2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Grid-Based Marking Prompt Framework for Spatial Understanding in Vision–Language Models

Ryo Terashima[1], Yuga Yano[1], Koshun Arimura[1] and Hakaru Tamukoh[1,2]

***Abstract*— In this study, we propose a grid-based marking prompt framework to enhance spatial understanding in vision-language models (VLMs). The framework integrates object detection, background masking, and number overlaying to enable VLMs to interpret spatial and contextual instructions more effectively. By inputting numbered images along with natural language instructions, a VLM selects the number corresponding to the most semantically appropriate location. The framework operates without requiring prior information such as 3D models or physical markers. Moreover, the proposed framework allows flexible rule adaptation through prompt engineering alone, providing general applicability across various objects and environments. We conducted two experiments for the object placement task. In experiment 1, shelf images captured by a service robot were used to evaluate the placement selection accuracy of a VLM. In experiment 2, the framework was implemented on a service robot and conducted the object placement task at positions selected by a VLM in a real-world environment. The framework achieved a high success rate in both experiments, demonstrating the effectiveness and practical utility of the framework in real-world environments.**

I. Introduction

In recent years, service robots have been used across various fields, including restaurants [1], [2], [3], hospitals [4], and homes [5]. To operate appropriately in the real-world environment, service robots must be capable of spatial understanding and estimating suitable actions [6], [7], [8]. Vision-language models (VLMs) [9], [10], [11], [12], [13], [14] have emerged as a promising approach to enabling such spatial understanding. Therefore, VLMs are increasingly being considered for integration into service robots to enhance their spatial understanding. However, VLMs still have issues with accurate spatial understanding [15], [16], [17].

To address this issue, Set-of-Mark (SoM) [18] was proposed as a prompting method to enhance the spatial understanding capabilities of VLMs. SoM overlays numbered on objects in images, allowing VLMs to reference and reason about the spatial relationships of objects using overlaying numbers. For example, a VLM using SoM can respond: "To the right of laptop number 9 is a lamp, which is number 12." However, SoM cannot mark spatial areas, making SoM unsuitable for tasks that target spatial areas. To address this issue, we propose a new framework to enhance the spatial understanding of VLMs.

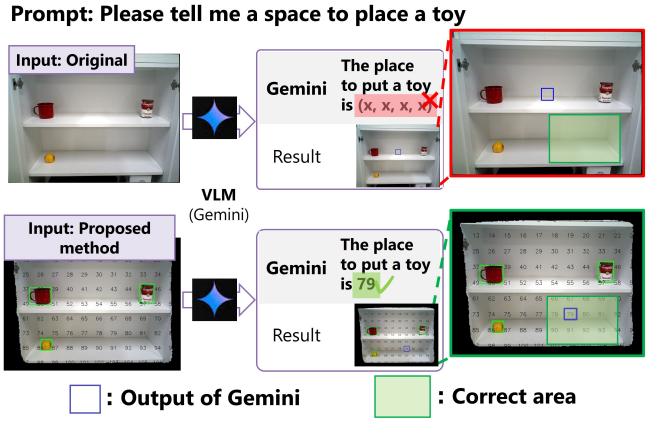

Fig. 1: Comparison of input formats for estimating empty spaces using a VLM (Gemini). **Upper**: Gemini directly estimates empty spaces using bounding boxes based on the original image input. **Bottom**: Gemini selects the place of empty spaces from the numbers in the image using the proposed method. In both cases, the bounding box size is 40 x 40 pixels, as indicated in the prompt.

Figure 1 shows an overview of the proposed framework. The framework consists of three main components: object detection, background masking, and number overlaying. The main feature of the proposed framework is the overlaying of numbers on the space, not only objects, as in SoM. This enables the framework to handle spaces without objects.

To evaluate the effectiveness of the proposed framework, we conducted experiments focusing on an object placement task. In this task, the VLM must recognize objects in an image and estimate appropriate empty spaces for placing a target object, based on the spatial arrangement and categories of the existing objects. The object placement task requires spatial understanding that incorporates semantic appropriateness. Therefore, the object placement task is well-suited for evaluating the spatial understanding of VLMs [19], [20].

We implemented the proposed framework on a service robot and evaluated its effectiveness in enhancing spatial understanding through an object placement task on a shelf. This study aims to enhance the spatial understanding of VLMs, which is important for enabling service robots to operate appropriately in the real-world environment.

This paper is based on results obtained from project JPNP16007 commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by JSPS KAKENHI Grant Numbers 23H03468 and 23K18495. This work was supported JST ALCA-Next Grant Number JPMJAN23F3. This work was supported by JST SPRING, Japan Grant Number JPMJSP2154.

All authors are with Kyushu Institute of Technology, Fukuoka, Japan. {terashima.ryo631, yano.yuuga158, arimura.koshun523}@mail.kyutech.jp and tamukoh@brain.kyutech.ac.jp

Hakaru Tamukoh is also affiliated with the Research Center for Neuromorphic AI Hardware, Fukuoka Japan.

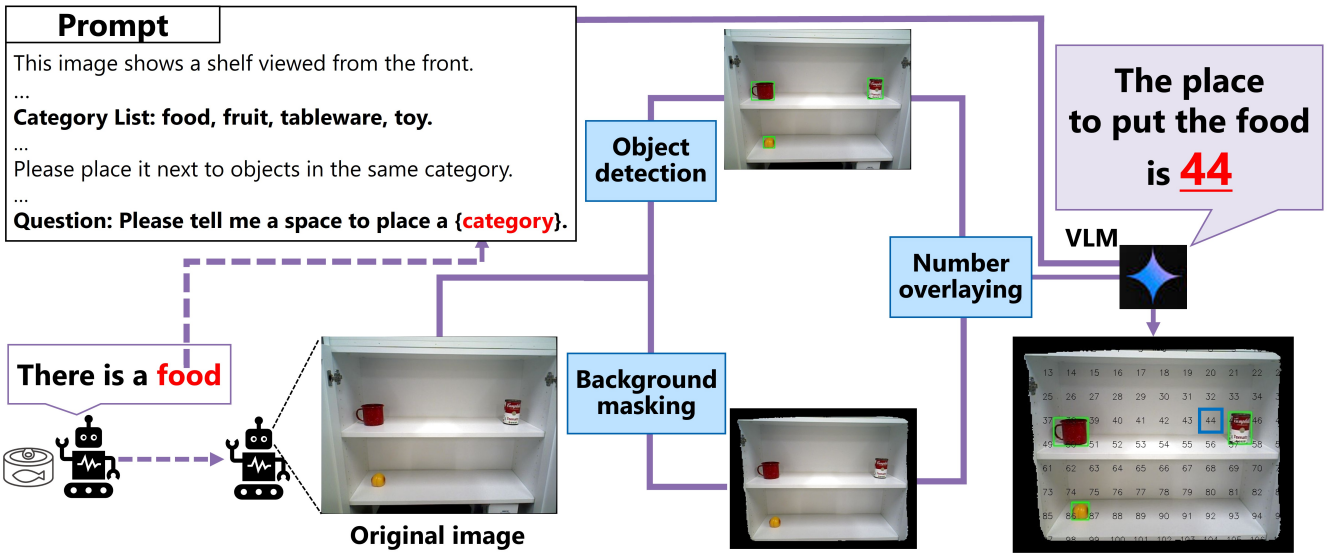

Fig. 2: Overview of the proposed framework

II. Related work

SoM [18] is a prompting method designed to assist VLMs in interpreting textual instructions and spatial relationships by marking objects in images. SoM uses a segmentation model to identify individual objects and overlays a unique number on each object. Using these numbered images, VLMs such as GPT-4V [10] can reference specific objects, enabling more precise spatial understanding of VLMs. For example, when asked, "What is to the left of the laptop on the right?" a VLM using SoM can correctly respond, "It is lamp number 12 to the left of laptop number 9." SoM shows high performance on referring expression tasks like RefCOCOg [21], even without additional training.

SoM is a method specialized for improving the precision of target object selection in VLMs. SoM clarifies the target candidates by overlaying numbers on each object in the image, helping VLMs identify the target object and improving interpretability and consistency. However, the object placement task in this study focuses on selecting semantically appropriate empty spaces in the image. Therefore, applying the SoM framework to enhance VLMs spatial understanding is difficult in the object placement task. To address this issue, we propose a grid-based marking prompt framework that focuses on estimating and reasoning about empty spaces.

III. Proposal

In this study, we propose a grid-based marking prompt framework to enhance the spatial understanding of VLMs. Figure 2 shows an overview of the proposed framework. The framework overlays numbers at equal spatial intervals across the image, allowing the VLM to select the appropriate number for a given instruction. To further enhance spatial understanding of VLMs, the framework incorporates background masking and adds bounding boxes (BBoxes) to indicate the positions of detected objects. The proposed framework consists of the following three components:

- **Background masking**: extraction of target areas (shelves, etc.) using a VLM.
- **Object detection**: detection of existing object locations using an object detection model.
- **Number overlaying**: divide the space into equal intervals, overlay numbers, and allow a VLM to select the placement location.

A. Background masking

The proposed framework extracts only the target area from the input image. Input images often contain irrelevant or distracting information, which can negatively affect the spatial understanding and reasoning of VLMs. To ensure robustness in real-world environments, the framework must handle diverse backgrounds and objects. Therefore, we utilize a zero-shot segmentation model.

Several zero-shot models for segmentation have been proposed, such as Nanosam [22] and Detic [23]. However, these methods often struggle to accurately identify the target area when elements like shelf doors or outer frames are included. To address these issues, we utilize Gemini [12]. Gemini is a VLM capable of segmenting areas based on natural language instructions ("e.g., detect a shelf without including doors"), while considering the structural features of objects.

B. Object detection

SoM has shown that prompting to indicate positions of objects is effective in enhancing the spatial understanding of VLMs. Based on this finding, we incorporate a component into the proposed framework that draws BBoxes of objects in the image. In this study, we use a light-weight zero-shot detection model by combining NanoSAM [22] with Grounding DiNO [24].

We input multiple prompts, such as [food, fruit, toy, tableware], as supercategories into NanoSAM to detect var-

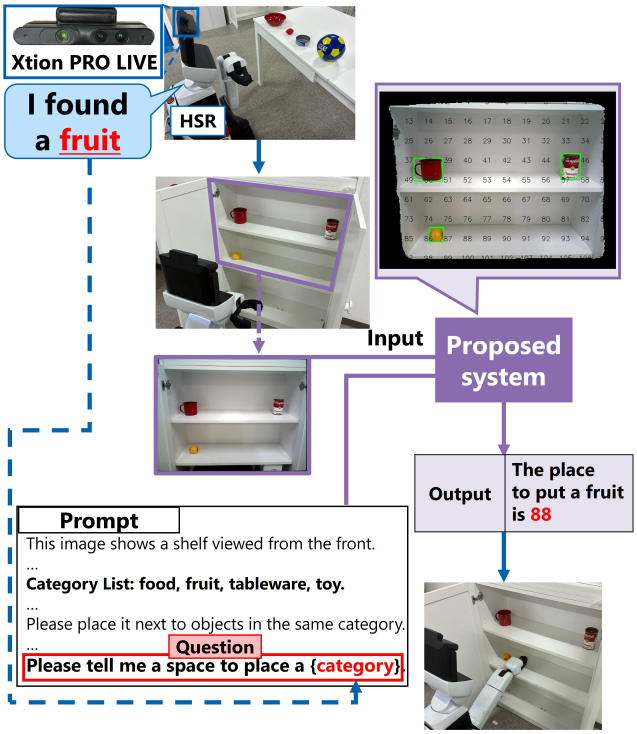

Fig. 3: Flow of experiments conducted. **Purple arrow**: Flow of Experiment 1. **Purple + blue arrow**: Flow of Experiment 2.

ious object categories in the image without prior training. However, NanoSAM often misclassifies object categories. To address this, we use only the BBox coordinates as input to the VLM, without classification labels. When the same object is detected under multiple supercategories, we retain only the detection with the highest confidence score, ensuring that a single BBox represents each object.

C. Number overlaying

To enhance the spatial context understanding of VLMs for object placement, we overlay numbers onto the input image. First, the target area is divided into a grid with equal intervals, and unique numbers are overlaid on each cell. The numbered image, along with a natural language instruction, is input into the VLM. This setup allows the VLM to select the appropriate number based on the semantic context. For example, given an instruction like "Place the cup near the other cups", the VLM can determine a semantically appropriate location from among the numbered grid cells in the image.

IV. Experimental Settings

The following section explains two experiments conducted to evaluate the effectiveness of the proposed framework. Figure 3 shows the experimental workflow.

In Experiment 1, we estimate empty spaces using the proposed framework and evaluate both estimation accuracy and processing time. Input images of VLMs were captured using a head camera(resolution 640×480) of Human Support Robot (HSR) [25]. In Experiment 2, we perform the complete task sequence, from object detection to placement, using HSR. In this experiment, we evaluate the practical effectiveness of the proposed framework in a real-world environment.

Fig. 4: Four scenes composed of YCB objects and two types of shelves used in experiments.

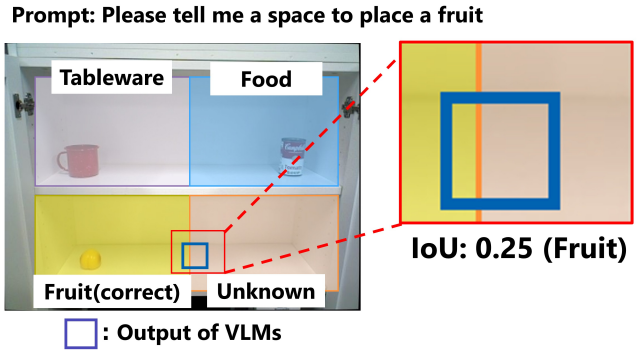

Fig. 5: Example of VLM output when targeting the fruit category in the prompt. The image compares the predicted location (blue box) generated by the VLM with the ground-truth location (yellow area) manually annotated as the appropriate location for fruit.

A. Experiment 1: Estimation of empty space

Figure 4 shows scenes used in the experiments. We manually divided the shelf into four areas, with category labels such as food, tableware, fruit, and unknown, for each area. We set the unknown area as the candidate for placing categories that do not belong on a shelf in the experiments.

For the object placement tasks, we used Yale-CMU-Berkeley(YCB) object set [26], a widely used benchmark for robot grasping tasks. YCB objects consist of various categories, such as food containers, tools, and daily goods, making YCB objects well-suited for simulating home environments. Figure 5 shows an example of the Intersection over Union(IoU) calculation. In this study, we evaluate the overlap between the 40×40 pixel region P estimated by

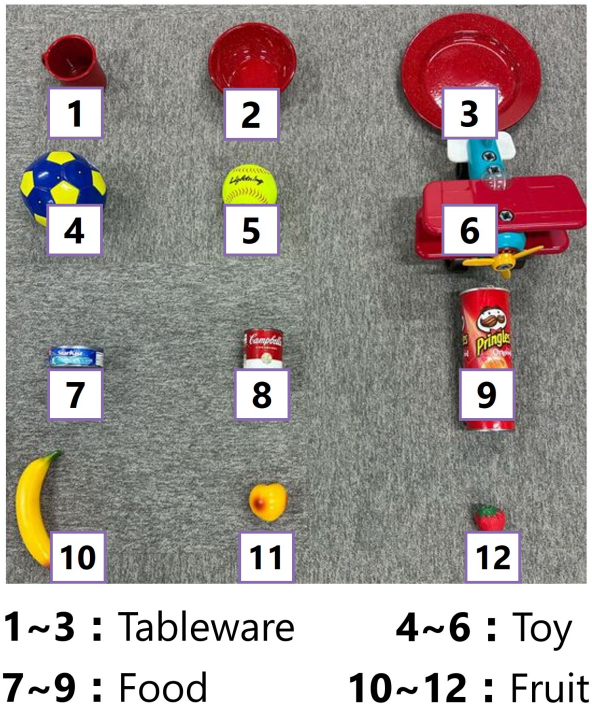

Fig. 6: List of objects used in the experiment.

the VLM and the manually defined GT region G, using IoU defined as:

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|}. \tag{1}$$

As shown in Equation 2, the mean IoU (mIoU) is the average IoU score over N trials for each scene. The mIoU serves as the metric to evaluate and compare the accuracy of the estimations of VLMs.

$$\mathrm{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IoU}_i. \tag{2}$$

Furthermore, the processing time for each component of the proposed framework is recorded to evaluate the overall practicality of the framework.

B. Experiment 2: Experiment in a real-world environment using HSR

In Experiment 2, we implemented the proposed framework on HSR and performed a series of object placement tasks on a shelf. The experiment aimed to confirm whether placement decisions based on spatial understanding could be executed in the real-world environment and to investigate the effect of any incorrect decisions.

Figure 6 shows the 12 objects used in this experiment, grouped in four categories. For each trial, we randomly selected three categories from the four categories and placed one object from each category on a shelf. We repeated this procedure four times to create four different placement patterns. In total, 16 object placement tasks were conducted,

TABLE I: Comparison of miou and average(avg.) processing time

Method	mIoU	Avg. processing time (s)
w/o marked (GPT-4o-latest)	0.500	7.091
w/o marked (Gemini-2.0-flash)	0.500	**2.881**
w/ marked (GPT-4o-latest)	0.625	7.542
w/ marked (Gemini-2.0-flash)	**0.719**	5.237

covering each category (tableware, toy, food, and fruit) across the different patterns. To evaluate the performance of empty space estimation, we assumed perfect accuracy in object detection and grasping. In this experiment, we define successful placement as follows:

- The object is placed within the area specified by the proposed framework.
- The object does not make contact with other objects.
- The object is stable after placement and does not fall.

V. Experimental Results

A. Experiment 1: Estimation of empty space

First, we evaluated the effectiveness of number overlaying, one of the key components of the proposed framework. We performed the empty space estimation task using images both with and without numbers as input. We used two types of VLMs and compared GPT [9] and Gemini [12].

Table I shows the accuracy and average processing time for each condition. When using images without number overlays, the mIoU was 50.0% for both GPT and Gemini. With the application of the proposed number overlaying, the mIoU increased to 62.5% for GPT and 71.9% for Gemini, showing improvements of 12.5 points and 21.9 points, respectively. These results indicate that number overlaying improves the spatial understanding performance of VLMs. Regarding processing time, the average was 7.091 seconds for GPT and 2.881 seconds for Gemini when using images without number overlays. With number overlaying applied, the average processing time increased to 7.542 seconds for GPT and 5.237 seconds for Gemini. Overall, Gemini outperformed GPT in both accuracy and processing time.

Based on these results, Gemini showed higher accuracy than GPT, confirming the effectiveness of number overlaying. Accordingly, we selected Gemini as the VLM for number estimation in the subsequent experiments.

Additionally, we evaluated empty space estimation using the proposed framework. The framework achieved an mIoU of 85.2% with an average processing time of 9.753 seconds. In the following section, we conduct an ablation study by individually removing the following components: (1) background masking and (2) object detection, to analyze their respective contributions to estimation accuracy and processing time.

B. Experiment 1: Ablation Study

Based on the results of the preliminary experiments, we conducted an ablation study using number overlaying and Gemini. We evaluated different combinations of the proposed

TABLE II: Ablation study of each component in the proposed system.

Components	Object detection	Background masking	IoU	Avg. processing time (s)
Object detection	✓		0.709	**5.774**
Background masking		✓	0.800	8.692
Proposed(Object detection + Background masking)	✓	✓	**0.852**	9.753

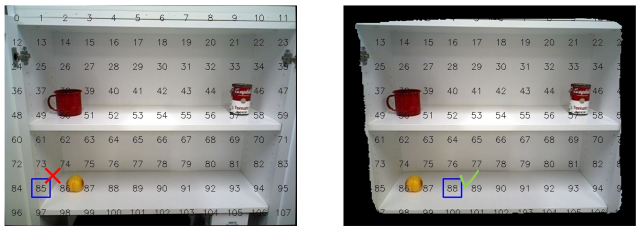

Fig. 7: Example of output VLM with and without a background mask.

framework by enabling or disabling its components, object detection, and background masking, to investigate the effect of each on spatial understanding.

Table II shows the experimental results of the ablation study. The framework, utilizing all proposed components, achieved the highest IoU of 85.2%. Background masking improved the accuracy of identifying placement areas by excluding irrelevant regions, thereby reducing errors such as selecting locations outside the shelf. Object detection contributed by providing the positions of existing objects, allowing the VLM to identify and avoid already occupied areas. In this way, Object detection reduces misestimation of VLMs during empty space selection. These results indicate that combining background masking, object detection, and number overlaying enables semantically valid and accurate empty space estimation. The fastest processing time, 5.774 seconds, was achieved by using only object detection.

C. Experiment 2: Experiment in a real-world environment using HSR

In this experiment, we implemented the proposed framework on HSR and performed object placement tasks on a shelf. We randomly selected three categories from the four categories and placed one object from each category on the shelf. Across four different placement patterns, a total of 16 object placement tasks were performed.

In 9 out of 16 placement tasks, HSR stably placed objects in the correct areas. Of the 7 failures, four were due to interference with existing objects on the shelf, and three occurred because objects were placed next to objects of a different category. Notably, no objects fell from the shelf in any trial. These results indicate the effectiveness of the proposed framework in real-world environments.

VI. Discussion

A. Effectiveness of object detection

BBoxes generated by object detection allow the VLM to identify the positions of existing objects in the image. By using these BBOX coordinates, the VLM can recognize areas occupied by objects as places to avoid. Consequently, the VLM selects empty spaces that do not overlap with existing objects. This indicates that object detection contributes to improving empty space estimation accuracy by preventing placement in occupied areas and allowing the selection of semantically appropriate spaces.

B. Effectiveness of Background masking

Figure 7 shows examples with and without background masking. Without background masking, the VLM occasionally selected areas outside the shelf. In contrast, applying background masking effectively guided the VLM to select areas within the shelf. By removing unnecessary background information, the masking process enables the VLM to identify semantically appropriate and stable locations more easily.

Additionally, background masking can be applied not only to shelves but also to other spatial structures such as desks and storage spaces. Therefore, the proposed framework has the potential to serve as a general-purpose tool for enhancing spatial understanding across various environments.

C. Reasons for failure in Experiment 2

We conducted the object placement task using HSR to evaluate the performance of the proposed framework in a real-world environment. Out of 16 autonomous placement trials, nine were successful, while seven failed.

The main factors contributing to failure are considered to be the following two. First, the proposed framework fixes the size of the object to be placed at 40×40 pixels as input to the VLM. Consequently, if the actual size of the target object exceeds this fixed size, there may not be sufficient space, causing objects to come into contact. To address this issue, we consider estimating the object size according to the actual dimensions of the target object.

Second, the low resolution (640×480) of the RGBD camera mounted on HSR may have caused the VLM to misidentify object categories [27], [28]. To address this, we implement a higher-resolution camera, and verifying its impact on reducing failures is necessary.

VII. Conclusions

In this study, we proposed a grid-based marking prompt framework to enhance the spatial understanding of VLMs. The framework consists of three components: (1) background masking, (2) object detection, and (3) number overlaying.

We evaluated the effectiveness of the framework through object placement tasks that require the integration of visual and linguistic processing. Experimental results showed that the proposed framework enhances VLMs in estimating empty spaces for semantically appropriate object placement. The ablation study showed that combining background masking, object detection, and number overlaying achieved the highest mIoU score of 0.852. Furthermore, experiments conducted with HSR showed that the framework enables semantically appropriate object placement in a real-world environment.

A key feature of the proposed framework is an extension of the SoM approach by expanding the number of overlays from object areas to include spatial areas without objects. This feature enables the framework to apply to tasks that require spatial decisions, where conventional methods are difficult to apply. Furthermore, the proposed framework can be applied to other spatial understanding tasks through prompt modification.

In future work, we plan to address more complex and ambiguous natural language instructions, such as "place it in a sunny spot" Additionally, the real-world experiments revealed two key challenges: the risk of object contact due to fixed object size, and reduced object identification accuracy caused by limitations in camera resolution.

REFERENCES

[1] Y. Yano, A. Mizutani, Y. Fukuda, D. Kanaoka, T. Ono, and H. Tamukoh, "Unified understanding of environment, task, and human for human-robot interaction in real-world environments," in *proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 224–230, 2024.

[2] Y. Yano, K. Isomoto, T. Ono, and H. Tamukoh, "Autonomous Waiter Robot System for Recognizing Customers, Taking Orders, and Serving Food," *RoboCup 2023: Robot World Cup XXVI*, pp. 252–261, 2024.

[3] K. Isomoto, S. Fumoto, R. Kobayashi, Y. Tanaka, and H. Tamukoh, "Lightweight hand-waving action recognition using reservoir computing in a cafeteria environment*," in *2025 IEEE/SICE International Symposium on System Integration (SII)*, pp. 1034–1040, 2025.

[4] J. Holland, L. Kingston, C. McCarthy, E. Armstrong, P. O ' Dwyer, F. Merz, and M. McConnell, "Service robots in the healthcare sector," *Robotics*, vol. 10, no. 1, p. 47, 2021.

[5] T. Ono, D. Kanaoka, T. Shiba, S. Tokuno, Y. Yano, A. Mizutani, I. Matsumoto, H. Amano, and H. Tamukoh, "Solution of World Robot Challenge 2020 Partner Robot Challenge (Real Space)," *Advanced Robotics*, vol. 36, no. 17, pp. 870–889, 2022.

[6] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, "Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics," in *Proceedings of the Computer Vision and Pattern Recognition Conference(CVPR)*, pp. 15768–15780, 2025.

[7] K. Yamao, D. Kanaoka, K. Isomoto, and H. Tamukoh, "Object positions interpretation system for service robots through targeted object marking," in *Proceedings of the 2025 IEEE/SICE International Symposium on System Integration (SII)*, pp. 998–1003, 2025.

[8] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, D. Kalashnikov, S. Levine, Y. Lu, C. Parada, K. Rao, P. Sermanet, A. T. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, M. Yan, N. Brown, M. Ahn, O. Cortes, N. Sievers, C. Tan, S. Xu, D. Reyes, J. Rettinghouse, J. Quiambao, P. Pastor, L. Luu, K.-H. Lee, Y. Kuang, S. Jesmonth, N. J. Joshi, K. Jeffrey, R. J. Ruano, J. Hsu, K. Gopalakrishnan, B. David, A. Zeng, and C. K. Fu, "Do As I Can, Not As I Say: Grounding language in robotic affordances," in *Proceedings of The 6th Conference on Robot Learning(CoRL)*, pp. 287–318, 2023.

[9] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[10] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, 2023.

[11] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2025.

[12] L. Team, A. Modi, A. S. Veerubhotla, A. Rysbek, A. Huber, A. Anand, A. Bhoopchand, B. Wiltshire, D. Gillick, D. Kasenberg, *et al.*, "Evaluating gemini in an arena for learning," *arXiv preprint arXiv:2505.24477*, 2025.

[13] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, *et al.*, "Open-world object manipulation using pre-trained vision-language models," in *in Proceedings of the 7th Conference on Robot Learning(CoLR)*, pp. 3397–3417, 2023.

[14] T. Xiao, H. Chan, P. Sermanet, A. Wahid, A. Brohan, K. Hausman, S. Levine, and J. Tompson, "Robotic skill acquistion via instruction augmentation with vision-language models," in *Proceedings of Robotics: Science and Systems(RSS)*, 2023.

[15] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," in *Proceedings of The 12th International Conference on Learning Representations*, 2024.

[16] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision-language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 135062–135093, 2024.

[17] Z. Wang, S. Zhou, S. He, H. Huang, L. Yang, Z. Zhang, X. Cheng, S. Ji, T. Jin, H. Zhao, and Z. Zhao, "Spatialclip: Learning 3d-aware image representations from spatially discriminative language," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29656–29666, 2025.

[18] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.

[19] J. Lee, S. Park, J. Park, K. Lee, and S. Choi, "Spots: Stable placement of objects with reasoning in semi-autonomous teleoperation systems," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17786–17792, 2024.

[20] I. Kapelyukh, Y. Ren, I. Alzugaray, and E. Johns, "Dream2Real: Zero-shot 3D object rearrangement with vision-language models," in *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4796–4803, 2024.

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

[22] NVIDIA-AI-IOT, "NanoSAM," 2023. https://github.com/NVIDIA-AI-IOT/nanosam (accessed 10. Aug. 2025).

[23] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, pp. 350–368, Springer, 2022.

[24] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Proceedings of the 18th European conference on computer vision*, pp. 38–55, 2024.

[25] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH journal*, vol. 6, no. 4, 2019.

[26] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.

[27] Y. Hao, H. Pei, Y. Lyu, Z. Yuan, J.-R. Rizzo, Y. Wang, and Y. Fang, "Understanding the impact of image quality and distance of objects to object detection performance," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11436–11442, 2023.

[28] P. K. A. Vasu, F. Faghri, C.-L. Li, C. Koc, N. True, A. Antony, G. Santhanam, J. Gabriel, P. Grasch, O. Tuzel, *et al.*, "Fastvlm: Efficient vision encoding for vision language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference(CVPR)*, pp. 19769–19780, 2025.