2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Object Positions Interpretation System for Service Robots Through Targeted Object Marking*

Kosei Yamao[1], Daiju Kanaoka[1], Kosei Isomoto[1] and Hakaru Tamukoh[1,2]

***Abstract*— Service robots are typically required to interpret and execute various complex tasks in home environments. Recognizing the environment, such as furniture, and understanding the relationships between object positions is critical for executing various tasks. Set of mark (SoM) is a visual prompting method that focuses on interpreting the relationship between semantic regions by overlaying marks in each region. However, SoM marks segmented regions that are not objects such as walls and floors. This marking creates noise when interpreting object positions. To address this problem, we propose a novel object-position interpretation system that combines an object detection model and a vision-language model (VLM). The proposed system incorporates an object detection model to mark only objects, allowing the VLM to efficiently interpret object positions. Furthermore, the proposed system improves the accuracy of the system by including the original image and label output by the object detection model in the input to the VLM. The experimental results show that the proposed system outperforms SoM in terms of interpreting object positions.**

I. Introduction

An aging population and low birth rates have considerably increased the demand for service robots [1]. Recently, various studies on service robots have been actively conducted [2]–[4]. Service robots are required to interpret complex commands from people and execute the desired tasks in real-time. For instance, when a robot receives a command "Bring me the item behind the lemon," the robot should understand the context of "behind" in real-world space and execute the task accurately. This study proposes a novel method for interpreting the positional relationships of objects by using a vision language model (VLM).

Yang et al. proposed a set of mark (SoM), which is a visual prompting method that improves the accuracy of vision and multimodal tasks using VLM [5]. SoM enhances the image interpretation performance of VLM by overlaying marks, such as numbers, on each region recognized through semantic segmentation methods. However, SoM also marks regions such as walls and floors. These marks are noise in the interpretation of the object-position relationship. We believe that this problem can be solved by marking only objects.

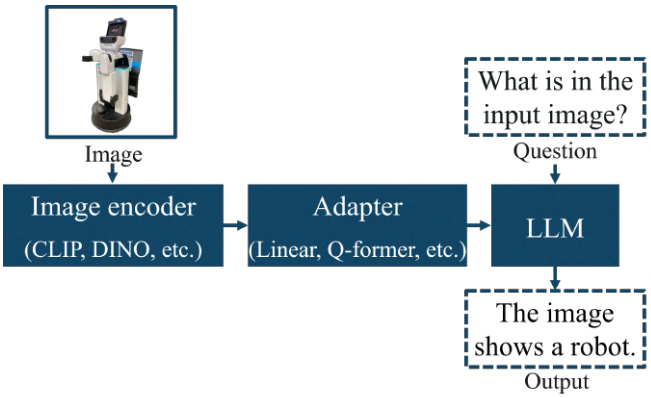

Fig. 1. Schematic of the VLM

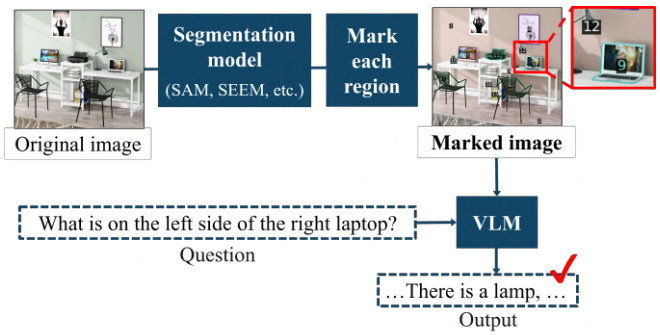

Fig. 2. **Schematic of the SoM.** SoM, a visual prompting method proposed, improves VLM performance by overlaying marks such as numbers and alphabetic letters on images.

In this study, we propose a novel system that combines object detection models and VLM to focus on object positions. This system enhances accuracy by overlaying marks only on objects by using an object detection model. The contributions of this study are as follows:

- We propose a novel object position interpretation system using object detection models and VLM.
- We confirmed that the proposed system can interpret object positions more accurately than the existing method in the experiment.

*This research is based on results obtained from a 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 as well as from JST ALCA Next Grant Number JPMJAN23F3.

[1]All authors are with Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan. {yamao.kosei665, kanaoka.daiju334, isomoto.kosei778}@mail.kyutech.jp, tamukoh@brain.kyutech.jp

[2]Hakaru Tamukoh is with Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196, Japan

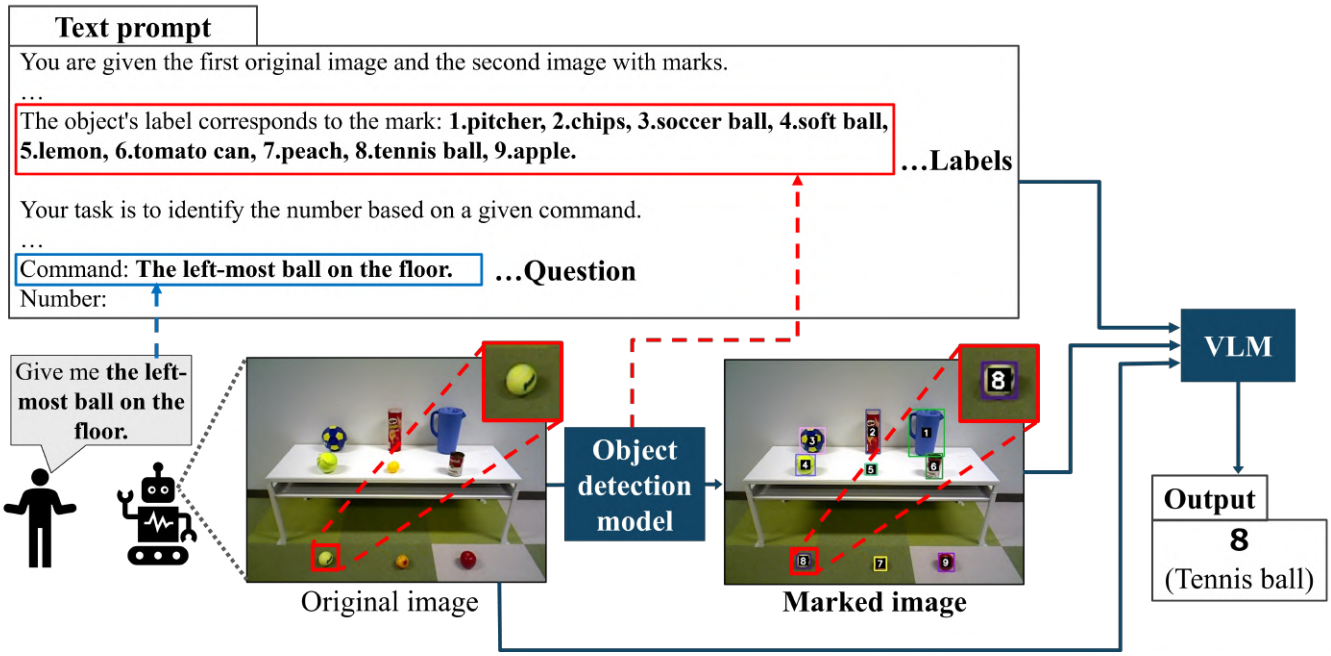

Fig. 3. **Schematic of the proposed object-position interpretation system.** The proposed system overlays marks only on the object, allowing the VLM to focus on the target objects.

II. Related Works

A. Vision-language model

A VLM is a machine learning model that integrates text and visual information and can solve tasks such as image captioning, image generation, and visual question answering (VQA). Fig. 1 shows the schematic of a VLM with VQA. First, the feature vector of an input image is calculated using an image encoder, such as contrastive language-image pretraining (CLIP) [6] or self-distillation with no labels (DINO) [7]. Next, the feature vector is inputted into an adapter such as a querying transformer (Q-former) [8] to transform it into a format suitable for a large language model (LLM). The transformed feature vector and feature vector of a question are inputted into an LLM, and the answer to the question is the output.

Numerous VLMs, such as GPT [9], [10], Gemini [11], [12] and LLaVA [13], [14] have been proposed. Their general versatility has been demonstrated experimentally and has attracted considerable attention [15], [16]. Studies in robotics have focused on integrating VLM-based systems into robots for accurate environmental and object recognition [17], [18].

B. SoM

VLMs have problems with visual tasks such as interpreting position relationships [19]. Various prompting methods were proposed to improve the performance of VLMs [5], [20], [21]. Numerous methods have been devised for enhancing the input text [22], [23] and visual prompting methods for designing input images [5], [20], [21].

SoM is a visual prompting method proposed by Yang et al. and is used to improve the performance of a VLM by overlaying marks such as numbers on images. Fig. 2 presents a schematic of the SoM. SoM uses segmentation models, such as the segment anything model (SAM) [24] and the segment everywhere all at once model (SEEM) [25] to segment an image into semantic regions. The VLM can then efficiently interpret the relationships between regions by overlaying the marks. SoM has been reported to improve the accuracy of VQA, particularly in experiments using GPT-4 [9].

However, SoM overlays marks on segmented regions that are not objects, such as walls and floors, among others. This phenomenon creates noise when interpreting the spatial relationships between objects. A solution to this problem is marking only objects instead of all the regions in the image.

III. Proposed System

In this study, we propose a novel object-position interpretation system using the VLM and an object detection model. Fig. 3 presents a schematic of the proposed system, which processes through the following 4 steps:

1) Object Detection: The system uses an object detection model to identify objects in an input image. Detected objects are referred to as ”target objects.” The use of an object detection model ensures that only objects, rather than non-object regions like walls or floors, are identified, thereby reducing noise.
2) Generating Marked Images: The system overlays numbers and bounding boxes on the detected target objects to create marked images. This approach improves the problem of SoM overlaying marks on segmented regions that are not objects, such as walls and floors.
3) Generating Text Prompts: The text prompt provided to the VLM consists of three parts:

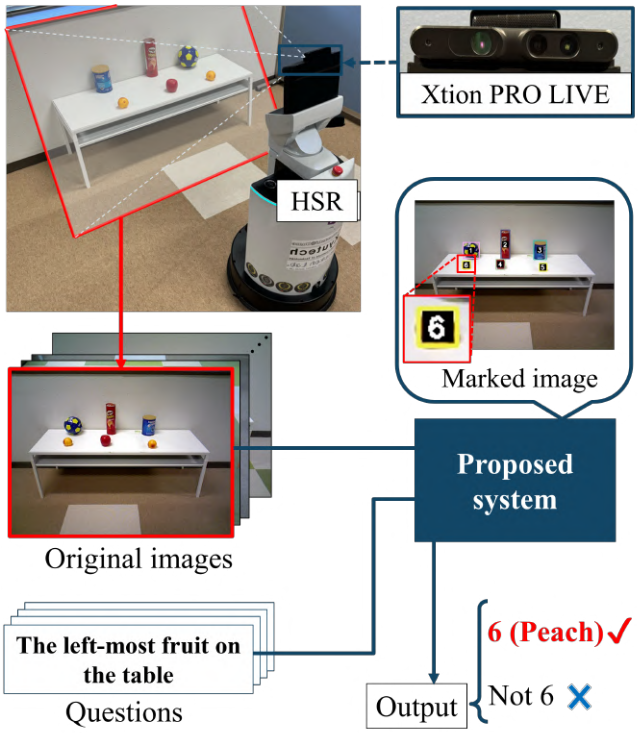

Fig. 4. **Schematic of the experiment.** Images taken by HSR and questions created from a template are input into the system, and the percentage of correct answers is evaluated.

Fig. 5. **List of objects used in the experiment.** 35 objects were selected from the YCB objects that were not too small.

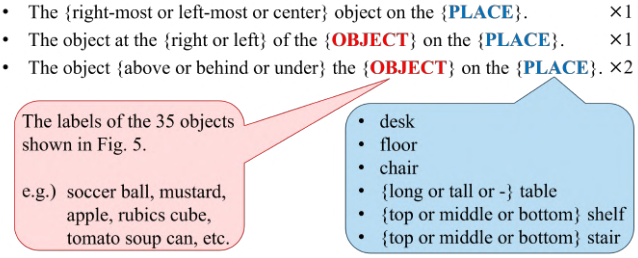

Fig. 6. Top: questions template and number of questions used in the experiment. Left: Describe and provide examples of nouns in the OBJECT of the questions template. Right: List of nouns in PLACE of the questions template.

- A description of the marked image, e.g., "The object's label corresponds to the mark: 1. pitcher, 2. chips, ..."
- A task description, e.g., "Your task is to identify the number based on a given command."
- A command indicating a target object, e.g., "Command: The left-most ball on the floor."

4) Input to the VLM: The marked image, original image, and text prompts are input to the VLM. This allows the VLM to focus only on the objects and improve the interpretation of the object positions.

Compared to SoM, which marks all segmented regions, including non-object areas, the proposed system focuses solely on target objects. This targeted marking improves the accuracy of object positions interpretation. Furthermore, the proposed system includes output labels from the object detection model in the text, improving accuracy. We focused on reducing the misidentification of similar objects by including labels. In addition, we input the original image, which marks do not hide the object, for an accurate object position interpretation.

IV. Experiment

A. Overview

Fig. 4 shows a schematic of the experiment. We evaluate the proposed system using 50 images, each paired with four questions. We inputted the prepared questions and images into the proposed system and evaluated the percentage of correct answers to the output numbers.

In this experiment, we used the you only look once v8 (YOLOv8) [26] trained on YCB objects, as the object-detection model. In addition, we used Gemini-1.5-pro, Gemini-1.5-flash [12], GPT-4o [10], and GPT-4 [9] as VLMs.

Furthermore, for the ablation studies on the proposed system, we conducted two types of evaluation experiments. First, the text prompts did not include labels from the object detection model. Second, only the marked images and text prompts were inputted into the VLM without the original image as the input.

B. Evaluation datasets

Images of 640 x 480 were captured with an RGB-D sensor (Xtion PRO LIVE) attached to a human support robot (HSR) [27] developed by Toyota Motor Corporation. There were 35 objects from the benchmark dataset, YCB objects [28], as shown in Fig. 5. The image contained between 5 and 15 objects.

Questions were generated based on the template shown in

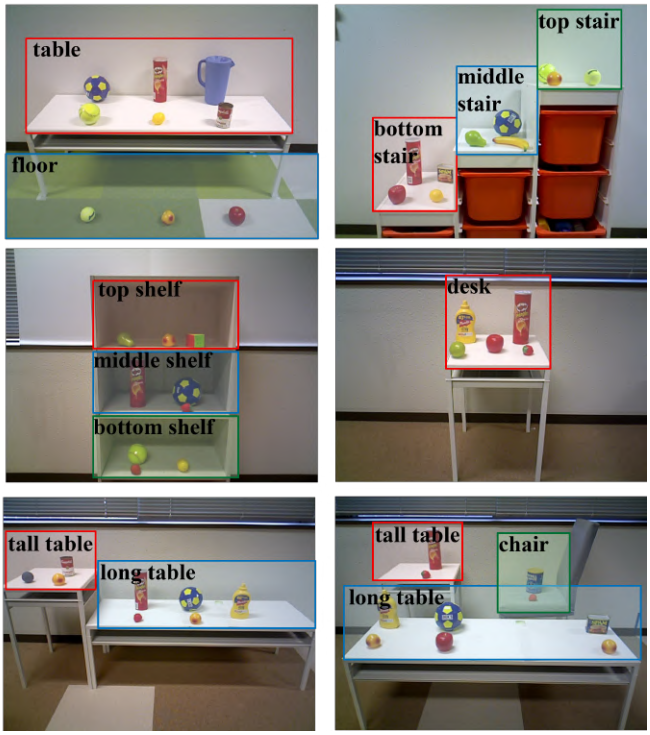

Fig. 7. **Areas assigned to PLACE.** The system is required to select the answer from the area corresponding to the PLACE.

Fig. 6. The labels of the 35 objects shown in Fig. 5 were used for the OBJECT of the question templates. As indicated in Fig. 6, enter 12 area names such as "desk" and "floor" in PLACE of the questions template. The system is required to select the answer from the area corresponding to the PLACE, as shown in Fig. 7.

C. Results

Table I lists the experimental results for interpreting the object positions. We confirmed that the proposed system achieved a higher correct answer rate than SoM did. In addition, the performance of the proposed method was improved by including the output labels by YOLOv8 in the text prompts.

In particular, Gemini 1.5-Flash improved the correct answer rate from 32.0 points to 76.5% with labels, compared with 44.5% without labels. Furthermore, we confirmed that including the original image improved the correct answer rate by several points for three models, namely GPT-4o, GPT-4, and Gemini-1.5-pro. However, Gemini-1.5-flash decreased the percentage of correct answers by 5.5 points to 39.0% when including the original image compared with 44.5% when excluding the original image.

V. Discussion

We confirmed that the proposed system is more accurate in interpreting object positions than SoM is. One reason for this phenomenon is that marks are only overlayed on the target objects. Fig. 8-A shows the marked image generated by the proposed system, and Fig. 8-B details an image generated

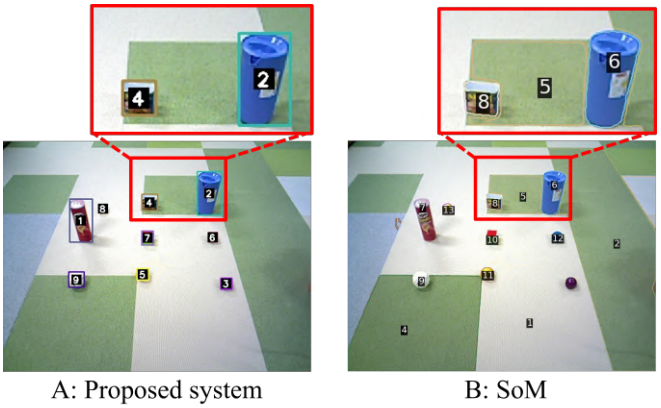

Fig. 8. **Marked images output from the proposed system and SoM.** Proposed system output is correct answers because it overlays marks only on the target object; SoM: may give wrong answers because it overlays marks even on the floor.

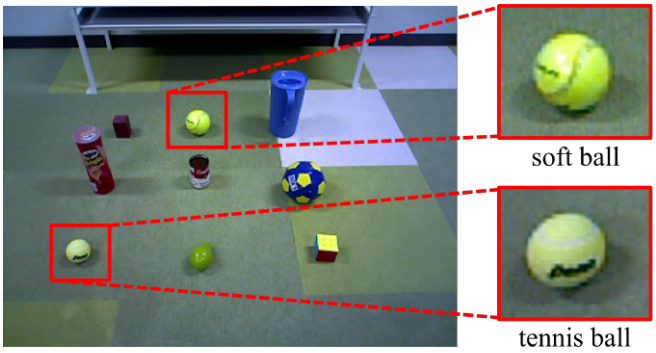

Fig. 9. **Example of similar objects.** Soft ball and tennis ball are similar in appearance and can be misidentified without labels.

by SoM. In Fig. 8-A, the proposed system outputs the correct answer, pitcher (#2), when it receives the question, "The object to the right of the spam can on the floor." However, in Fig. 8–B, the system outputs an incorrect answer for #5 marked on the floor. In SoM, the number is overlaid on the floor on the right side of the spam, which is the cause of the error. This case is frequent and contributes to the high accuracy of the proposed system.

In addition, we confirmed that including labels from the object detection model in text prompts improved accuracy. We speculated that this phenomenon could be because the inclusion of labels would reduce the misidentification of similar objects. Fig. 9 details a soft ball and tennis ball as examples of similar objects. Without labels, when the proposed system receives a question, "The object to the right of the soft ball on the floor," output the incorrect answer, an object to the right of the tennis ball. This problem occurs frequently and is a crucial for including labels to improve accuracy.

Furthermore, we confirmed that GPT-4o, GPT-4, and Gemini-1.5-pro improved the accuracy by several points when inputting the original image. We expected this result

TABLE I

EXPERIMENTAL RESULTS OF INTERPRETING OBJECT POSITIONS. WE USED A CORRECT ANSWER RATE [%] AS A EVALUATION METRIC.

		Models			
		GPT-4o	GPT-4	Gemini-1.5-pro	Gemini-1.5-flash
Proposed system	w/ Labels + Original image	**84.0**	**69.0**	**74.0**	74.5
	w/ Labels	79.0	68.0	68.5	**76.5**
	w/ Original image	58.0	41.5	55.5	39.0
	-	55.0	40.5	51.0	44.5
SoM	w/ Original image	55.5	41.5	36.5	41.5
	-	53.5	36.5	29.5	40.5

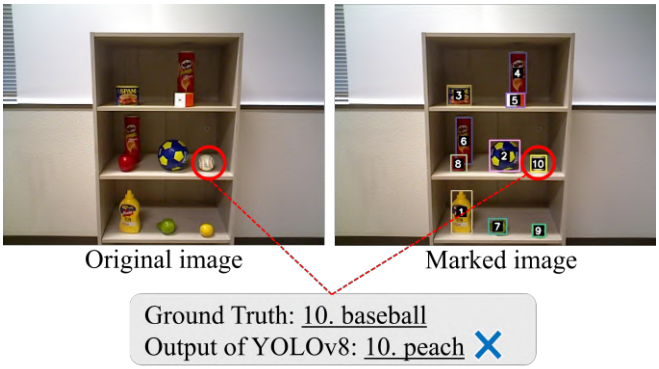

Fig. 10. Example of an image in which the answer is improved by inputting the original image.

because the marks on the image would hide the objects. Therefore, more accurate object recognition can be achieved by inputting an original image that is not covered. Fig. 10 shows an example of an image in which the answer was improved by inputting the original image. In Fig. 10, YOLOv8 incorrectly recognizes "baseball" (#10) as a "peach." Therefore, without the original image, the proposed system outputs the incorrect answer 8 for the question "The object at the left of the baseball on the middle shelf." However, when using the original image, the proposed system outputs the correct answer of 2. These results suggest that inputting the original image can reduce the effect of mislabeling because of the misrecognition of the object detection model.

Moreover, we consider that the decrease in the correct answer rate for the Gemini-1.5-flash could be attributed to the increased number of input tokens. Mosh et al. reported that in models such as Gemini, the inference performance decreases as the number of input tokens increases [29], indicating that more data input does not necessarily improve accuracy. On the other hand, Gemini-1.5-pro, with its more advanced token processing capabilities, is presumed to be able to effectively use additional information (labels and original images). However, many unresolved issues regarding VLM remain, and detailed research and verification are necessary. We focused on achieving accurate object recognition by inputting the original image in which the object is not hidden because the marks on the marked image hide the object.

VI. CONCLUSION

In this study, we proposed a novel system that interprets the relationship between target object positions by using an object detection model and a VLM. The proposed system for interpreting the relationship between object positions in an image uses an object detection model to overlay marks only on target objects. The VLM interprets the relationship between the marked object positions in the image and outputs the appropriate answers to the questions. Through the experiment, we confirmed that the proposed system outperformed SoM in terms of interpreting object positions. However, the percentage of correct responses was only 84%.

In the future, we plan to further improve accuracy by fine-tuning VLMs with marked images and their descriptions. Yan et al. propose a novel learning method utilizing SoM [30]: fine-tuning VLMs with marked images generated by SoM, which significantly improves visual inference performance and reduces misrecognition. We consider that fine-tuning the VLM with the marked image output from our proposed system will result in higher accuracy. Furthermore, we aim to implement the system on a robot and make it actually work.

REFERENCES

[1] "Fuji keizai group." https://www.fuji-keizai.co.jp/report/detail.html?code=162208813. (Accessed on 29/08/2023).

[2] 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 (RO-MAN)*, 2024.

[3] 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, pp. 870–889, 2022.

[4] K. Yamao, D. Kanaoka, K. Isomoto, A. Mizutani, Y. Tanaka, and H. Tamukoh, "Development of a saycan-based task planning system capable of handling abstract nouns," in *Proceedings of the International Conference on Artificial Life and Robotics (ICAROB)*, pp. 430–434, 2024.

[5] 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*, October 2023.

[6] 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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, pp. 8748–8763, 18–24 Jul 2021.

[7] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[8] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, vol. 202, pp. 19730–19742, 23–29 Jul 2023.

[9] OpenAI et al, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, March 2024.

[10] OpenAI, "Hello GPT-4o." `https://openai.com/index/hello-gpt-4o/`, 2024. (Accessed on 25/11/2024).

[11] Gemini Team et al, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, December 2024.

[12] Gemini Team et al, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 34892–34916, 2023.

[14] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge." `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024. (Accessed on 25/11/2024).

[15] C. Li, H. Liu, L. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, H. Hu, Z. Liu, Y. J. Lee, and J. Gao, "Elevater: A benchmark and toolkit for evaluating language-augmented visual models," in *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 9287–9301, Curran Associates, Inc., 2022.

[16] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8253–8280, May 2022.

[17] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, C. Finn, and K. Hausman, "Open-world object manipulation using pre-trained vision-language models," in *Proceedings of the 7th Conference on Robot Learning (CoRL)*, vol. 229, pp. 3397–3417, November 2023.

[18] T. Xiao, H. Chan, P. Sermanet, A. Wahid, A. Brohan, K. Hausman, S. Levine, and J. Tompson, "Robotic skill acquisition via instruction augmentation with vision-language models," *arXiv preprint arXiv:2211.11736*, 2023.

[19] 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 Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[20] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11987–11997, October 2023.

[21] 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*, April 2023.

[22] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11048–11064, 2022.

[23] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 22199–22213, 2022.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

[25] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," in *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 19769–19782, Curran Associates, Inc., 2023.

[26] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO." `https://github.com/ultralytics/ultralytics`, 2023. (Accessed on 25/11/2024).

[27] 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. 1, pp. 1–15, 2019.

[28] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *Proceedings of the 2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517, 2015.

[29] M. Levy, A. Jacoby, and Y. Goldberg, "Same task, more tokens: the impact of input length on the reasoning performance of large language models," *arXiv preprint arXiv:2402.14848*, 2024.

[30] A. Yan, Z. Yang, J. Wu, W. Zhu, J. Yang, L. Li, K. Lin, J. Wang, J. McAuley, J. Gao, *et al.*, "List items one by one: A new data source and learning paradigm for multimodal llms," *arXiv preprint arXiv:2404.16375*, 2024.