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

Dawei Cai
Department of Electrical and Computer Engineering
University of Washington
Seattle, WA, USA

***Abstract*—Humanoid robots need to identify task-relevant objects from natural-language instructions before any grasp planning can begin. This poster paper evaluates LocateAnything [1] as a language-guided pre-grasp 2D object localization module for 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 study is limited to the perception stage and does not test physical grasping or robot control.**

***Keywords*—vision-language grounding, object localization, humanoid robot perception, pre-grasp localization, LocateAnything**

I. Introduction

In everyday indoor settings, people rarely describe objects using only fixed category labels. A person may refer to "the cup on the left side of the keyboard," "the charger behind the tablet," or "the small black object near the paper." For a humanoid robot, this kind of instruction requires more than category detection. The robot first has to localize the intended object in the scene before any grasp or motion planning step can be considered. This problem is closely related to natural human-robot interaction in shared spaces [2].

Traditional object detectors such as YOLO can recognize predefined categories such as bottle, cup, laptop, phone, or chair [3]. They are useful when the target category is known in advance, but they are less flexible when the user describes an object through attributes, spatial relations, or context. In a cluttered tabletop scene, detecting all cups or bottles is not enough if the instruction refers to one specific object among several similar items.

Vision-language grounding models offer a possible way to connect these referring expressions to image regions. This direction builds on vision-language representation learning such as CLIP [4] and open-vocabulary grounding systems such as Grounding DINO and OWL-ViT [5], [6]. LocateAnything is a recent grounding model that predicts target regions from image–text inputs [1]. The current evaluation studies whether LocateAnything can support a narrow robotics perception task: pre-grasp 2D object localization from natural-language prompts in cluttered indoor or tabletop images.

The guiding question is: *Can LocateAnything accurately localize task-relevant objects from natural-language prompts in cluttered indoor scenes?* The study does not introduce a new model and does not claim to implement full robotic grasping. Instead, it reports a small mini-pilot evaluation of an existing vision-language grounding model for the perception step before grasp planning.

II. System Overview

The localization pipeline is:

Indoor Scene Image + Natural-Language Prompt $\rightarrow$ LocateAnything $\rightarrow$ Predicted Box / Point $\rightarrow$ Target Center Extraction $\rightarrow$ Ground Truth Comparison $\rightarrow$ Evaluation Metrics

For each image–prompt pair, LocateAnything returns a predicted target location [1]. When the output is a bounding box, it can be compared directly with the manually annotated ground-truth box. If the output is a point or mask, it can be converted into a target center or bounding region before evaluation.

The extracted target center is treated only as a 2D pre-grasp visual target. It is not a full grasp pose. A complete manipulation pipeline would still require depth sensing, 3D pose estimation, grasp planning, motion planning, and robot control. Although prior work has studied language-guided grasping and manipulation [7], this mini-pilot focuses only on whether the perception module can identify the correct object.

III. Pilot Experiment Design

The mini-pilot uses cluttered indoor and tabletop scenes collected from a desk setup. The final dataset contains 39 valid image–prompt pairs after excluding one prompt where no valid target object existed. The scenes include common tabletop items such as bottles, cups, phones, keyboards, papers, chargers, cables, pens, keys, and small containers.

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

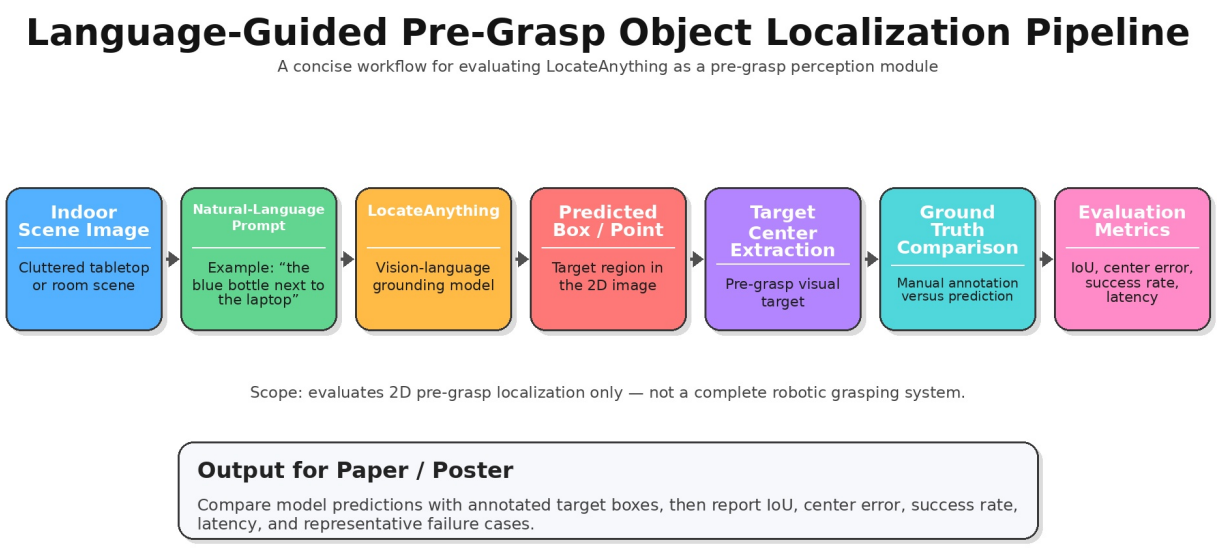

Figure 1: Proposed language-guided pre-grasp object localization pipeline. The system takes an indoor scene image and a natural-language prompt, uses LocateAnything to predict a 2D target region, extracts the target center, and compares the prediction with ground-truth annotations for evaluation.

prompts had no valid samples in the final dataset, so they are included in Table I only as a zero-sample category.

The evaluation procedure was:

1) Collect tabletop images with clutter and repeated objects.
2) Write one to two natural-language prompts per image.
3) Manually annotate the target bounding box for each prompt.
4) Run LocateAnything on every image–prompt pair.
5) Save the predicted box or point output.
6) Compare each prediction with the ground-truth annotation.
7) Compute IoU, center-point error, success rate, and qualitative failure cases.

IV. Evaluation Metrics

The evaluation uses IoU, center-point error, success rate, and qualitative failure analysis.

Intersection over Union (IoU). IoU measures overlap between the predicted bounding box and the ground-truth box. A prediction is counted as successful when IoU ≥ 0.5.

Center-point error. Center-point error is the pixel distance between the predicted target center and the ground-truth target center. This metric is useful for pre-grasp localization because a robot perception system often needs a 2D target center before estimating a possible 3D grasp point.

Success rate. Success rate is the percentage of prompts where the model localizes the correct target according to the IoU ≥ 0.5 criterion.

Failure analysis. Quantitative metrics are paired with a qualitative review of common errors. The review considers prompt ambiguity, similar objects, small or thin targets, partial occlusion, spatial relations, and background clutter.

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

V. Preliminary Results and Failure Analysis

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.

Performance varied by prompt type. Spatial-relation prompts had 16 valid samples and reached 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. These results suggest that LocateAnything can often localize clearly described tabletop targets when the prompt provides enough visual or relational information.

The harder cases were small-object and ambiguous prompts. Small-object prompts had 12 valid samples and a 58.3% success rate, while ambiguous prompts had 3 valid samples and a 33.3% success rate. Occlusion prompts achieved 75.0% success across 4 valid samples. Attribute prompts had no samples in the final valid dataset, so no measured conclusion is reported for that category.

The qualitative examples in Fig. 2 are consistent with the numerical results. LocateAnything can localize some clear spatial and small-object targets, but errors appear when the object is small, visually weak, or ambiguous among nearby objects. Spatial language is important for language-conditioned

Table I: Quantitative Results by Prompt Type

Prompt Type	N	Mean IoU	Median Center Error (px)	Success Rate (%)	Main Failure Mode
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

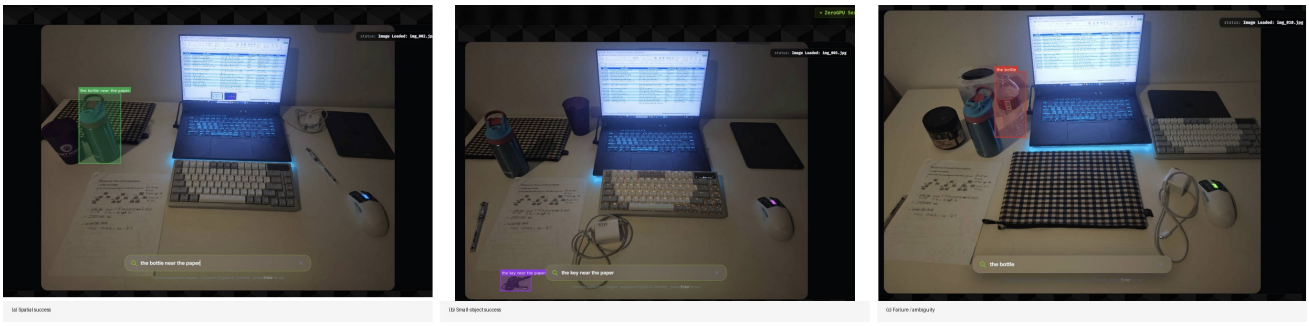

Figure 2: Representative LocateAnything outputs from the mini-pilot dataset. The examples show successful localization for clear prompts and more difficult cases involving small objects, ambiguity, or clutter.

robot manipulation [8], but this mini-pilot should not be interpreted as evidence of robust spatial reasoning across all indoor scenes.

VI. Limitations and Future Work

The current evaluation has several limitations. First, the images were collected from a limited tabletop and desk setup. The results should therefore be interpreted as an early feasibility test rather than a general benchmark for humanoid robot perception.

Second, the setup only evaluates 2D localization. A bounding box or center point can identify the target in the image, but it is not a complete 3D grasp pose. Full robotic grasping would require depth sensing, object pose estimation, grasp planning, motion planning, and closed-loop control.

Third, the method has not been tested on a full humanoid robot. Even when the object is localized correctly in 2D, a physical robot may still fail during grasp execution. For this reason, the study should be understood as a pre-grasp localization evaluation, not a complete manipulation system.

Future work should expand the dataset, compare LocateAnything with baselines such as YOLO, Grounding DINO, or OWL-ViT [3], [5], [6], add RGB-D camera input, connect the 2D output to a robot manipulation pipeline, and test whether better 2D localization improves real robot grasp planning.

VII. Conclusion

This paper presented a mini-pilot evaluation of LocateAnything for language-guided pre-grasp 2D object localization. On a small tabletop dataset, LocateAnything showed promising preliminary performance for spatial-relation and similar-object prompts. 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. The study remains limited to a small tabletop dataset and does not evaluate physical grasping or robot control. Future work should test larger and more varied scenes, compare against additional baselines, integrate depth information, and evaluate 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.

References

[1] 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.

[2] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human-robot interaction," *Foundations and Trends in Robotics*, vol. 5, no. 4, pp. 261–349, 2017.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.