Evaluating LocateAnything for Language-Guided Pre-Grasp Object Localization in Humanoid Robot Perception

Dawei Cai / David Cai
Department of Electrical & Computer Engineering
University of Washington, Seattle
daweic@uw.edu

July 1, 2026

Abstract

This report evaluates LocateAnything as a language-guided visual grounding module for humanoid robot pre-grasp perception. The study focuses on 2D object localization from natural-language prompts in cluttered tabletop scenes. The mini-pilot contains 39 valid image–prompt pairs after excluding one prompt with no valid target object. LocateAnything achieved a mean Intersection over Union (IoU) of 0.646, a median center-point error of 17.2 pixels, and a 76.9% overall success rate using IoU ≥ 0.5 as the success criterion. Spatial-relation prompts reached 93.8% success and similar-object prompts reached 100.0% success, while small-object and ambiguous prompts were more difficult. Inference latency is not reported quantitatively because the web demo did not provide a reliable per-query timing value. This report is limited to perception-stage 2D localization and does not test physical grasping, robot control, or full humanoid deployment.

1 Introduction

In everyday indoor settings, people rarely describe objects using only fixed category labels. A person may ask for "the bottle near the paper," "the key near the pen," or "the small black object near the grid file bag." For a humanoid robot, this type of instruction requires more than detecting object categories. The robot must first localize the intended target in the camera image before any downstream grasp planning or motion planning can be considered.

Traditional object detectors such as YOLO are useful for recognizing predefined categories [1]. However, closed-set detection is limited when a user describes an object through language, spatial relations, or scene context. In a cluttered tabletop scene, detecting that several bottle-like objects exist is not enough; the robot must identify which object the user meant.

Vision-language grounding models offer a possible way to connect natural-language referring expressions to image regions. This direction builds on vision-language representation learning such as CLIP [2] and open-vocabulary grounding systems such as Grounding DINO and OWL-ViT [3, 4]. LocateAnything is a recent vision-language grounding model that predicts target regions from image–text inputs [5]. This report evaluates LocateAnything for a narrow robotics perception task: language-guided pre-grasp 2D object localization in cluttered tabletop scenes.

1.1 Research Question

This project studies the following question:

> Can LocateAnything localize task-relevant tabletop objects from natural-language prompts accurately enough to support pre-grasp perception?

The study does not introduce a new model and does not claim to implement a complete humanoid grasping system. It is a mini-pilot evaluation of an existing visual grounding model for the perception stage before grasp planning.

1.2 Contributions

- Evaluate LocateAnything on a small set of cluttered tabletop images and natural-language prompts.
- Report IoU, center-point error, and success rate across prompt types.
- Analyze qualitative examples involving clear localization, spatial relations, small objects, and ambiguous prompts.
- Discuss limitations for using 2D language-guided localization as a pre-grasp perception component.

2 Background

2.1 Language-Guided Object Localization

Language-guided object localization aims to identify an image region that corresponds to a natural-language expression. Unlike fixed-category detection, the target can be specified by category, attribute, relation, or task context. For example, a prompt such as "the bottle near the paper" requires the model to connect visual object appearance with a spatial relation in the scene.

Open-vocabulary grounding systems are relevant for robot perception because they reduce the need for manually defining every possible target category. They can also support more natural human instructions. However, grounding a prompt in a cluttered scene remains difficult when objects are small, visually similar, partially occluded, or ambiguously described.

2.2 Humanoid Robot Pre-Grasp Perception

Before a humanoid robot can grasp an object, it must determine which object is relevant and where that object is located. A 2D bounding box or center point is not a complete grasp pose, but it can serve as an early perception output for later depth estimation, segmentation, or grasp planning. Prior work has explored language-guided grasping and manipulation [6], but the current evaluation focuses only on the 2D localization step.

3 Methodology

3.1 System Overview

Figure 1 shows the evaluation pipeline. A tabletop image and a natural-language prompt are passed into LocateAnything. The predicted region is converted into a bounding box or center point and compared with a manually annotated ground-truth box.

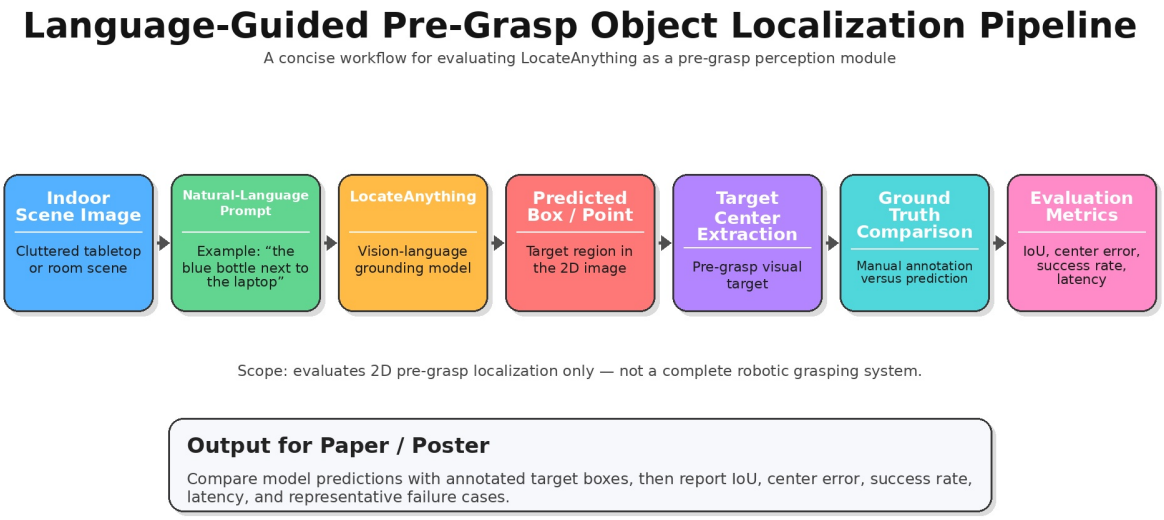


Figure 1: Evaluation framework for language-guided pre-grasp object localization. The pipeline evaluates 2D target localization only and does not perform physical grasp planning or robot control.

3.2 Dataset and Prompts

The mini-pilot uses cluttered desk and tabletop scenes containing common objects such as bottles, cups, keyboards, papers, chargers, cables, pens, keys, and small containers. The final dataset contains 39 valid image–prompt pairs. One prompt was excluded because there was no valid target object for the instruction.

Each image–prompt pair was manually annotated with a ground-truth bounding box for the intended target object. The prompt categories include spatial relation, similar object, small object, occlusion, and ambiguous prompt cases. Attribute prompts had no valid samples in the final dataset, so that category is reported as zero-sample in Table 3.

Table 1: Prompt categories used in the mini-pilot evaluation.

Category	Example Prompt
Spatial relation	the bottle near the paper
Similar object	the water bottle next to the grid file bag
Small object	the key near the pen
Occlusion	the charger behind the iPad
Ambiguous prompt	the bottle

3.3 Evaluation Metrics

The evaluation uses the following metrics:

- **Intersection over Union (IoU):** overlap between the predicted bounding box and the ground-truth box.

- **Center-point error:** pixel distance between the predicted target center and the ground-truth target center.
- **Success rate:** percentage of prompts where IoU ≥ 0.5.
- **Failure mode:** qualitative category of the main localization error.

Inference latency is not quantitatively reported in this pilot version because the LocateAnything web demo did not provide a reliable per-query timing value.

4 Experimental Setup

4.1 Scene Collection

The scene images were collected from a limited tabletop and desk setup. This setup was chosen because tabletop manipulation is a common setting for pre-grasp perception, and it naturally includes clutter, repeated objects, small items, and partial occlusion. The examples include prompts involving paper, bottles, water bottles, keys, chargers, pens, and desk objects.

The evaluation should be interpreted as an early feasibility test rather than a general benchmark. The camera viewpoint, lighting, and object set are limited, and the dataset is too small to support broad claims about humanoid robot deployment.

4.2 Software and Annotation Environment

Table 2: Experimental environment.

Item	Configuration
Model	LocateAnything web demo
Input data	Desk/tabletop RGB images
Prompt type	Natural-language referring expressions
Ground truth	Manually annotated 2D boxes
Success criterion	IoU ≥ 0.5
Latency reporting	Not reported quantitatively

5 Results

5.1 Quantitative Results

After excluding one prompt that did not have a valid target object, the mini-pilot contained 39 image–prompt pairs. Using IoU ≥ 0.5 as the success criterion, LocateAnything achieved a 76.9% overall success rate, with a mean IoU of 0.646 and a median center-point error of 17.2 pixels.

Spatial-relation prompts showed strong performance, with 16 valid samples and a 93.8% success rate. Similar-object prompts also performed well in this small dataset, with 4 valid samples and a 100.0% success rate. Small-object prompts were more difficult, reaching 58.3% success across 12 valid samples. Ambiguous prompts were the hardest group, with 33.3% success across 3 valid samples. Attribute prompts had no valid samples in the final dataset, so no measured conclusion is reported for that category.

Table 3: Quantitative results by prompt type.

Type	N	IoU	Error	Success	Main Failure
Attribute	0	–	–	–	no samples
Spatial relation	16	0.792	12.8	93.8%	one wrong target/reference
Similar object	4	0.818	11.3	100.0%	none observed
Small object	12	0.490	22.2	58.3%	small/thin targets
Occlusion	4	0.650	33.5	75.0%	partial occlusion
Ambiguous prompt	3	0.262	167.2	33.3%	ambiguous target selection
Overall	39	0.646	17.2	76.9%	small or ambiguous targets

5.2 Qualitative Successful Examples

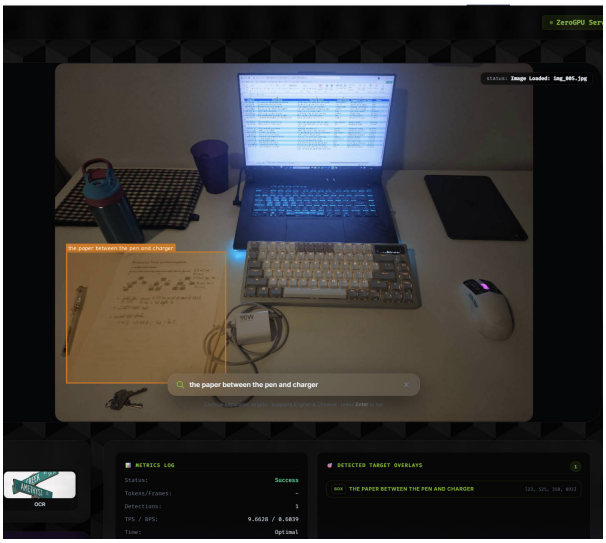
Figure 2 shows successful localization examples. The filenames are preserved from the original image annotations. The first two examples are included as successful qualitative outputs, but the final quantitative table still reports zero valid attribute samples because the final dataset did not include measured attribute-prompt rows.

6 Failure Case Analysis

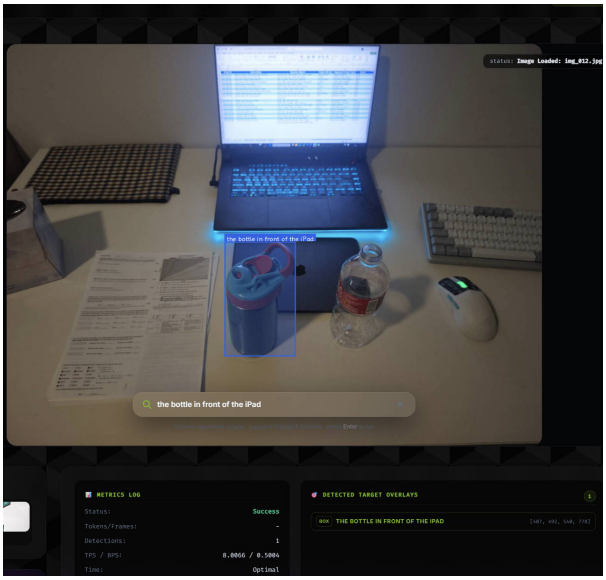
Failure cases are important because robotic systems need to know when perception is unreliable. In this mini-pilot, the main weak points were small or thin targets and ambiguous prompts. Spatial language also creates risk when multiple objects could satisfy the same relation.

Table 4: Failure case categories.

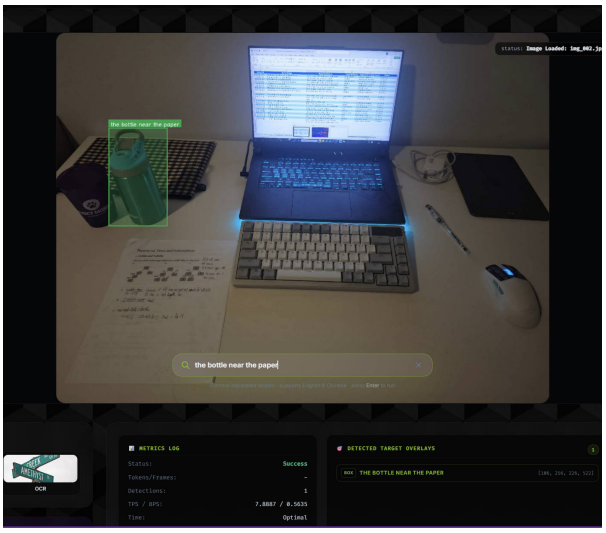
Failure Type	Description
Object ambiguity	Multiple objects match the prompt.
Occlusion	The target object is partially hidden.
Spatial confusion	The model selects the wrong object or reference relation.
Small object size	The target occupies a small region of the image.
Background distraction	Visually salient non-target objects attract the prediction.



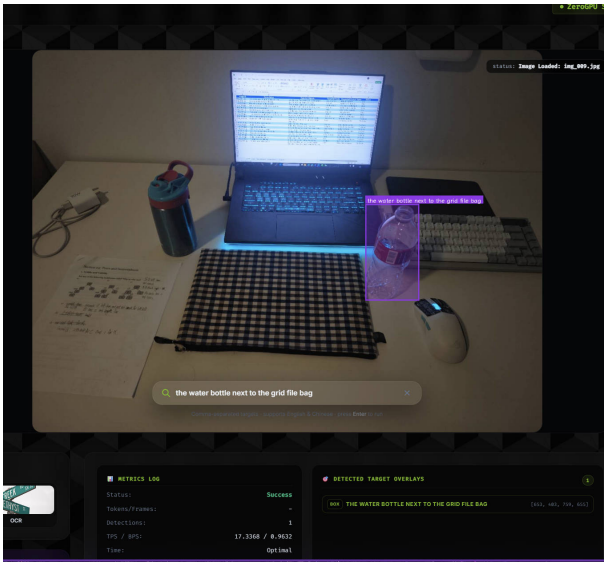
(a) Clear target region selected from a spatial prompt.



(b) Bottle localized from a relation prompt.

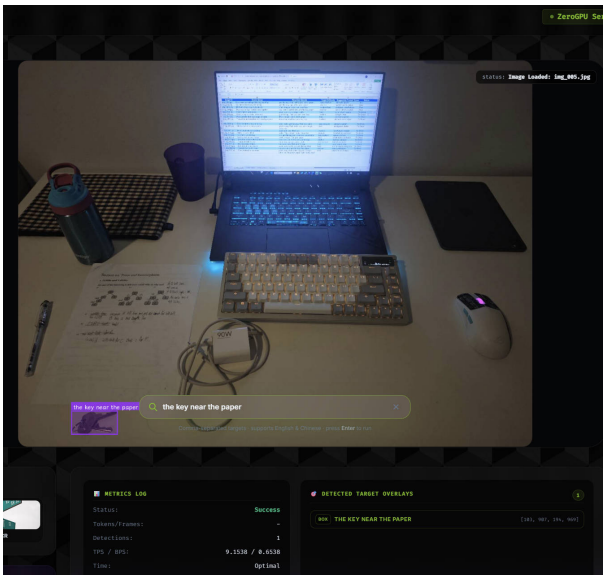


(c) Bottle near paper.

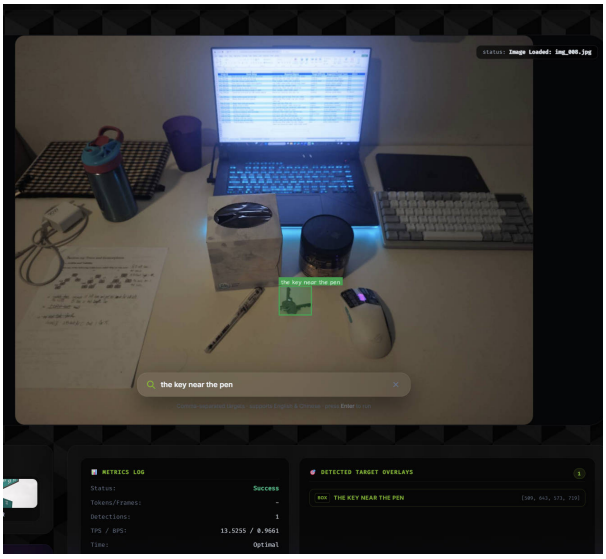


(d) Water bottle next to grid file bag.

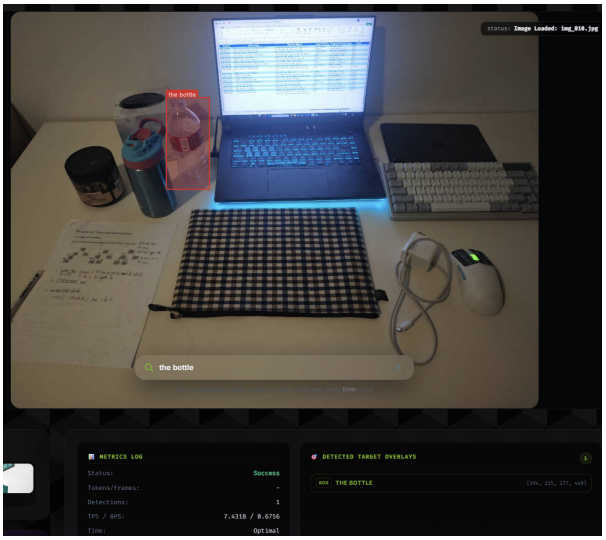
Figure 2: Representative successful LocateAnything outputs from the mini-pilot dataset.



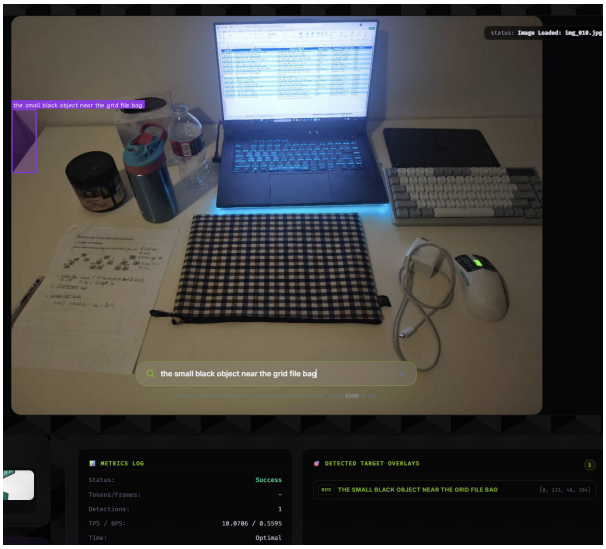
(a) Small-object success: key near paper.



(b) Small-object success: key near pen.



(c) Ambiguous prompt: "the bottle".



(d) Ambiguous small black object near the grid file bag.

Figure 3: Representative small-object and ambiguous-prompt examples.

7 Discussion

The results suggest that LocateAnything can often localize clearly described tabletop targets, especially for spatial-relation and similar-object prompts in this small dataset. This is encouraging for language-guided pre-grasp perception because many user commands naturally describe objects by nearby references.

At the same time, the weaker performance on small-object and ambiguous prompts shows that 2D visual grounding alone is not enough for robust manipulation. A real robot would need additional checks, such as depth sensing, segmentation, confidence estimation, or clarification dialogue when the instruction is unclear. Spatial language remains important for language-conditioned robot manipulation [7], but this mini-pilot does not prove robust spatial reasoning across all indoor scenes.

8 Limitations

- The dataset is small and collected from a limited tabletop/desk setup.
- The evaluation uses 2D image localization only.
- The study does not test physical grasping, robot control, or humanoid deployment.
- Ground-truth annotations were manually produced and may contain small labeling error.
- Inference latency is not reported because the web demo did not provide reliable per-query timing.

9 Conclusion

This report evaluated LocateAnything as a language-guided pre-grasp 2D localization module for humanoid robot perception. On a small tabletop mini-pilot dataset, LocateAnything achieved a mean IoU of 0.646, a median center-point error of 17.2 pixels, and a 76.9% success rate. Performance was strongest for spatial-relation and similar-object prompts, while small-object and ambiguous prompts were more challenging.

These results suggest that LocateAnything may be useful as a perception-stage component when the target is clearly described. However, the current study remains limited to a small tabletop dataset and does not evaluate full robotic grasping. Future work should expand the dataset, compare against additional baselines, integrate depth information, and test whether improved 2D localization helps real robot grasp planning.

AI-Assisted Writing Disclosure

The author used AI-assisted tools for language editing, formatting support, and LaTeX preparation. All research design, experiments, results, and final claims were reviewed and verified by the author.

Acknowledgments

No external acknowledgments are included in this version.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[3] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[4] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple open-vocabulary object detection with vision transformers," in *European Conference on Computer Vision*, 2022.

[5] S. Wang, S. Liu, Y. Kuang, X. Wei, Y. Liu, Z. Li, Y. Man, G. Chen, A. Tao, G. Liu, J. Kautz, L. Zhang, and Z. Yu, "LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding," *arXiv preprint arXiv:2605.27365*, 2026.

[6] G. Tziafas, Y. Xu, A. Goel, M. Kasaei, Z. Li, and H. Kasaei, "Language-guided robot grasping: CLIP-based referring grasp synthesis in clutter," *arXiv preprint arXiv:2311.05779*, 2023.

[7] Q. Luo, Y. Li, and Y. Wu, "Grounding object relations in language-conditioned robotic manipulation with semantic-spatial reasoning," *arXiv preprint arXiv:2303.17919*, 2023.