Daily Assistive View Control Learning of Low-Cost Low-Rigidity Robot via Large-Scale Vision-Language Model

Kento Kawaharazuka[1], Naoaki Kanazawa[1], Yoshiki Obinata[1], Kei Okada[1], and Masayuki Inaba[1]

Abstract— In this study, we develop a simple daily assistive robot that controls its own vision according to linguistic instructions. The robot performs several daily tasks such as recording a user's face, hands, or screen, and remotely capturing images of desired locations. To construct such a robot, we combine a pre-trained large-scale vision-language model with a low-cost low-rigidity robot arm. The correlation between the robot's physical and visual information is learned probabilistically using a neural network, and changes in the probability distribution based on changes in time and environment are considered by parametric bias, which is a learnable network input variable. We demonstrate the effectiveness of this learning method by open-vocabulary view control experiments with an actual robot arm, MyCobot.

I. INTRODUCTION

Robots for daily assistive tasks have been developed in various forms [1]. They can perform tasks such as cooking [2] to cleaning [3] and serving [4]. In this study, we consider a simple daily assistive robot that controls its own vision according to linguistic instructions (Fig. 1). The robot performs several daily assistive tasks such as recording images of the user's face, hands, and screen, remotely capturing images of desired locations, illuminating the user's hands, and so on.

Several studies have been conducted on robotic view control. The most common one is on the planning task of "Next-Best-View" for autonomous 3D exploration of objects and environments [5], [6]. There are also studies on endoscope control in surgery [7] and gaze control in social robots [8]. In recent years, robots exploring a 3D space to find answers to linguistic questions have been studied [9]. There is a study on constructing 3D maps that include language information for the navigation of mobile robots [10]. There are also some examples of pick and place tasks based on linguistic instructions [11]–[13]. On the other hand, this study differs from previous tasks in that it controls the robot's view in a direction appropriate to the linguistic instructions. While it is true that some existing methods may perform view control implicitly, this study aims to develop a system that explicitly links linguistic instructions with physical information in order to achieve more precise and intentional view control.

In this study, we develop an open-vocabulary view control system using a low-cost low-rigidity robot arm. A web camera is attached to the arm-tip of MyCobot, a low-cost low-rigidity robot arm suitable for daily assistive tasks. The robot can perform actions based on linguistic instructions

Fig. 1. Open-vocabulary view control of a low-cost low-rigidity robot arm for daily assistive tasks. The lower figures show the image from the web camera attached to the arm-tip.

utilizing a pre-trained large-scale vision-language model that has been remarkably developed in recent years [14]–[16]. In addition, to ensure the performance of the low-cost low-rigidity robot, we introduce an experience-based learning mechanism using a neural network. The correlation between the visual information based on the vision-language model and the physical information of the low-cost low-rigidity robot is trained. In order to consider the stochastic correlation due to small changes in the visual field and the low-rigidity body, we construct a predictive model that outputs the mean and variance of sensor values. In addition, changes in the probability distribution of the visual information due to changes in time and environment are considered by parametric bias (PB) [17], which is a learnable network input variable. It is also possible to simply continue to collect and search images, but this would increase the cost of data management and memory, and would not capture its stochastic nature and changes in its probability distribution. Several experiments on actual robots show that the robot can respond to a variety of linguistic instructions and environments.

The structure of this study is as follows. In Section II, we describe the construction of the probabilistic model between physical and visual information, data collection and network training, update of parametric bias representing the changes in probability distribution, and open-vocabulary view control. In Section III, we describe simple quantitative evaluation experiments and more practical advanced experiments. In Section IV, we discuss the experimental results and some limitations of this study, and conclude in Section V.

[1] The authors are with the Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan. [kawaharazuka, kanazawa, obinata, k-okada, inaba]@jsk.t.u-tokyo.ac.jp

Fig. 2. The system overview including Vision-Language Model CLIP, Data Collector, Network Trainer, PB Updater, and Controller.

II. VIEW CONTROL LEARNING OF LOW-COST LOW-RIGIDITY ROBOT VIA LARGE-SCALE VISION-LANGUAGE MODEL

The overall system of this study is shown in Fig. 2. We call our network Stochastic Predictive Network with Parametric Bias (SPNPB). Information from the Vision-Language Model (VLM) and the robot's physical information are collected by Data Collector. SPNPB is trained by Network Trainer, and the Parametric Bias (PB) is updated online by Network Updater. SPNPB is used by Controller for open-vocabulary view control. Based on the network of [18], time-series information is removed and a large-scale vision-language model is applied.

The setup of MyCobot, a low-cost low-rigidity robot used in this study, is shown in the right figure of Fig. 2. A web camera is attached at the end of the arm to obtain an image V. From the arm, the joint angle θ can be obtained (the arm has six degrees of freedom, but only the first four are used in this study with limited angle ranges: {[-165, 165], [-45, 45], [-22.5, 0], [-22.5, 0]} [deg]). Although MyCobot cannot measure the joint torque, the necessary torque τ can be calculated from the current joint angle θ through its geometric model.

A. Stochastic Neural Network with Parametric Bias

SPNPB can be expressed by the following formula,

$$(s, \sigma) = h(u, p) \quad (1)$$

where s is the sensor state, σ is the variance of s under the assumption of Gaussian distribution, u is the control input, p is the parametric bias [17], and h expresses SPNPB. In this study, s denotes $\left(v^T \quad \tau^T\right)^T$ with a vector v ($\in \mathbb{R}^{512}$) where the current image V is transformed by a vision-language model CLIP [19], and τ denotes the joint torque required for the gravity compensation ($\in \mathbb{R}^4$). For u, the target joint angle θ ($\in \mathbb{R}^4$) is used. Note that the values of s and u are normalized using all obtained data points. Since σ represents the variance and must always be positive, the network outputs σ through exponential function. p is responsible for representing the change in the probabilistic distribution due to changes in time and environment, and is assumed to be two-dimensional in this study. SPNPB consists of four fully-connected layers. The number of units is set to $\{N_u + N_p, 100, 300, 500, 2N_s\}$ ($N_{\{u,s,p\}}$ is the number of dimensions of $\{u, s, p\}$). The activation function is hyperbolic tangent, and the update rule is Adam [20].

B. Training of SPNPB

We collect a dataset of s and u by random robot motions. By collecting data at different times of the day and in different environments, these differences can be embedded in the parametric bias as implicit information so that various data points with different distributions can be represented in a single model. In a series of motions in a trial k in the same environment, the dataset $D_k = \{(s_1^k, u_1^k), (s_2^k, u_2^k), \cdots, (s_{N_k}^k, u_{N_k}^k)\}$ is collected ($1 \leq k \leq K$, where K is the total number of trials and N_k is the number of data points for the trial k). Then, we create the dataset $D_{train} = \{(D_1, p_1), (D_2, p_2), \cdots, (D_K, p_K)\}$ for training. p_k ($1 \leq k \leq K$) is the parametric bias for the trial k, which is a variable with a common value during the trial but a different value for other trials. The dataset D_{train} and the following loss function are used to train SPNPB,

$$P(s_{i,n}^k | D_{k,n}, W, p_k) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{i,n}}} \exp\left(-\frac{(\hat{s}_{i,n}^k - s_{i,n}^k)^2}{2\hat{\sigma}_{i,n}}\right) \quad (2)$$

$$L_{likelihood}(W, p_{1:K} | D_{train}) = \prod_{k=1}^{K} \prod_{n=1}^{N_k} \prod_{i=1}^{N_s} P(s_{i,n}^k | D_{k,n}, W, p_k) \quad (3)$$

$$L_{train} = -\log(L_{likelihood}) \quad (4)$$

where P is the probability density function, $\{s, \sigma\}_i$ is $\{s, \sigma\}$ of the i-th sensor, $D_{k,n}$ is the n-th data point in D_k, W is the network weight of SPNPB, $\{\hat{s}, \hat{\sigma}\}$ is the value of $\{s, \sigma\}$ predicted from SPNPB using the dataset $D_{k,n}$, the current weight W, and the current parametric bias p_k for the trial k, and $p_{1:K}$ is a vector of p_k within $1 \leq k \leq K$. $L_{likelihood}$ denotes the likelihood function for W and p given D_{train}, and we consider the problem of maximizing it. This function is a modification of the loss function in [21]. We simplify the computation to the summation of $\log(P)$ by performing the transformation as in Eq. 4, making it the problem of minimizing L_{train}. In the usual training, only the network weight W is updated, but in this study, W and p_k are updated at the same time. Note that each p_k is optimized with an initial value of 0.

C. Update of Parametric Bias

By continuously updating the parametric bias p, the robot can always adapt to changes in time and environment. While

Fig. 3. The setup of the basic experiment. The upper figures show the changes in physical state (two changes in the angle of the web camera attached to the tip of the robot arm) and the lower figure shows the changes in environmental state (three changes in the arrangement of the five target objects).

Fig. 4. The arrangement of the trained parametric bias and its trajectory during the online update of parametric bias regarding three environmental and physical states in the basic experiment.

the robot is moving, we always collect the dataset D_{update} of s and u. The update of p starts when the number of the collected data N_{data}^{update} exceeds the threshold N_{thre}^{update}. Data exceeding N_{max}^{update} is discarded from the oldest ones. We use Eq. 2 – Eq. 4 as a loss function, and train SPNPB with N_{batch}^{update} and N_{epoch}^{update} as the number of batches and epochs, respectively. Here, the network weight W is fixed and only p is updated. By updating only p, we can avoid catastrophic forgetting and overfitting while adapting to the current environment. In this study, we set $N_{batch}^{update} = N_{data}^{update}$, $N_{thre}^{update} = 100$, $N_{max}^{update} = 200$, and $N_{epoch}^{update} = 3$, and use Momentum SGD [22] as the update rule.

D. Open-Vocabulary View Control

By using the trained SPNPB, it is possible to control the robot view from linguistic instructions. First, we prepare a linguistic instruction Q, and this is transformed into a latent vector q by using CLIP which is in the same latent space as v. Next, the initial value u^{init} of the control input u^{opt} to be optimized is determined, and the control input u^{opt} is updated by repeating the following process,

$$L_{control} = -q \cdot \hat{v} + C_\tau \|\hat{\tau}\|_2 \quad (5)$$

$$u^{opt} \leftarrow u^{opt} - \gamma \frac{\partial L_{control}}{\partial u^{opt}} \quad (6)$$

where $\{\hat{v}, \hat{\tau}\}$ is the mean of $\{v, \tau\}$ predicted from the current SPNPB and u^{opt}, C_τ is the weight of the loss function, and γ is the learning rate. The first term on the right-hand side of Eq. 5 is the loss to obtain the control input closest to the given linguistic instruction, and the second term on the right-hand side is the loss to reduce the required torque

of the posture. Eq. 6 updates u^{opt} using backpropagation technique and the gradient descent method. Here, the initial value u^{init} uses $N_{init}^{control}$ number of random u. By starting the optimization from various u^{init}, we can avoid the solution from falling into the local minima. Although γ can be a fixed value, in this study, each u^{opt} is updated by using $N_{batch}^{control}$ number of γ which are the logarithmically divided values in $[0, \gamma^{max}]$, and then u^{opt} with the smallest loss when running Eq. 5 is used repeatedly to obtain faster convergence. Eq. 5 – Eq. 6 are performed $N_{epoch}^{control}$ times, and the obtained u^{opt} is sent to the actual robot.

In this study, we set $N_{init}^{control} = 30000$, $N_{batch}^{control} = 100$, $N_{epoch}^{control} = 2$, $\gamma^{max} = 0.1$, and $C_\tau = 0.0001$.

III. EXPERIMENTS

First, we conduct a basic quantitative experiment in a small area surrounded by objects and walls. Next, we conduct an advanced experiment in a wider and more cluttered environment to demonstrate the effectiveness of the method.

A. Basic Experiment

The setup of the basic experiment is shown in Fig. 3. Five objects – 1. mug, 2. headphones, 3. bottle, 4. tissue box, and 5. clock – are arranged in front of the robot on a desk. By surrounding the desk with walls, the robot can uniquely determine the direction in which it should look at to face the object corresponding to each linguistic instruction. Parametric bias can embed not only environmental changes but also the physical changes of a low-rigidity robot. In this study, the angle of the web camera attached to the tip of the robot arm is changed to 0 degrees (state B0) and 30 degrees (state B1). Also, we prepare three environments, E0, E1, and E2, in which the positions of the five objects are shifted one by one as shown in the lower right figure of Fig. 3.

Fig. 5. The open-vocabulary view control of the basic experiment. The lower figures show the images from the web camera attached to the arm-tip.

Fig. 6. The comparison of view control errors among various neural network models: **PB+ST** - the proposed SPNPB, **ST** - stochastic predictive model without parametric bias, **PB** -the normal predictive model with parametric bias, and **None** - the general neural network without parametric bias, under three different environmental and physical states.

A dataset is collected for 100 seconds for each of the six states, which are combinations of two physical changes and three environmental changes. The joint angles are fed randomly, and data collection is performed at 10 Hz, obtaining a total of 6000 data points. SPNPB is trained based on this dataset. The arrangement of the trained parametric bias when applying Principle Component Analysis (PCA) is shown in Fig. 4. Note that the parametric bias for each state is expressed as E{0, 1, 2}-B{0, 1}. It can be seen that each PB is regularly arranged along the environmental and physical changes. Since no information on environmental and physical changes is given during the training, various state changes can be implicitly self-organized in the space of PB.

Next, we conducted an experiment in which parametric bias is updated online for three states, E0-B1, E1-B0, and E2-B1. The trajectories "-traj" when updating the parametric bias from random motions are shown in Fig. 4. The initial value of p is 0, and it can be seen that the parametric bias gradually approaches the appropriate value trained for the current physical and environmental conditions. In other words, the robot can gradually recognize the current state correctly even if its body and surrounding environment change.

Finally, we conducted an experiment of view control using the trained SPNPB. The experiment is performed for the state E1-B0 after the correct PB is recognized. Here, linguistic instructions of "Check the clock.", "Where are the headphones?", "See the tissue box.", "Find the mug.", and "Look at the bottle." are given in that order. The result of open-vocabulary view control is shown in Fig. 5. It can be seen that the robot's view is correctly controlled so that the object mentioned in the linguistic instruction fits into

the camera image. The results of comparative experiments of this view control for various physical and environmental conditions of E0-B1, E1-B0, and E2-B1, while changing the neural network model used, are shown in Fig. 6. The models used are SPNPB of this study (**PB+ST**), SPNPB without parametric bias (**ST**), SPNPB with the loss function Eq. 4 changed to a general mean squared error (**PB**), and a general neural network model excluding parametric bias and setting the loss function as mean squared error (**None**). For five target objects O, five linguistic instructions, "Look at the O.", "See the O.", "Find the O.", "Check the O.", and "Where is the O?" are used. The mean and variance of the distance between the camera's line-of-sight vector and the predefined location of the target object (the distance from a point to a line) in the 25 experiments are shown in Fig. 6. From the results, it is found that the error of view control in **PB+ST** is the lowest, while the accuracy in **ST**, **PB**, and **None** is much lower. In particular, the errors for E0-B1 and E1-B0 when using **ST**, **PB**, and **None**, are more than twice as large as those when using **PB+ST**.

B. Advanced Experiment

We conducted an advanced experiment in a setting closer resembling our living space. A monitor, a keyboard, a mug, a bottle, a tissue case, and various other objects were placed on the desk. We obtained data from random motions in various environments at different times of the day. The environments are divided into combinations of two states: one with or without a person sitting at a desk (Human or None), and one with all lights on (Bright) or some lights on (Dark). Data collection is performed at 10 Hz for 100 seconds for each of eight different time periods E0–E7, obtaining a total of 8000 data points. E0, E2, and E3 are Human/Bright cases, E1, E4,

Fig. 7. The arrangement of the trained parametric bias and its trajectory during the online update of parametric bias in the advanced experiment.

and E5 are None/Bright cases, E6 is a Human/Dark case, and E7 is a None/Dark case. SPNPB is trained based on this dataset. The arrangement of the trained parametric bias when applying PCA, and the corresponding environmental states are shown in Fig. 7. It is found that each PB is regularly arranged along the environmental changes, and the space of PB is implicitly self-organized.

Next, we conducted an experiment to update the parametric bias online. The trajectories "-traj" when updating parametric bias from random motion for a new Human/Bright environment E8 and a new None/Dark environment E9 are shown in Fig. 7. For both cases, we can see that p_k is gradually updated toward that of the same previously trained environment: in case of E8, toward p_k of Human/Bright environment E0, E2, and E3, and in case of E9, toward p_k of None/Dark environment E7. The same performance as that of Section III-A can be achieved even in a cluttered environment.

Finally, we conducted a view control experiment for the environment of E8 using the trained SPNPB and updated PB. Here, linguistic instructions of "Find the bottles.", "See the red chair.", "Where are my mugs?", "Check the bookshelf.", "See my hands.", "Watch my monitor.", "Look at the keyboard and mouse." are given in that order. The result of open-vocabulary view control is shown in Fig. 8. Similar to Section III-A, it is possible for the robot to view the objects and environments as indicated.

IV. Discussion

The obtained experimental results are discussed. First, we conducted basic quantitative experiments in a controlled environment where the direction of the target object is uniquely determined. The changes in the probability distribution of the network based on changes in the body and the environment can be regularly self-organized in the space of parametric

bias. By updating PB online, the current state of the body and the environment can be appropriately identified. The obtained SPNPB can then be used to direct the robot's gaze in the direction of appropriate objects based on linguistic instructions. It is found that this performance is achieved only when both the probabilistic predictive model and parametric bias are used, and when either one of them is not used, the control performance is reduced to about half. Next, we conducted advanced view control experiments in a cluttered environment that is more similar to that of daily life. By collecting and learning data at various times of the day with different parameters such as the presence of a human and the brightness of the room, these environmental changes become self-organized in the space of parametric bias. As in the basic experiment, the current environment is able to be understood appropriately, and the robot can perform view control for various objects and environments based on linguistic instructions. These results indicate that even with a low-cost low-rigidity body structure, and even with a large-scale vision-language model whose output changes with slight changes in the image, appropriate and adaptive view control is possible by considering the relationship between vision and body in a stochastic form, and by incorporating large changes in the form of parametric bias.

Limitations and future prospects of this study are described. First, this study mainly deals with view control, and it does not actually perform a task such as recording the image in response to the command "please record". Of course, such a command can be easily implemented by recognizing the word "record", but there is no limit to the variety of commands, such as "send an image via chat", "read the sentence out loud", or "turn on the lights". In the future, we would like to develop this system into a more practical system that automatically uses multiple APIs and view control according to linguistic instructions, using large-scale language models. Second, the system currently does not accept commands such as "a little more to the right" or "look at the back". This is because there is no embodiment in the large-scale vision-language model itself, which is an interesting issue to be addressed in the future. In addition, we would like to construct a system that can always keep moving by regularly collecting data and learning networks. For a more practical system, it is also necessary to consider obstacle avoidance and motion planning. Finally, although this study has dealt mainly with the two modalities of vision and body, we would like to extend these modalities in the future. We believe that if it becomes possible to handle not only images but also videos, sounds, and tactile sensations in the same way, it will be possible to perform a wider variety of tasks based on linguistic instructions. We would like to develop a system that acquires correlations among various sensors and grows autonomously.

V. Conclusion

This study has described the development of a low-cost low-rigidity robot system that performs daily assistive tasks through view control based on linguistic instructions. A

Fig. 8. The open-vocabulary view control of the advanced experiment. The lower figures show the images from the camera.

neural network is used to learn the correlation between the visual information from CLIP, one of the large-scale vision-language models, and the physical information including target joint angles and necessary joint torques of a low-rigidity robot. Here, its probabilistic correlations caused by small changes in the visual field and the low-rigidity body are considered by a predictive model with mean and variance network outputs. Changes in the correlations due to changes in time and environment are considered by parametric bias, which is a learnable network input variable. The actual robot experiments show that the robot can control its vision from appropriate motion commands according to linguistic instructions, and that open-vocabulary view control is possible even with a low-cost low-rigidity robot. In the future, we would like to consider the correlation among language, sound, image, tactile sensation, etc., to enable more advanced robot body control based on linguistic instructions.

REFERENCES

[1] K. Okada, T. Ogura, A. Haneda, J. Fujimoto, F. Gravot, and M. Inaba, "Humanoid Motion Generation System on HRP2-JSK for Daily Life Environment," in *Proceedings of the 2005 IEEE International Conference on Mechatronics and Automation*, 2005, pp. 1772–1777.

[2] M. Wächter, E. Ovchinnikova, V. Wittenbeck, P. Kaiser, S. Szedmak, W. Mustafa, D. Kraft, N. Krüger, J. Piater, and T. Asfour, "Integrating multi-purpose natural language understanding, robot's memory, and symbolic planning for task execution in humanoid robots," *Robotics and Autonomous Systems*, vol. 99, pp. 148–165, 2018.

[3] J. Kim, A. K. Mishra, R. Limosani, M. Scafuro, N. Cauli, J. Santos-Victor, B. Mazzolai, and F. Cavallo, "Control strategies for cleaning robots in domestic applications: A comprehensive review," *International Journal of Advanced Robotic Systems*, vol. 16, no. 4, pp. 1–21, 2019.

[4] M. Saito, H. Chen, K. Okada, M. Inaba, L. Kunze, and M. Beetz, "Semantic Object Search in Large-scale Indoor Environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Active Semantic Perception and Object Search in the Real World*, 2011.

[5] C. Connolly, "The determination of next best views," in *Proceedings of the 1985 IEEE International Conference on Robotics and Automation*, 1985, pp. 432–435.

[6] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding Horizon "Next-Best-View" Planner for 3D Exploration," in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation*, 2016, pp. 1462–1468.

[7] Y. He, B. Zhao, X. Qi, S. Li, Y. Yang, and Y. Hu, "Automatic Surgical Field of View Control in Robot-Assisted Nasal Surgery," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 247–254, 2021.

[8] A. Zaraki, D. Mazzei, M. Giuliani, and D. D. Rossi, "Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 157–168, 2014.

[9] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied Question Answering," in *Proceedings of the 2018 IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2018.

[10] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual Language Maps for Robot Navigation," arXiv preprint arXiv:2210.05714, 2022.

[11] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation*, 2018, pp. 3774–3781.

[12] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and Where Pathways for Robotic Manipulation," in *Proceedings of the 2021 Conference on Robot Learning*, 2021, pp. 1–13.

[13] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do As I Can and Not As I Say: Grounding Language in Robotic Affordances," arXiv preprint arXiv:2204.01691, 2022.

[14] F. Li, H. Zhang, Y. Zhang, S. Liu, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, "Vision-Language Intelligence: Tasks, Representation Learning, and Large Models," arXiv preprint arXiv:2203.01922, 2022.

[15] K. Kawaharazuka, Y. Obinata, N. Kanazawa, K. Okada, and M. Inaba, "Robotic Applications of Pre-Trained Vision-Language Models to Various Recognition Behaviors (in press)," in *Proceedings of the 2023 IEEE-RAS International Conference on Humanoid Robots*, 2023.

[16] K. Kawaharazuka, Y. Obinata, N. Kanazawa, K. Okada, and M. Inaba, "VQA-based Robotic State Recognition Optimized with Genetic Algorithm," in *Proceedings of the 2023 IEEE International Conference on Robotics and Automation*, 2023, pp. 8306–8311.

[17] J. Tani, "Self-organization of behavioral primitives as multiple attractor dynamics: a robot experiment," in *Proceedings of the 2002 International Joint Conference on Neural Networks*, 2002, pp. 489–494.

[18] K. Kawaharazuka, K. Shinjo, Y. Kawamura, K. Okada, and M. Inaba, "Environmentally Adaptive Control Including Variance Minimization Using Stochastic Predictive Network with Parametric Bias: Application to Mobile Robots," in *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021, pp. 8381–8387.

[19] 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," arXiv preprint arXiv:2103.00020, 2021.

[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015, pp. 1–15.

[21] S. Murata, J. Namikawa, H. Arie, S. Sugano, and J. Tani, "Learning to Reproduce Fluctuating Time Series by Inferring Their Time-Dependent Stochastic Properties: Application in Robot Learning Via Tutoring," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 4, pp. 298–310, 2013.

[22] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1, pp. 145–151, 1999.