Wearable Robot Control Method Based on Vision-Language Models

Hyeon-Seok Seong[1], Seif Farag[2], Sang-Wook Lee[3], Igor Gaponov[4] and Jee-Hwan Ryu[1]

***Abstract*— This study introduces a novel application of Vision-Language Models (VLMs) in the field of human-robot interaction, specifically in controlling wearable robots with only a *single camera* setup. Thanks to the pre-trained knowledge of VLMs, our approach can estimate the weight of grasped objects in an industrial setup and accordingly adjust the robot's assistance mode. We detail the methodology of the control framework, including prompts that outline the hand gesture detection, identification of grasped objects, weight estimation, *training-less* datasets, and the response format for the robot control. This allows the system to adapt to a specific user environment without the need for extensive dataset collection, model training, or fine-tuning. Our method has been demonstrated in real-world applications with the wearable robot to confirm its feasibility.**

I. INTRODUCTION

In the field of human-robot interaction (HRI), the challenge of harmonizing human intentions with the assistance of robots is critical. The pursuit of intuitive interfaces has prompted the investigation of control mechanisms beyond conventional manual inputs. Notably, advancements in vision-language models (VLMs) present new opportunities for enhancing interaction between humans and robots. These models excel in processing complex visual scenes and interpreting linguistic commands, facilitating more autonomous and adaptable robotic tasks [1][2].

Traditional sensor-based control methods, though effective, often require complex and uncomfortable setups that are both annoying for the end user and require extensive experience for correct control and integration. Moreover, these methods can be restrictive and less adaptable to dynamic, real-world environments. Straightforward voice control is usually not feasible in an industrial or noisy setting while requiring the users to explicitly state what they want, which introduces further constraints. Our research seeks to address these limitations by introducing a vision-based control mechanism that utilizes the advanced image processing and object recognition capabilities of VLMs that are paired with their semantic understanding of these objects.

As shown in Figure 1, traditional methods for detecting hand gestures such as Google's Mediapipe model completely fail in detecting any hand key points when the hand is occluded by itself (in a 2D image) or by an object or a tool. In contrast, VLMs can detect the correct gesture as well as what object is being held which enables our system to estimate object weights and adjust wearable robot states accordingly. This approach fosters a more natural interaction, allowing the robot to proactively support the user's intentions by providing higher or lower assistance as needed.

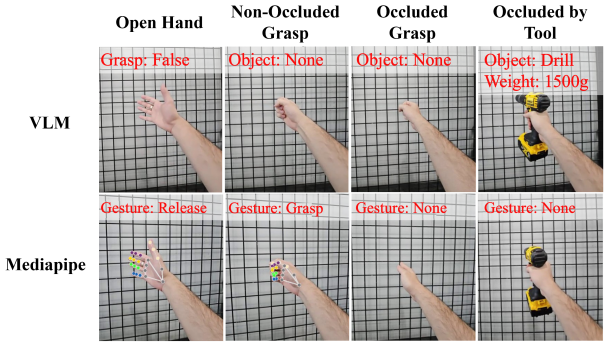

Fig. 1. Comparison of results using Mediapipe-based classification model and Visual Langauge Model(VLM) for hand poses captured by a first-person view camera. Even if the fingers are obscured by other parts of the hand or a tool, VLM can identify the hand pose and infer information about the object being held.

A key aspect of our study is a 'training-less' system that facilitates the quick incorporation of new tools and tasks without extensive retraining or data annotation. This adaptability is especially valuable in industries with diverse tools and task requirements [3][4]. This *training-less* dataset approach, where entering the name and weight of a new tool allows the system to immediately incorporate this data without additional training (unlike traditional models such as Grounding-Dino, SAM, or standard YOLOv8) or the need to do Retrieval Augmented Generation (RAG) which is a method used to retrieve data stored in a vector-database and gives the autoregressive model memory-like capabilities.

This approach has proven to be effective as it can better 'understand' specific situations where traditional methods fail, in particular hand movement recognition (e.g. grasping), under occlusions (while holding tools) and in a cluttered environment (industrial setting). The system can also easily adapt to any environment without the need for retraining. Another useful feature of this system is that it can estimate the approximate weight of objects fairly accurately, which is an emergent property coming from the VLM's vast training set, offering a flexible solution for numerous applications.

In the following sections, we will first describe in detail the methodology of the algorithms we implemented, and then demonstrate the results of applying VLM-based object weight estimation to a real-world system.

*This work was supported by Samsung Heavy Industries

The authors are with the [1]Dept of Civil and Environmental Engineering and [2]Robotics Program, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea {hysk.seong,s.farag,jhryu}@kaist.ac.kr

[3]Sang-Wook Lee with the Samsung Heavy Industry, Geoje, 53206, South Korea sw2412.lee@samsung.com

[4]Igor Gaponov with the Dept of Computer Science, University College London, London, WC1E 6BT, UK i.gaponov@ucl.ac.uk

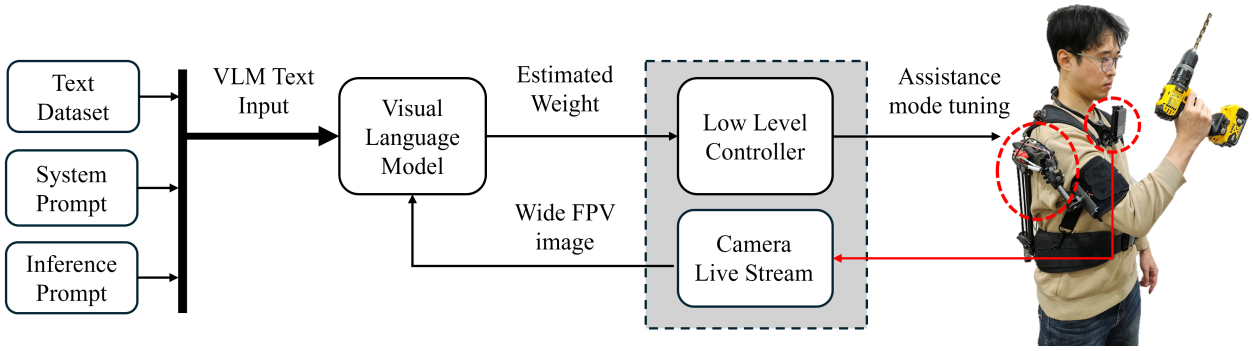

Fig. 2. A high-level control framework that adjusts the assistance level of the upper-arm wearable robot based on the weight of the object in the user's hand, estimated through prompts sent to a Vision-Language Model (VLM) using images streamed in real-time from a camera attached to the body.

II. Methodology

In the proposed approach, we intended to use a single camera to capture images of the user interacting with various objects, without any dedicated control sensors. These images were then processed by a VLM, with both GPT-4V(vision) and LLaVA tested as they are currently the state-of-the-art proprietary and open-source models. The VLM then tried to recognize whether the user was grasping an object and, if that was the case, attempted to identify the object and estimate its weight based on pre-existing knowledge acquired at the pre-training phase. Based on the information obtained as the result of this approach, the wearable robot could dynamically adapt the level of the support it provided to the user during manipulation.

To test the effectiveness of this system we used a general prompt that requested the model to estimate the weight of the objects being actively grasped by a human hand. This approach showed promise; after the model thoroughly described the image, we found that both the object name (or category) and its approximate weight were quite accurate. To parse the weight with a consistent unit and constantly identify the name of the object within a similar category, we devised a novel prompt. This prompt instructed the VLM to follow a strict answer guide more rigorously and to focus more on the steps it should take to provide the answer. It follows multiple guidelines published by several leading research groups such as Meta's open-source guide for LLM prompting, OpenAI's Meta-Prompting [5], and several studies by Google Brain and Deep Mind [6], [7].

The prompting techniques employed here incorporate the

- **Chain of Thought** approach which supports the use of logical steps in inference.
- **Capitalization** which forces the attention mechanism to give importance to specific parts of the prompt.
- **few-shot prompting** which provides multiple examples for the model to detect a pattern.

This combination improved the VLMs' analytical abilities, returning the response in a parseable format that includes three elements:

1) **Holding flag:** The model returns *True* or *False* depending on if there is an object being held or not.
2) **Object name:** The model returns the name of the object (or the closest category), which is mostly used for debugging and constructing a simple dataset.
3) **Estimated weight:** The model returns the object's weight as a number in *grams*.

The prompt is separated into three main parts as shown in Figure 2:

- A system prompt that describes the core behavior of the system and how to use the Chain of Thought to get a correct solution,
- An inference prompt, which asks the system to identify the situation and estimate the object's weight; and finally,
- A text dataset (training-less) which is a collection of names and corresponding weights for all objects of interest.

In the end, all prompts are combined at the inference stage to represent a single prompt, but initial separation supports adjusting the behavior or adding items to the dataset *during operation* in *real-time*. An example of the prompt used is provided in the appendix.

III. Demonstration

We verified the performance of our developed VLM-based control system applied to a wearable robot for upper arm assistance. As depicted in Figure 2, a camera mounted on the front of the wearable robot-wearing user transmits real-time images and pre-set prompts to the VLM, which then estimates the weight of the object held by the user in real-time.

A typical workflow unfolds as follows: Initially, the user wearing the wearable robot grasps a tool from a workstation to undertake a task. The scene of the hand holding a tool is captured by the chest-mounted camera and is transmitted in real-time to a wireless remote computer, where the latest video frame is decoded and tokenized and is then fed to the Vision-Language Model (VLM) alongside the combined prompts. Subsequently, the VLM processes this data to deduce the weight of the tool being handled and returns this information to the wearable robot, which then adjusts the assistance levels accordingly. The assistance mode is sustained throughout the task if there is no change in the

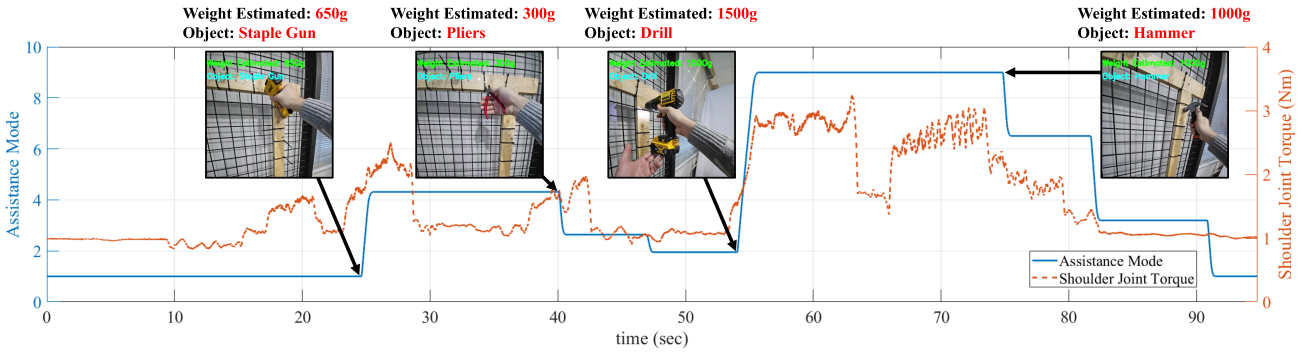

Fig. 3. As a demonstration result of applying a VLM-based controller to a wearable robot, the moment the weight of the object held in hand is estimated, the robot's assistance mode begins to change, subsequently altering the shoulder joint torque provided to the user.

inferred weight information. Upon the user returning the tool to the workstation and the VLM determining the absence of the tool in the user's hand, the wearable robot reverts to the minimal assistance mode.

The actual experiment was conducted in a workspace simulating various tool-handling work environments, with the users wearing an assistive robot with a camera while performing various industrial tasks using a staple gun, pliers, an electric drill, and a hammer. Figure 3 presents experimental results of comparing labeled images with data measured during the experiment as tools were switched, alongside the VLM-estimated object information. Initially, as the weight of the staple gun is estimated, the wearable robot's assistance level increases, leading to a change in the shoulder joint torque provided to the user. Subsequently, when handling the pliers, a lower measured weight prompted an adjustment of the assistance mode. While holding the electric drill, the increase in measured weight led to a peak in the shoulder joint torque, and finally, while holding the hammer, a decrease was observed.

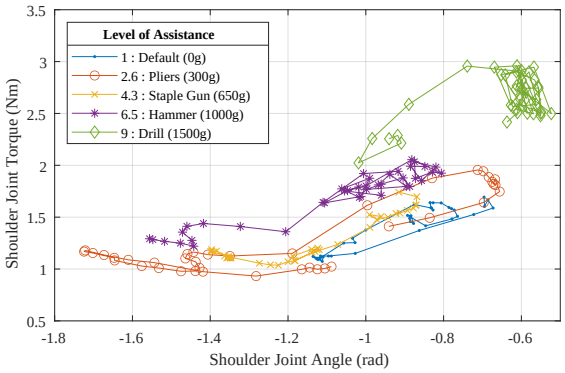

Fig. 4. Changes in shoulder joint torque of the wearable robot according to the assistance mode

Figure 4 shows the variations in shoulder joint torque of the wearable robot, adjusted following the weight information estimated by the VLM throughout the demonstration. The shoulder joint angle is set to 0 when the arm is parallel to the ground, and the user receives assistance from the wearable robot for compensating the weight of the tool held in the hand while lifting and lowering the arm during the task.

IV. Discussion and Conclusion

Our results showcase substantial benefits of using VLMs for task and object detection over conventional sensor-based configurations, notably in terms of simplifying setup processes, reducing costs, and enhancing user comfort, while also offering seamless adaptability to new tasks and environments. Nonetheless, we acknowledge certain limitations, particularly the reliance on the camera's field of view and challenges related to low-bandwidth control. The latter issue may be addressed through the employment of more advanced models, a subject currently experiencing vigorous research activity.

This study attempts to make another step towards intuitive human-robot interaction, highlighting the efficacy of vision-language models (VLMs) as a novel control mechanism for robotic systems. Our findings emphasize the critical role of adaptability and straightforward implementation in the evolution of assistance robots of the future, suggesting that the integration of VLMs could significantly augment the ways these systems understand and interact with their human counterparts. With faster, more accurate and coherent models being published daily, the possibilities of using VLMs for HRI are endless and more promising than ever.

Acknowledgments

This research was supported by Samsung Heavy Industries.

References

[1] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[2] L. Bärmann, R. Kartmann, F. Peller-Konrad, A. Waibel, and T. Asfour, "Incremental learning of humanoid robot behavior from natural interaction and large language models," *arXiv preprint arXiv:2309.04316*, 2023.

[3] M. Ficocelli, J. Terao, and G. Nejat, "Promoting interactions between humans and robots using robotic emotional behavior," *IEEE Transactions on Cybernetics*, vol. 46, pp. 2911–2923, 12 2016.

[4] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014, pp. 407–414.

[5] M. Suzgun and A. T. Kalai, "Meta-prompting: Enhancing language models with task-agnostic scaffolding," *arXiv preprint arXiv:2401.12954*, 2024.

[6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[7] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

APPENDIX

A. System Prompt

```
Below is an image which includes a hand and an object, your task is to determine if the hand is holding an object or not, and if so, estimate the mass of this object:
First, you should determine if the there is hand in the image
Second, you should determine if that hand is holding an object
Third, if the hand is indeed holding an object, you should estimate the mass of that object
Your answer should always include the TRUE if hand is holding an object, or FALSE if not holding an object
Your answer should always include the mass in grams, or '0' if no object can be found
Your answer should always include the name or category of the object being held.
YOU ARE ONLY ALLOWED TO RESPOND IN THE FOLLOWING FORMAT 'Holding: <TRUE or FALSE>, Object: <object_name>, Weight: <object_estimated_weight>', and the MAXIMUM object_estimated_weight is '3000 grams.'
```

B. Text Dataset

```
Drill, 2000g
Vaccum Cleaner, 1430g
Hacksaw, 323g
Screwdriver, 220g
...
```

C. Few-Shot Examples & Inference Prompt

```
Here are multiple examples for different situations:
example 1 (person holding a cup in his hand):
User: What is the weight of the object in the image?
Assistant: Holding: TRUE, Object: cup, Weight: 200 grams
example 2 (person in the image but his hands are not visible):
User: What is the weight of the object in the image?
Assistant: Holding: FALSE, Object: None, Weight: 0 grams
example 3 (person holding a power tool):
User: What is the weight of the object in the image?
Assistant: Holding: TRUE, Object: drill, Weight: 1500 grams
example 4 (person holding a hammer-like object):
User: What is the weight of the object in the image?
Assistant: Holding: TRUE, Object: hammer, Weight: 2000 grams

User: What is the weight of the object in the image?
Assistant: ...
```