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

Utilizing Knowledge in Vision-Language Model for Robotic Grasp Force Estimation through Robot-to-Human Image Translation

Shohei Hagane[1], Shigeaki Goto[1] and Yoshihiro Ohama[1]

Abstract— **In recent years, general-purpose robots have been introduced into domains requiring delicate manipulation, such as materials experimentation. While advances have been made in automating specific processes, generalized robotic pick-and-place operations still pose a challenge due to the diversity of target objects and the need for appropriate force control. This study proposes a novel approach for zero-shot estimation of target grasping force by utilizing the prior knowledge about human motions embedded in Vision-Language Model (VLM). The key idea is to convert robot manipulation images into human-action images using a style transfer approach based on a fine-tuned Variational Auto-Encoder (VAE), enabling the VLM to better infer grasping force requirements. The VLM, specifically GPT-4o, is prompted to estimate target grasping force in discrete categories (no grasp, light grip, firm grip).**

Experimental results demonstrate that converting robot images into human representations improves the accuracy not only of target grasping force estimation but also of understanding the target objects. Furthermore, the inclusion of target object information in the prompt improves estimation accuracy across all input image types. These findings highlight the effectiveness of utilizing human-knowledge-trained VLM for robotic force control and open new avenues for general-purpose, cost-efficient manipulation without relying on large-scale robot force datasets.

I. INTRODUCTION

Recently, general-purpose robots have begun to appear in places that need delicate handling, such as cooking and material experimentation. However, achieving fully generalized robotic implementations in these contexts remains challenging. Take materials-synthesis experiments as an example. While many processes have been automated using specialized equipment, general transport tasks which are primarily pick-and-place operations, remain a challenge for robots because of the large variety of target objects[1].

As promising methods for general pick-and-place tasks, learning-based methods that leverage large-scale models, such as Large Language Model (LLM), large-scale Vision-Language Models (VLM), and Robot Foundation Model (RFM), have attracted attention. By exploiting large training datasets, these methods can adapt to a variety of tasks in a zero-shot fashion. However, they tend to fail when tasks require force control, such as grasping and lifting heavy objects or delicately handling fragile items. This limitation arises because the models have not adequately learned from robotic force-sensing data and therefore cannot output the target gripping force needed for the gripper. In addition, it is difficult to assemble a sufficiently large, format-uniform force database for training such large models. It is because proper robotic force data formats vary in their requirements depending on the embodiment of the system and the control objectives. Consequently, this study aims to develop technology that can universally generate target control values for robotic force control. Specifically, our goal is to zero-shot estimate the "target grasping force," which is the amount of force the robot should apply at the moment during general pick-and-place tasks in materials experiments.

In this study, a new approach to effectively use the large-scale models with knowledge of human actions for force control is proposed. The proposed approach is utilizing a style transfer technique that converts robot information to human information. Training data for large-scale generative AI models such as VLM are collected at an internet scale, and the amount of content is heavily inclined toward human-related data compared to robot-related data. In other words, it is assumed that large-scale models generally accumulate extensive knowledge about typical human actions. Given this, it may be more effective to use images of humans rather than images of robots in order to utilize the knowledge embedded in existing VLM. In this paper, as shown in Fig. 1, we propose a novel approach for zero-shot estimation of target grasping force by converting robot images into human images and feeding them into a VLM that has been pre-trained on human action knowledge.

To evaluate the feasibility of the proposed approach, the following experiment was conducted. For a pick-and-place task commonly observed in the field of materials experimentation, we converted images of tasks performed by a robot into human-converted images. The converted images, along with a prompt for grasp force estimation, were input into a VLM trained on human data, which inferred the appropriate target grasping force under the given conditions. The estimated target grasping force is compared with the ground-truth labels, and its performance was evaluated in terms of the F1 score derived from the confusion matrix.

The contributions of this work are as follows:

- This approach introduces a new research framework that compensates for the chronic lack of high-quality force data and force policies in the robotics field using "robot data + style translation + VLM inference." It shows the way for applying multimodal large-scale models to robotic force control.
- This paper showed the applicability of the proposed approach to robotic pickking tasks and highlighted its potential extensions to fields involving force-sensitive manipulation.

[1]Authors are with Toyota Central R&D Labs., Inc., 41-1, Yokomichi, Nagakute, Aichi 480-1192, Japan Shohei.Hagane.mz@mosk.tytlabs.co.jp

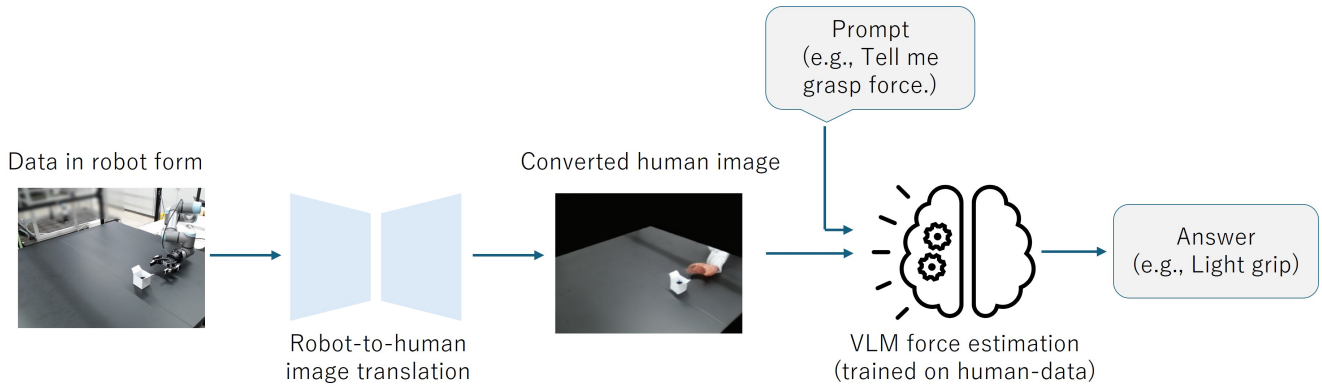

Fig. 1. Overview of proposed approach

II. Related Work

A. Robot Control with Vision-Language Model

VLM are being adapted to control robotic manipulators, enabling robots to interpret visual scenes and follow natural language instructions during manipulation tasks. Since VLM are trained on large amount of data, it has ability to adapt various task settings. Models like RT-X[2] and VIMA[3] exploit large-scale training to achieve versatile pick-and-place manipulation, yet they omit force data for training, so they cannot generate explicit force targets during execution. In contrast, $\pi^{0.5}$[4] and ForceVLA[5] can handle tasks that require force control. However, it still depends on additional task-specific fine-tuning with supervised demonstrations. Consequently, there remains a significant challenge in building up diverse and scalable force control datasets, which are essential for robots to acquire generalized capabilities in force-aware manipulation.

B. Force Control in Robot Manipulation

Recent surveys have shown that robotic force control methods still heavily depend on task-specific modeling and parameter tuning, which limits their generalizability [6]. Although numerous studies have explored learning-based approaches to improve the generality of force control, such as adaptive impedance control[7] and reinforcement learning for adaptive force control[8], the format of force data remains inconsistent across studies. As a result, the scalability of these methods is often confined to in-domain tasks.

C. Visual Domain Adaptation for Robot

Generative models that can convert images of robots performing tasks into images of humans performing the same tasks do exist. This is a technique known as image-to-image translation, and CycleGAN[9] is one of the state-of-the-art methods. This research aims to learn a pixel-level mapping between the image domain of robot tasks and that of human tasks. The technology is used to train robots via reinforcement learning using human demonstration videos. On the other hand, as far as the author is aware, there are still no studies that convert robot images into human images for the purpose of leveraging VLM.

this research is filling the gap between robot control with VLM and force control using image-to-image translation.

III. Proposed Approach

As shown in Fig. 1, the proposed approach consists of ”Robot-to-Human Image Translation” and ”Vision-Language Force Estimation”. The first process converts images of a robot into images of a human performing the same task. As the second process, the VLM receives the converted image along with a prompt that instructs it to infer the proper grasping force.

A. Robot-to-Human Image Translation

To convert robot images to human images, VAE was used. The VAE is pretrained on Open-Images-data-set v4[10]. The encoder part of this VAE is fine-tuned to map similar human tasks and robot tasks into the same feature space using pairs of robot and human images. In this study, we used 234 pairs of robot and human images which are prepared by the author for fine-tuning. The contents of the images are pick-and-place tasks of 14 objects related to material experiments (see TABLE I). Fig. 2 shows an example of the conversion of a newly acquired validation robot image using the fine-tuned VAE.

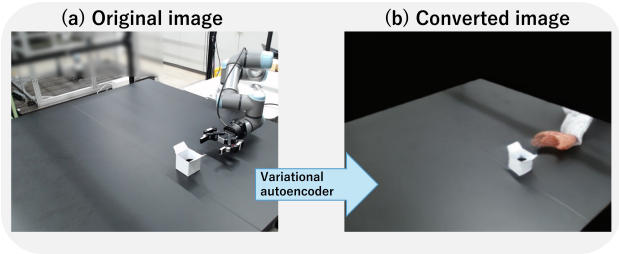

Fig. 2. Result of robot image transformation

B. Vision-Language Force Estimation

To illustrate the concept of the proposed approach, a VLM that satisfies the following requirements is necessary:

- A model trained on data that includes human actions

TABLE I

SPECIFICATIONS OF ITEMS USED IN THIS STUDY

ID	Explanation
1	Paper book (height is 250 mm, width is 150 mm, thickness is 70 mm)
2	Empty paper box (height is 50 mm, width is 50 mm, depth is 50 mm)
3	Empty aluminum-foil dish (height is 40 mm, diameter is 40 mm)
4	Plastic reacher grabber (length is 600 mm)
5	Empty plastic reagent bottle (height is 250 mm, diameter is 75 mm)
6	Empty plastic box (height is 50 mm, width is 130 mm, thickness is 70 mm)
7	Water-full plastic squeeze wash bottle (height is 200 mm, diameter is 60 mm)
8	Empty plastic squeeze wash bottle (height is 200 mm, diameter is 60 mm)
9	Empty plastic beaker (height is 130 mm, diameter is 60 mm)
10	Empty plastic reagent bottle (height is 150 mm, diameter is 40 mm)
11	Empty plastic vial (height is 30 mm, diameter is 35 mm)
12	Empty plastic beaker (height is 90 mm, diameter is 40 mm)
13	Empty plastic reagent bottle (height is 150 mm, diameter is 40 mm)
14	Water-full plastic reagent bottle (height is 250 mm, diameter is 75 mm)

- A model with the capability to understand and respond to the instruction ”estimation of the target grasping force”

In this study, we employ GPT-4o as the VLM that meets the above requirements. GPT-4o is a general-purpose multimodal generative AI model, capable of accepting not only natural language prompts but also images and audio as input. In addition to being trained on vast amounts of data, GPT-4o also utilizes autonomous internet search functionality to comprehensively cover common knowledge. Therefore, it is considered to have accumulated an extensive amount of data related to human actions, and has the capability to follow any given instruction.

The prompt used as input to GPT-4o to infer target grasping force is shown in Fig. 3. This prompt is designed to estimate the target grasping force in a step-by-step manner. Furthermore, the prompt specifies that the estimated target grasping force should be expressed using predefined terms. By extracting the specified grasping force term from the output of GPT-4o, the final target grasping force is obtained.

Note that the labels “Almost none” and “Light grip” are treated as equivalent (i.e., 1: light grip). As a result, the final set of target grasping force labels is defined as: 0: No grasp, 1: Light grip, 2: Firm grip.

IV. EXPERIMENTAL SETUP

To clarify the effect of input image type on the VLM’s estimation of target grasping force in pick-and-place tasks, evaluations under the following three conditions are conducted:

- Estimation using images of the robot performing tasks
- Estimation using human images converted from robot images
- Estimation using images of a human performing tasks

You are given an image showing a moment during a materials experiment. Your task is to analyze the scene and describe the hand activity being performed. Please follow the steps below in order to construct your response:

1. Understand and describe the task:
Carefully observe the image and infer what kind of task is being performed by the hands. Describe the hand activity in detail.

2. Identify the manipulated object:
Based on your description, identify the object (if any) currently being manipulated.

* If no object is being manipulated, output:
“Nothing”
and end your response.

3. Determine hand contact:
If an object is being manipulated, examine the image to determine whether the hand is in contact with the object.

4. Estimate gripping force:
If the hand is touching the object and appears to be grasping or about to grasp it, select the appropriate gripping force that should be applied at this moment from the following list:
[“Zero”, “Almost none”, “Light grip”, “Firm grip to avoid dropping”]

Please provide your answer in a clear and structured format, addressing each step explicitly.

Fig. 3. Actual prompt text that is input to GPT-4o

Experiments were conducted on 14 objects as listed in TABLE I. Images of pick-and-place actions, captured at 0.5-second intervals, are input to GPT-4o in chronological order to estimate both the target grasping force and the target object at each moment. An overview of the experiment is shown in Fig. 4. To compensate for the probabilistic variability of GPT-4o’s output, each image is evaluated 10 times, and the most frequent grasping force among the outputs is adopted as the final target grasping force. If the situation cannot be determined (i.e., GPT-4o declines to provide an answer), the target grasping force from the previous frame is retained. If the initial frame cannot be determined, the target grasping force is set to ”0: No grasp”.

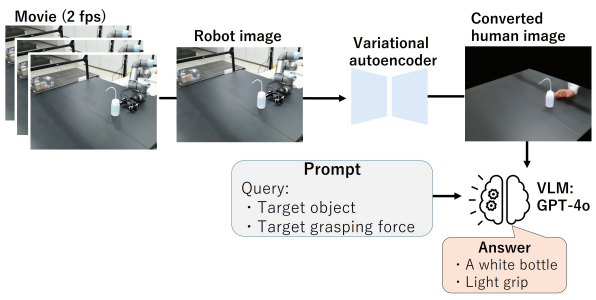

Fig. 4. Overview of the experimentation

A. Evaluation method

1) Target Object Estimation Accuracy: To evaluate whether GPT-4o is making appropriate situational judgments,

TABLE II

GROUND TRUTH LABELS ANNOTATION RESULTS

ID	Maximum measured grasping force [N]	Ground truth labels
1	> 12	2:"Firm grip"
2	< 0.1	1:"Light grip"
3	< 0.1	1:"Light grip"
4	> 12	2:"Firm grip"
5	2.8	1:"Light grip"
6	1.7	1:"Light grip"
7	1.7	1:"Light grip"
8	4.9	1:"Light grip"
9	1.0	1:"Light grip"
10	1.1	1:"Light grip"
11	< 0.1	1:"Light grip"
12	2.2	1:"Light grip"
13	0.6	1:"Light grip"
14	> 12	2:"Firm grip"

the accuracy of its estimation for grasped objects is investigated. Since variations in naming of objects (i.e., expression differences) must be accounted for when determining the correctness of grasped object estimation, we utilized another GPT-4o as the evaluator to avoid subjective bias from human judgment. The prompt used for this evaluation is shown in Fig. 5.

```
The target object is "{target_object}".
Please determine whether this object or an equivalent object
(e.g., a paraphrase or something in the same category) appears
in the following sentence as manipulated object.
Consider whether the meaning matches, and output only "True"
or "False".

Sentence: "{sentence}"
```

Fig. 5. prompt for checking correctness of grasped object estimation

To verify the reliability of this evaluation process, we manually assessed the correctness of predictions specifically for the case in which the grasped object was a "paper book," and compared the results with those judged by GPT-4o. Since the results matched exactly, we concluded that the correctness evaluation was valid.

2) Target Grasping Force Estimation: The prediction performance for the target grasping force was assessed using the macro-averaged F1-score, considering the imbalance between the class categories. Since the average is taken regardless of the number of samples per class, the performance of minority classes is also emphasized.

B. Data Collection and Annotation

In this study, the ground truth labels for the target grasping force of each object are determined using a force sensor (GN20H, manufactured by OOTAHIRO Co., Ltd.[11]). The experimenter manually grasps each object 2 times, and the maximum value of the applied grasping force is measured. Using the median value of the sensor's measurement range (6.0 N) as a threshold, grasps are classified into either "light grip" or "firm grip." TABLE II shows the maximum measured grasping force for each object and the corresponding ground truth labels.

V. RESULTS AND DISCUSSION

A. Results of Target Grasping Force Estimation

Fig. 6 shows the classification results of grasping force estimated by GPT-4o. As indicated by the comparison of F1-scores, the accuracy of target grasping force estimation improves as the input images become more human-like. In addition, the ability to correctly infer the timing of grasping versus non-grasping actions also improves with human-like images. Furthermore, it was observed that the likelihood of selecting "1: Light grip" increases as the images resemble human figures more closely.

Fig. 6 shows the classification results of grasping force estimated by GPT-4o. As indicated by the comparison of F1-scores, the accuracy of target grasping force estimation improves as the input images become more human-like. In addition, the ability to correctly infer the timing of grasping versus non-grasping actions also improves with human-like images. Here, "grasping at the correct timing" indicates that GPT-4o inferred a "grasp" action in the same frame where the ground truth data show the object being grasped. In the case of robot images, it is considered that GPT-4o may have had difficulty recognizing the gripper itself or distinguishing it from the background, possibly due to the limited amount of gripper-related information in its training data. Furthermore, it was observed that the likelihood of selecting "1: Light grip" increases as the images more closely resemble human figures.

B. Results of Estimation Results of Grasped Objects

TABLE III presents the accuracy of grasped object estimation.

TABLE III

CORRECTNESS OF GRASPED OBJECT ESTIMATION

	Robot image n=504	Human-converted image n=504	Human image n=518
Correct answer rate	0.492	0.183	0.637

As shown in TABLE III, inputting real human images results in the highest accuracy in identifying the grasped object. On the other hand, when robot images are converted into human images, the accuracy drops significantly. This decline is likely due to the image conversion process using a VAE, which causes blurring around the hands and makes the edges between space and objects ambiguous, thereby negatively affecting the object recognition performance of the VLM.

Interestingly, as seen in the results for the human-converted images in Fig. 6 and Table III, GPT-4o was able to estimate the target grasping force more accurately than in the case of robot images, even when the object information was incorrect. This phenomenon is likely attributed to a

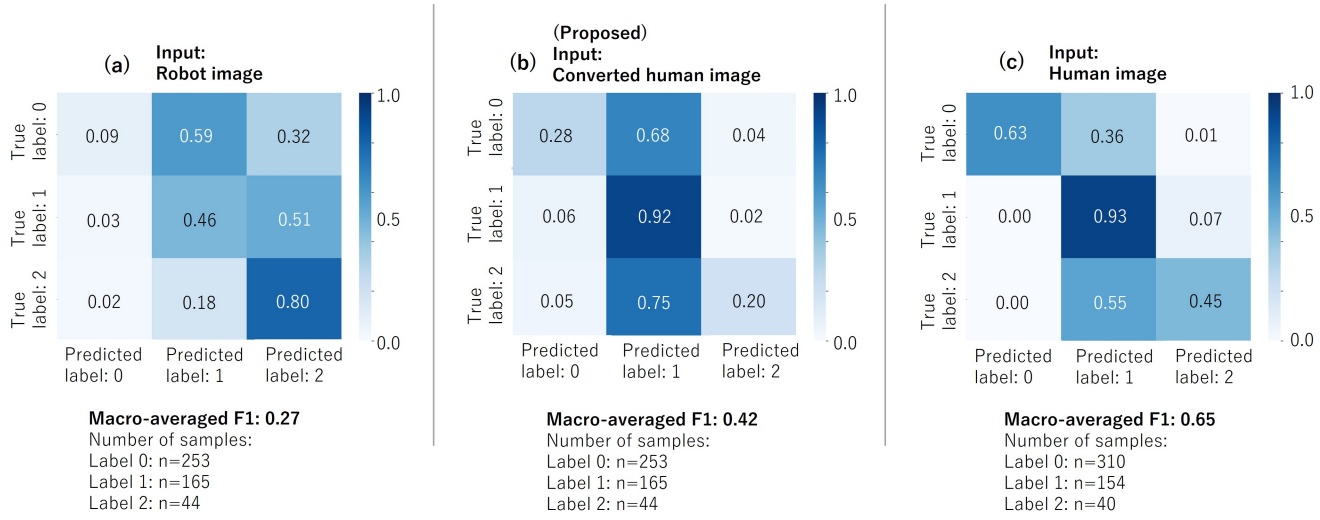

Fig. 6. Confuison matrices of target grasping force estimation

combination of the tendencies in GPT-4o's output behavior and the imbalance in the distribution of ground truth labels.

In this study, the model tended to predict the label "2: Firm grip" more frequently for robot images, and label "1: Light grip" more frequently for human images. There was also a notable imbalance in the number of samples for each label: only about 10% of the ground truth samples corresponded to firm grips, whereas approximately 35% corresponded to light grips.

As a result, when GPT-4o receives human-like images and increasingly outputs "light grip" at appropriate moments, its performance in estimating the target grasping force (measured by the F1-score) improves, even if its understanding of the grasped object itself is unclear.

C. Target Grasping Force Estimation with Target Object Information

Object information is considered useful for estimating the target grasping force. This is supported by the fact that, as shown in Fig. 6, the F1-score is higher when real human images are used, where object information is more accurately understood, compared to the case with human-converted images (see TABLE III).

To test this, the object properties and sizes (as listed in TABLE I) are added to the prompt in Fig. 3 and input to GPT-4o to estimate the target grasping force. TABLE IV shows the comparison of confusion matrices' F1-scores for each image input condition.

TABLE IV
CORRECTNESS OF GRASPED OBJECT ESTIMATION

Macro-ave F1	Robot image	Human-converted image	Human image
w/o object info	0.27	0.42	0.65
with object info	0.36	0.53	0.75

As shown in TABLE IV, the F1-score improved by approximately 0.1 points in all cases. This indicates that contextual information, such as the size and nature of the grasped object, is useful for estimating the target grasping force.

VI. Limitations and Future Work

A. Method for Converting Robot Information into Human Representations

In this study, due to the limited performance of the VAE, GPT-4o was unable to sufficiently understand information about the situation. In future work, we aim to improve accuracy by employing more advanced image conversion techniques, such as Cycle-GAN, which can generate higher-resolution images. Furthermore, generating human representations that better reflect the actual context—based on textual descriptions of the robot's working environment—may also contribute to improved accuracy, especially for dynamic actions such as throwing objects. In this direction, diffusion-based image generation models could also serve as a promising approach.

B. Evaluation Method for Grasping Performance

In this study, the effectiveness of the proposed approach was evaluated based on the F1-score derived from the confusion matrix of the target grasping force estimations. However, the threshold F1-score required for practical deployment, as well as the actual grasping success rate achievable, must be validated using a physical robotic system. In addition, future work will include verifying the proposed method across a wider range of tasks and objects to confirm its generality.

C. Discrepancy in Ground Truth Due to Embodiment Gap

In simple pick-and-place tasks, the appropriate grasping force can remain consistent regardless of finger shape, as long as the properties of the object and its intended use are understood. However, for more complex manipulations beyond simple force adjustment, the force actions generated by models trained on human data do not necessarily represent

the correct actions when applied to a robot with a different embodiment. Therefore, identifying a function that maps the appropriate grasping forces in the human domain to those in the robot domain will also be an important direction for future work.

On the other hand, it could be possible to utilize VLM not as a controller of detailed actions, such as joint-level position data, but as a more abstract, goal-directed motion planner. In this role, the VLM could provide insight into the appropriate force amount and place to apply, based on the notion that "a human would manipulate this part of the object in this way," regardless of the robot's form. In the future, we also plan to explore this potential through the approach proposed in this study.

VII. CONCLUSIONS

In this study, we proposed a novel approach for zero-shot estimation of robotic target grasping force by utilizing the knowledge of VLM trained on human-related data. The key idea behind the method is to convert robot images into human-like representations using image-to-image translation, thereby enabling VLM to more effectively infer target grasping force.

Experimental results using GPT-4o demonstrated that the accuracy of target grasping force estimation improves as the input image becomes more human-like. Also, when human images are input to GPT-4o, the estimation accuracy of the grasped object is higher than in the case of robot images. Furthermore, the addition of contextual object information to the prompt further enhanced prediction performance, highlighting the importance of both visual and semantic cues.

This approach presents a new direction for robot force control by enabling the use of existing VLM without additional fine-tuning or force data collection. By translating robotic information into a format more accessible to such models, our method opens up new opportunities for generalizable, cost-efficient robotic manipulation in unstructured environments.

Future work will focus on improving the quality of robot-to-human image translation, as well as validating real-world robotic performance through physical implementation.

APPENDIX

A. Grasping Experiment

To evaluate the current performance of the proposed approach, a simplified grasping experiment was conducted. In this experiment, only the picking motion is executed to verify whether grasping can be achieved with the estimated target gripping force. An experiment overview is shown in Fig. 7.

In this experiment, UR3e from Universal Robots[12], and gripper is 2f-85 from Robotiq[13] were used. The experimental setup is described below. At the start of each trial, the object to be grasped is placed 5 cm away from the gripper, and the control cycle is executed. At every control cycle, the robot image observed by the camera is converted into a human-view image, and GPT-4o estimates the target

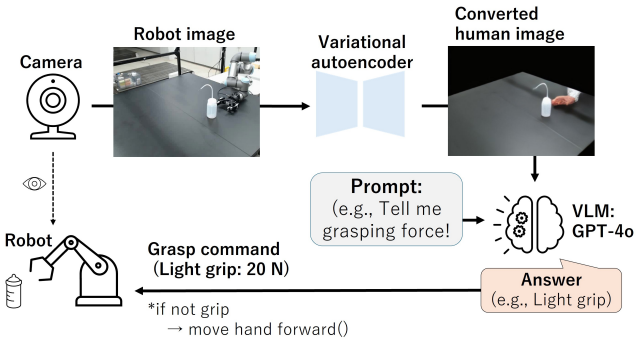

Fig. 7. Overview of the grasping experiment

gripping force using the prompt shown in Fig. 3, which is then executed. If the estimated target gripping force is "0: no grasp," the end-effector is moved 5 cm closer to the object, and the control cycle is repeated.

The gripping forces corresponding to the target states were set to 20 N for a "light grip" and 230 N for a "Firm grip." The force for a light grip was determined from the average maximum gripping force (20.5 N) measured for the "light grip" objects in TABLE I. The force for a Firm grip was set to the maximum gripping force of the gripper.

The objects to be grasped were selected along two axes: weight (Heavy/Light) and fragility (Solid/Fragile). In this experiment, the following four objects are grasped:

- Paper book (Heavy-Solid)
- Empty paper box (Light-Fragile)
- Empty plastic squeeze wash bottle (Heavy-Fragile)
- Empty plastic reagent bottle (Light-Solid)

Five trials are conducted for each object, and success is defined as grasping the object without breaking it or spilling its contents. In addition to the number of successful grasps, cases in which the system correctly judged the timing to start grasping ("Timing" in TABLE V) are counted. Here, "grasping at the correct timing" refers to the case where the hand begins to close for the first time when the grasped object is already in range of robot's fingers. Cases in which the system correctly estimated the target gripping force amount ("Force" in TABLE V) are also counted.

B. Result

TABLE V shows the grasping success rates for four types of objects, categorized by weight and softness, using (a) robot operation images and (b) human-transformed images converted from robot operation images. Fig. 8 is one example of experiment result of human-converted image case.

When robot images were input into GPT-4o, there were no successful grasping cases. In contrast, when the images were converted into human operation images, at least one successful case was observed for each of the four target objects. The estimation of grasping timing also showed higher accuracy in the human-converted image condition, demonstrating the same trend observed in Section Fig. 6.

TABLE V

NUMBER OF SUCCESSFUL GRASPING TRIALS

Object	(a)Robot image			(b)Human-converted image		
	Timing	Force	**Result**	Timing	Force	**Result**
Paper book	0/5	0/5	**0/5**	3/5	2/5	**2/5**
Empty paper box	0/5	5/5	**0/5**	2/5	5/5	**2/5**
Empty plastic squeeze wash bottle	0/5	3/5	**0/5**	1/5	5/5	**1/5**
Empty plastic reagent bottle	0/5	3/5	**0/5**	1/5	5/5	**1/5**

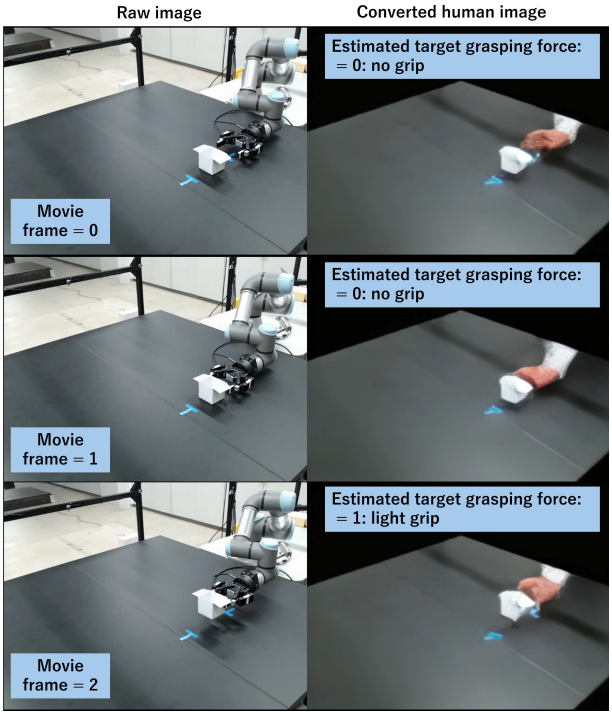

Fig. 8. Example of experiment result of human-converted image case

REFERENCES

[1] L. Hung, J. A. Yager, D. Monteverde, D. Baiocchi, H.-K. Kwon, S. Sun, and S. Suram, "Autonomous Laboratories for Accelerated Materials Discovery: A Community Survey and Practical Insights," *Digital Discovery*, vol. 3, no. 7, pp. 1273–1279, 2024.

[2] A. O'Neill *et al.*, "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," 2025, arXiv:2310.08864.

[3] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: General Robot Manipulation with Multimodal Prompts," 2023, arXiv:2210.03094.

[4] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: A Vision–Language–Action Model with Open-World Generalization," 2025, arXiv:2504.16054.

[5] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, C. Lu, and W. Zhang, "ForceVLA: Enhancing VLA Models with a Force-aware MoE for Contact-rich Manipulation," 2025, arXiv:2505.22159.

[6] M. Suomalainen, Y. Karayiannidis, and V. Kyrki, "A Survey of Robot Manipulation in Contact," *Robotics and Autonomous Systems*, vol. 156, p. 104224, 2022.

[7] M. F. Karim, S. Bollimuntha, M. S. Hashmi, A. Das, G. Singh, S. Sridhar, A. K. Singh, N. Govindan, and K. M. Krishna, "DA-VIL: Adaptive Dual-Arm Manipulation with Reinforcement Learning and Variable Impedance Control," 2024, arXiv:2410.19712.

[8] S. Sakaino, N. Masuya, H. Sato, K. Yamane, T. Kusume, M. Konosu, and N. Imazu, "Practical Implementations of Bilateral Control-Based Imitation Learning at iREX2023," in *Proceedings of the 10th IEEJ International Workshop on Sensing, Actuation, Motion Control, and Optimization (SAMCON)*, 2024, pp. 213–218.

[9] L. Smith and M. Zhang, "Learning to Imitate Human Demonstrations via CycleGAN," https://bair.berkeley.edu/blog/2019/12/13/humans-cyclegan/, 2019, bAIR Blog (Accessed: 2025-07-23).

[10] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, mar 2020.

[11] Ootahiro Co., "GN20H Tactile Sensor," https://www.ootahiro.co.jp/products/products01_06.html, 2025, accessed: 2025-07-23.

[12] Universal Robots, "UR3e Collaborative Robot," https://www.universal-robots.com/products/ur3e/, 2025, accessed: 2025-07-23.

[13] Robotiq, "2F-85 Adaptive Gripper," https://robotiq.com/products/adaptive-grippers, 2025, accessed: 2025-07-23.