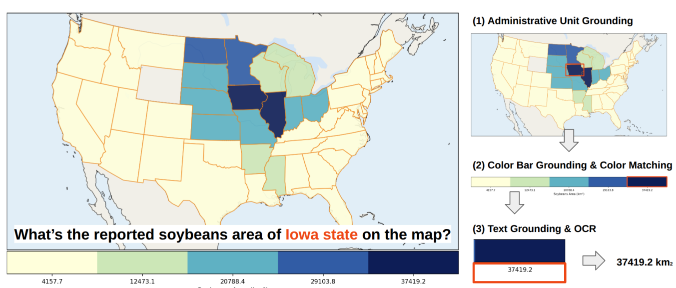

Figure 1: REC is a fundamental capability for map reasoning.

Although recent vision–language models (VLMs) excel at natural scene understanding and grounding Yu et al. (2016), we find that their strong general performance does not fully transfer to geospatial maps. Unlike natural images, maps are structured, symbolic, and geospatial knowledge-dependent, posing distinct challenges for multimodal reasoning. Existing map benchmarks (e.g., MapQA Chang et al. (2022), MapWise Mukhopadhyay et al. (2024), MapBench Xing et al. (2025), and MapIQ Srivastava et al. (2025)) are important first steps, but their predominant question answering (QA) formulations are often too high level to diagnose why a model succeeds or fails. Specifically, if a model cannot *refer* to the correct map element (e.g., a legend/colorbar), *localize* the referenced region, or *reason* about spatial relations (e.g., "left"/"right" mapping to west/east, "north of," "near" vs. "far"), it can fail for a simple question or merely appear correct for the wrong reason via dataset biases or shortcuts. Consequently, QA-only benchmarks provide only limited signals for map-centric reasoning capability evaluation, and further, are ill-suited for the purpose for training VLMs to enhance such reasoning capability.

With the above insights, we argue that effective map understanding hinges on a progression of **referring expression comprehension (REC)** capabilities (as shown in Fig. 1): (i) grounding of common geographic entities (e.g., basic units within annotated boundaries such as administrative ones), (ii) composed grounding and interpretation of thematic layers (e.g., localize a region, interpret color and find the colorbar, and read a choropleth value), and (iii) chains of referring involving finding relational, multi-hop, or temporal cues, evaluation of which all current map datasets lack.

[*] Equal contribution

To address this limitation, we introduce **MapRef**, a large dataset benchmark for REC on maps. MapRef spans domains covering geography, weather, infrastructure, population, and more. It unifies multiple data sources into a single geospatial REC benchmark, including programmatically generated administrative maps at different spatial scales (world view, country, state, etc), choropleth maps derived from both synthetic and real statistical attributes, and OpenStreetMap-based maps capturing diverse urban structure with vector annotations for points of interest (POIs) and infrastructure. MapRef additionally poses challenging referring expression tasks based on annotated map images collected from in-the-wild, that require richer spatial (and occasionally temporal) reasoning. In total, MapRef collects 8,431 maps with 92,628 annotations for REC evaluation, serving as the first benchmark toward reliable and scalable map REC evaluation and potential training.

We evaluate three leading VLMs on MapRef and analyze performance across reasoning levels, from single-entity localization to multi-entity relational grounding. The results expose substantial gaps in spatial reference resolution, topological reasoning, and cross-layer association that are largely masked by QA-only settings. For example, at 0.20 IoU GPT-5 can achieve only 13.9% accuracy for grounding global countries, and 10.0% grounding accuracy for answering simple multi-hop questions about US, Brazilian, and German states as in Figure 1. These findings suggest that current VLM pretraining and post-training data has not yet provided sufficient supervision for map-centric grounding. By systematically evaluating and revealing the limitations of VLMs on the map modality, we aim for MapRef to be a foundation for further research in geospatial map grounding, spatial reasoning, and cross-modal learning.

2 The MapRef Dataset

2.1 Taxonomy and Hierarchy

As shown in Figure 2, MapRef is collected under a three-level hierarchical design, which serves as a set of explicit guidelines for dataset construction. This hierarchy ensures that MapRef is scalable, multi-aspect, and comprehensive, while remaining systematically controlled. Each level isolates a core dimension of map understanding, enabling fine-grained coverage of map variability and reasoning requirements. Specifically, **Level-1** defines spatial granularity, governing the geographic scale of the mapped administrative units. It spans global views, subnational regions in foreign countries, U.S. states, and finer-grained county- or city-level units, thereby controlling the scope and resolution at which spatial entities are represented. **Level-2** specifies cartographic representation, describing the visual layers included in each map image. These layers comprise administrative boundary polygons, choropleth encodings of statistical attributes, point-of-interest symbol layers, and basemap layers derived from OpenStreetMap. This level regulates how spatial information is visually encoded and combined across maps. **Level-3** captures the cognitive reasoning operations required to interpret and ground references in maps. It includes topological and directional reasoning, quantitative spatial comparison, symbolic matching, and textual extraction, reflecting the diverse semantic and geometric competencies needed for map understanding.

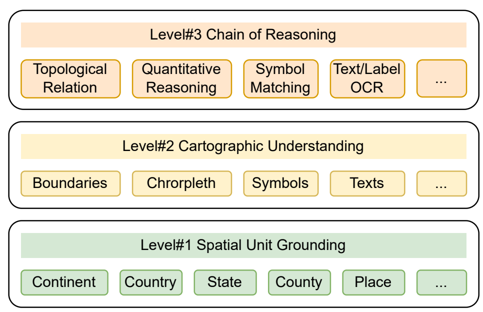

Figure 2: MapRef Taxonomy.

2.2 Dataset Creation

Programmatic Map Generation. We construct a large-scale, systematically controlled map-generation engine using Cartopy Met Office (2010 - 2015) and geoBoundaries geoBoundaries Team (2024). Guided by the MapRef hierarchy, the engine enumerates geographic templates, cartographic styles, and thematic overlays, exporting raster maps together with machine-readable metadata. This design enables precise supervision for referring expression comprehension while avoiding heavy manual annotation. Additional implementation details are provided in Sec. A.1.

Using this engine, we generate three principal classes of maps: (1) *administrative maps* containing only boundary polygons, isolating shape and topological structure; (2) *administrative maps with* *synthetic continuous fields*, which overlays environmental or weather style textures while preserving boundaries; and (3) *choropleth maps* with synthetic or real statistical values, supporting both synthetic and real-world patterns (e.g., land cover, fire counts, cropland, temperature, elevation). In total, we generate 1,523 maps with 85,369 annotations (administrative regions and colorbars), rendered at consistent DPI and figure size. Example outputs are shown in Appendix Fig. A1.

OpenStreetMap (OSM) Data Generation. We additionally construct map instances from OpenStreetMap data across 500 U.S. cities. City administrative boundaries are obtained from the 500 Cities dataset Centers for Disease Control and Prevention (2020), and OpenStreetMap queries are used to extract bounding boxes for eight object categories, including airports, energy infrastructure, entertainment venues, and universities, etc. These maps support fine-grained object grounding and urban-scale spatial reasoning. Additional details are provided in Sec. A.3.

Open-Ended Maps "From the Wild." To complement programmatically generated maps with real-world cartographic variability, we collect 548 Creative Commons–licensed map images using diverse region- and task-specific queries (e.g., "<region> weather map," "GIS map with statistics"). Licensing records are retained for all images. Each image is assigned to one of nine semantic categories (weather, land-use, health, disaster, geographic, economic, demographic, environmental, other) to ensure broad domain coverage. Thirteen geospatial experts annotated and verified 899 question–answer pairs grounded in these images. After annotation, this subset of dataset includes 466 single-box QA, 235 multi-box QA, and 198 text-based QA. These examples constitute the Level 3 portion of MapRef, emphasizing open-ended spatial and temporal reasoning beyond synthetic settings. Further details are provided in Sec. A.2.

3 MAPREF BENCHMARKING

Benchmark Setup We evaluate one closed-source baseline model, GPT-5 OpenAI (2025), and two open-source models, Qwen3-VL Yang et al. (2025) and CogVLM-Grounding[1] Wang et al. (2024). 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 1–3b. For brevity, we leave the OSM evaluation results table to the appendix, Sec. A1.

We evaluate models across the three MapRef reasoning levels according to Fig. 2. Levels 1–2 use synthetic maps with structured JSON grounding prompts, while Level 3 uses human-annotated maps from the wild. Spatial grounding accuracy is computed using intersection-over-union (IoU) (across administrative scales, where available). For choropleth tasks, we additionally measure whether the predicted discrete value for a region matches the ground truth, conditional on sufficient localization. See Appendix, Sec. A.4 for further details and prompt templates.

Some example questions in level 1 to 3 (on maps with self-contained info) are shown below :

- Level 1: *"Identify where the target object 'german state of Sachsen-Anhalt' is located."*
- Level 2: *"What region had the highest maximum temperature anomaly?"*
- Level 3: *"Which water body does the Mississippi river drain into?"*, *"Locate the largest contiguous area of snow and ice."*, *"Where will the typhoon be on Nov 6 at 00:00 GMT?"*

3.1 BENCHMARK PERFORMANCE

Level 1 Reasoning: Administrative Grounding Table 1 reports Level 1 grounding performance across administrative units and standard bounding box IoU thresholds. As we can see, all models struggle to localize units within non-labeled administrative boundaries. GPT-5 performs best overall, but still only 31.5% at country-level with 0.1 IoU. At 0.5 IoU threshold, the numbers drop to 0.3%. This highlights a significant gap for grounding the basic geospatial domain map elements.

Level 2 Reasoning: Choropleth Value Retrieval Table 3a summarizes performance when models must jointly localize a region and retrieve its choropleth value. With compound questions, colorbar grounding shows strong performance, yet administrative localization weakens further under the

[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		
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

(a) Level-2 REC (admin + choropleth).

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

(b) Level-3 In-the-wild REC.

Model/*IoU*	0.10	0.20	0.50
GPT-5	53.4	40.6	12.1
Qwen3-VL-8B	55.8	42.6	12.7
CogVLM	30.0	17.2	8.1

Figure 3: Level-2 and 3 referring expression comprehension (REC) results. (a) Administrative grounding and choropleth value retrieval. (b) REC accuracy on maps from the wild.

combined task. Overall, due to that models cannot conduct geospatial grounding effectively, discrete value extraction in a composed pipeline is further affected with only 0% - 16.7% REC accuracy.

Level 2 Reasoning on OSM Maps OSM-based results (Appendix Table A1) show higher performance on feature categories such as airports, stadiums, and parks, reflecting reliance on familiar shapes and text cues from natural-image pretraining. All models degrade sharply at IoU 0.5, indicating persistent difficulty with fine-grained geometric precision. Higher performance on OSM compared to administrative boundaries suggests stronger priors learning for typical map symbols.

Level 3 Reasoning: Maps from the wild REC Table 3b shows results on human-authored questions spanning reasoning levels 1–3, requiring multi-object reasoning and semantic interpretation. Despite low absolute accuracy, models appear to generalize well. 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.

Figure 4 shows example maps and questions collected from the wild and annotated by human. Facing diverse questions, GPT-5 can suffer from significant performance drop from 0.8 to 0.0 IoU, even though the map understand questions might be easy for human intelligence. We show more representative examples of REC successes and failures with GPT-5 on the human-annotated dataset in the Appendix, Figure A3.

4 Conclusions

The MapRef benchmark reveals a substantial performance gap between geospatial map reasoning and referring expression comprehension on natural images such as RefCOCO. Reading maps demands precise geometric grounding, symbolic interpretation, and multi-level, multi-hop reasoning across spatial and semantic abstractions, which current models appear not to have acquired during pretraining. We hope MapRef serves as a robust diagnostic benchmark for exposing these limitations and as a principled framework for systematic data generation, enabling future supervised fine-tuning and reinforcement learning approaches to more effectively develop map-specific reasoning skills.

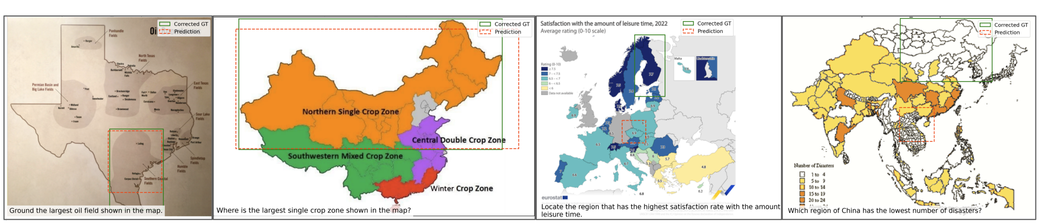

Figure 4: Example maps from the wild and questions. Facing different level of difficulties, GPT-5's grounding IOU can degrade from 0.8 to 0.0 from left to right (best viewed in color).

REFERENCES

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

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 513–524, 2025.

geoBoundaries Team. The geoboundaries global database of political administrative boundaries. https://www.geoboundaries.org, 2024.

Met Office. *Cartopy: a cartographic Python library with a Matplotlib interface*. Exeter, Devon, 2010 - 2015. URL https://cartopy.readthedocs.io.

Microsoft. Microsoft planetary computer. https://planetarycomputer.microsoft.com, 2024. Accessed: 2025.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Vivek Gupta, and Dan Roth. Mapwise: Evaluating vision-language models for advanced map queries. *arXiv preprint arXiv:2409.00255*, 2024.

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

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. Mapiq: Benchmarking multimodal large language models for map question answering. *arXiv preprint arXiv:2507.11625*, 2025.

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human? *arXiv preprint arXiv:2503.14607*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

A APPENDIX

A.1 PROGRAMMATIC MAP GENERATION

A.1.1 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).

A.1.2 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.

A.1.3 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.

A.2 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 466 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 reasoning beyond synthetic settings.

A.3 OPENSTREETMAPS

OSM maps were obtained 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. This resulted in a set of 6360 local maps, with one annotation per map.

A.3.1 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.

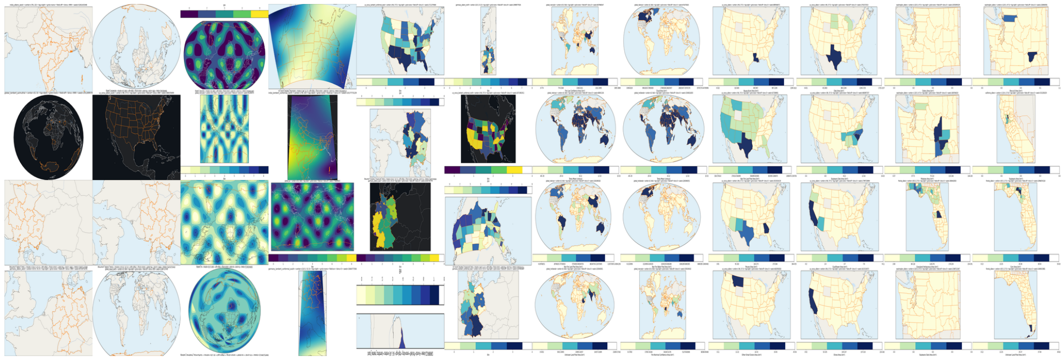

Figure A1: MapRef Cartopy-based Gallery: Diverse and Collected Questions/Answers. Layout similar to level 1 to level 3.

A.3.2 "MAPS IN-THE-WILD"

The final dataset resulted in a total of 548 map images and 1168 human labeled question-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.

A.4 MAPREF EVALUATION DETAILS

A.4.1 DATASET SAMPLING

From the synthetic map annotations (and the OSM dataset) we sample 10% of all synthetic map annotations, producing 7605 (636) evaluation examples. For the human-annotated higher-level reasoning benchmark, we use the full set of 466 single-object QA pairs.

A.4.2 REASONING LEVELS AND PROMPTING STRATEGY

Our evaluation spans Levels 1–3 of the MapRef reasoning taxonomy (Fig. 2). Levels 1 and 2 rely on our 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."

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. A2, we show examples of prompts used for Qwen3-VL Level 1 and Level 2 tasks (GPT-5 used analogous prompts; CogVLM required simplified versions).

A.4.3 GROUNDING ACCURACY AT IOU THRESHOLDS

To analyze scale sensitivity and 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.

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,y⌋
↪  0,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 A2: Prompt templates for Level 1 and Level 2 grounding and value prediction.

Table A1: 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

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.

A.4.4 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.

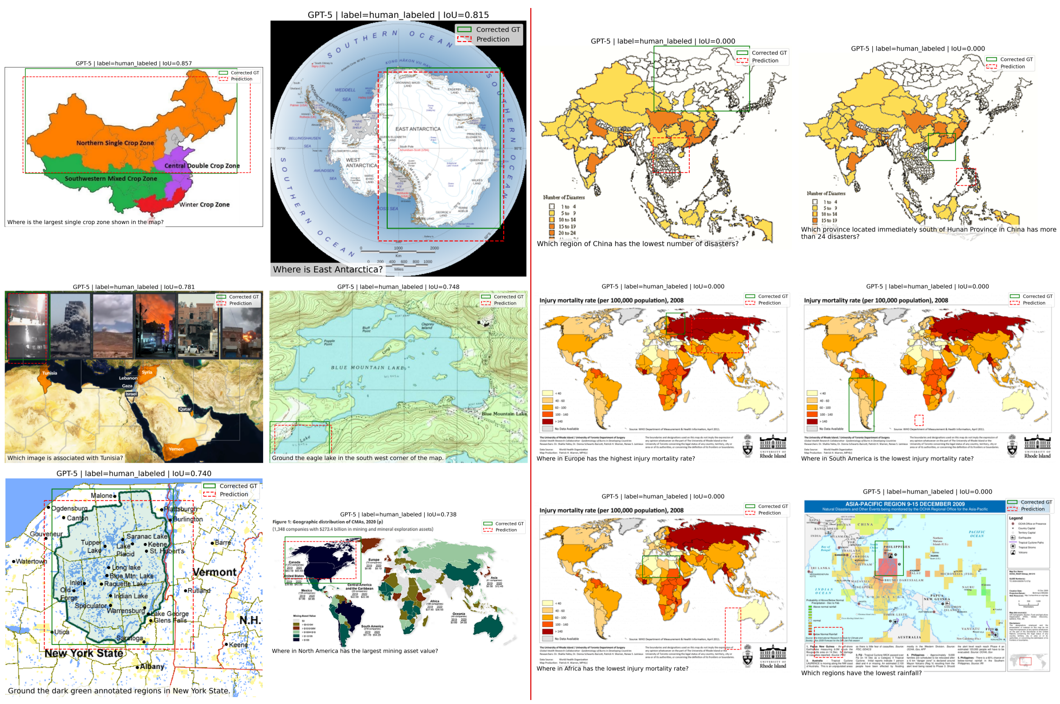

Figure A3: Examples of best and worst predictions from GPT-5 over the *in-the-wild* dataset (see Sec. 2.2). Best is shown on left of dividing red line, and worst on right. The referring expression is on the lower right of each figure, and the ground truth and predicted bounding boxes are shown.