An Analysis of Multimodal LLMs for Object Localization in Earth Observation Imagery

Darryl Hannan, John Cooper, Dylan White, Henry Kvinge, Timothy Doster, and Yijing Watkins

Introduction

- Multimodal LLMs (MLLMs) offer impressive performance across many zero-shot computer vision tasks but struggle with tasks that require fine-grained detection and spatial reasoning.
- Prior work[1] has demonstrated that the same trend holds true for earth observation (EO) tasks.
- Recent MLLMs[2,3] now include explicit localization capabilities, making them better suited for these tasks..
- First study to benchmark these new models on EO object localization tasks and compare their performance to traditional detectors.

Zero-shot Results

Model	RarePlanes mAP@30pix	AAP mAP@30pix	xBD mAP@15pix
Molmo 7B O	62.62	**30.26**	2.97
Molmo 72B	**72.12**	29.82	**4.22**
Qwen 2.5-VL 7B	46.62	30.01	0.49
Qwen 2.5-VL 72B	50.03	12.09	0.50
Llama 3.2 11B	0.00	0.00	0.00
Llama 3.2 90B	0.00	0.00	0.00

Table 1: Object detection results for various MLLMs across three different datasets.

Key Takeaways:

- MLLMs offer strong performance when objects are objects are sufficiently large, the shape is relatively distinct, and the class is not too specific.
- Larger models do not always outperform smaller models.
- The Molmo family of models offers the strongest localization performance in the EO domain.
- MLLMs that are not explicitly tuned to output object coordinates, do not possess the innate ability to do so, despite strong performance across other tasks.

Failure Scenarios and Limitations

- Models tend to produce more false negatives than false positives.
- Objects more likely to be missed when: very small, partially obscured, or close to other objects.
- Potential reasons for false negatives: Lack of extreme precision in point placement, no object confidence scores (unlike traditional detectors), issues scaling to scenes with many objects.
- False positives typically occur with reasonable distractors, but we occasionally see catastrophic failures.

Figure 3: RarePlanes example with Molmo 72B labels. The model successfully predicts most planes but misses a plane that is in close quarters to others and misses two that are partially obscured.

Figure 4: xBD example with Molmo 72B labels. Example of a catastrophic failure, where models will sometimes generate a sequence of many detections in a line. We are uncertain what results in this behavior, but we notice it more with small models.

Models, Datasets, and Metrics

- MLLMs evaluated:
  - Includes localization capabilities
    - Molmo 7B O and Molmo 72B[2]
    - Qwen 2.5-VL 7B and 72B[3]
  - Does not include localization capabilities
    - Llama 3.2 11B and 90B[4]
- Datasets:
  - RarePlanes (RP): 1-class aircraft detection via satellite imagery
  - Aerial Animal Detection (AAP): 3-class animal detection via imagery taken from a helicopter.
  - xBD: 1-class building detection via satellite imagery.
- Metrics:
  - Center Mean Average Precision (mAP): Modified mAP metric that uses a pre-determined pixel distance between center points, rather than bounding box overlap, to compute precision and recall.

Figure 1: Sample outputs using various MLLMs (green dots=ground truth and red Xs=predictions) across three tasks: building detection (left), animal detection (middle), and plane detection (right).

Comparison to Standard Detectors

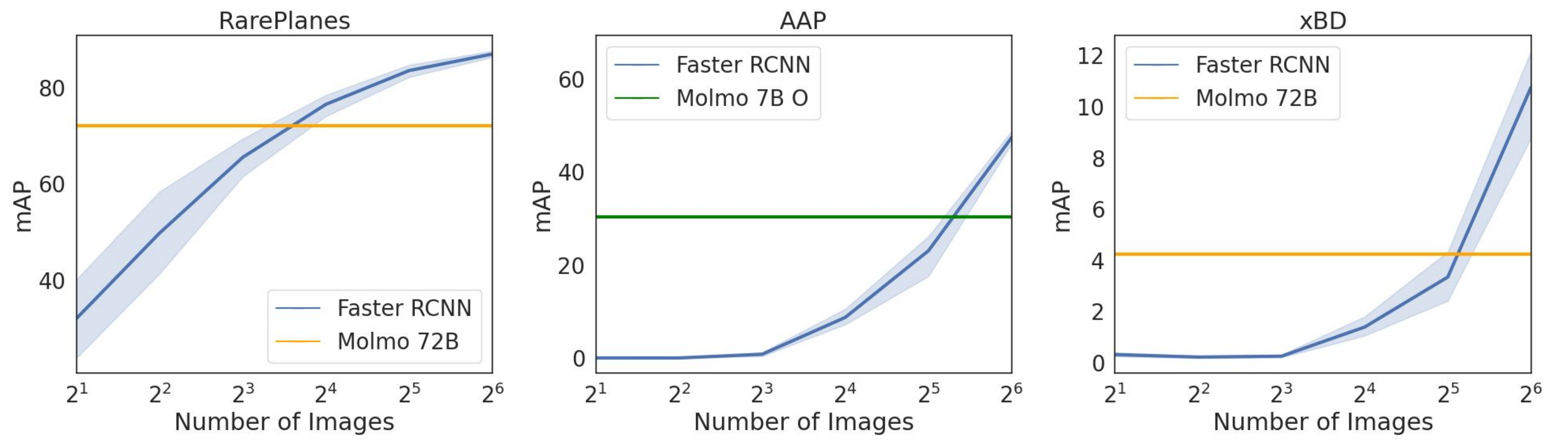

Figure 2: Few-shot Faster RCNN performance with varying amounts of training images (blue lines) vs. top-performing MLLM's performance (alt-color lines) for each task.

Key Takeaways:

- If an MLLM struggles with a task, a standard detector is likely to as well.
- MLLMs offer utility in few-shot and limited data scenarios but are quickly surpassed by standard object detectors as more data becomes available (< 64 examples for the datasets we explored).

References:

[1] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. arXiv preprint arXiv: 2401.17600, 2024.
[2] Matt Deitke, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. arXiv preprint arXiv:2409.17146, 2024
[3] Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/ qwen2.5-vl/.
[4] Abhimanyu Dubey, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Acknowledgements:

The research described herein was funded by the Generative AI for Science, Energy, and Security Science & Technology Investment under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.
This work was also supported by the Center for AI.

For additional information, contact:

Darryl Hannan
darryl.hannan@pnnl.gov