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

Grasping Motion Generation for Deformable Objects under Dynamic Position Changes via Variance Prediction

Riko Kawata[1], Hyogo Hiruma[2,3], Hiroshi Ito[2], Tetsuya Ogata[2,4], Shigeki Sugano[1]

Abstract— **Because of labor shortages, robots are expected to provide work assistance in a variety of settings, including the home environment. At home we often deal with flexible objects, but flexible objects are characterized by their tendency to change position and shape. Because of this nature, data dealing with flexible objects involves uncertainty. Although deep learning has been used to perform a variety of complex tasks, the deterministic nature of conventional RNN makes it difficult to handle data with a probabilistic structure. In this study, we propose a method based on deep predictive learning that enables real-time motion generation and predicts the variance of joint angles, which facilitates learning of probabilistic structures and can handle dynamic changes. Experimental results show that the robot is able to generate motions that are adaptive to flexible objects with dynamic position changes.**

I. INTRODUCTION

Labor shortages has recently become a challenge for human society and robots are considered as a promising technology for human assistance. However, the variability and the complexity of the human environment has been a long standing challenge. One example is house works, which includes the use of deformable objects such as clothing and cables. Since such objects inherently have the characteristic to dynamically change through various causes, the robots are required to adaptively perform specific tasks. Therefore, to enable effective robotic assistance in tasks that involve handling such objects, it is essential for robots to adapt to dynamic environmental changes in real time.

Traditionally, robot motions were created based on rule-based methods. For example, techniques that specify the end-effector position [1] or utilize hand-crafted trajectories [2] have been widely adopted in industrial applications. While these methods are capable of executing fast or highly precise motions, they struggle to generalize to complex tasks. This is mainly because such tasks require the combination of multiple motion primitives, making the implementation costly and inflexible.

In recent years, deep learning techniques have lowered the costs of implementing complex robot motions [3]. This widened the area of robot application to more challenging task settings, such as opening doors [4] , operating mechanical zippers [5], or folding laundries [6]. Among these approaches, Deep Predictive Learning [7][8] has emerged as a promising method that can achieve real-time motion generation with a small amount of training data. This method trains neural network models to imitate human-demonstrated motion data that are collected via teleoperation. The robot then learns the associated sensorimotor sequences using Recurrent Neural Network (RNN). RNN models embed the dynamics of the demonstrated behaviors into its internal context, enabling robust generalization to environmental variations even with limited training data.

There are two key challenges when deploying such systems in real-world robotic applications. The first is accurate recognition of object position, which is essential for generating context-appropriate actions. Spatial Attention mechanisms [9] are employed to extract task-relevant information from images. Several attention-based visual processing approaches, such as methods for specifying the location of points of interest on a pixel-by-pixel basis [10] and methods for dynamically changing the object of attention based on the internal state of RNN [11][12], have been proposed to mimic human visual attention and improve perception performance.

The second challenge is adapting to environments with probabilistic and dynamic characteristics, such as those commonly found in human living spaces. In these situations, the robot must be capable of predicting and reacting to changes in real time. However, conventional RNN are inherently deterministic and tend to struggle in environments that require real-time adaptation. This often results in unstable behavior generation. To address this issue, models such as Stochastic Continuous-Time RNN (S-CTRNN) [13] and Predictive coding inspired Variational RNN (PV-RNN) [14] have been proposed, which incorporate stochasticity into the network dynamics.

Therefore, a model that addresses both challenges simultaneously—accurate attention and stable behavior in dynamic environments—is highly desirable. However, naively combining the two often leads to unstable attention point extraction and unreliable motion generation.

In this paper, we propose a novel model that extends the existing Spatial Attention RNN (SARNN) [9] by incorporating stochastic prediction mechanisms. To stabilize the learning of attention points, we further introduce a *stereo disparity constraint* which is an additional training loss to force the individually predicted attention points to have high consistency between stereo images. This design allows the

[1]Riko Kawata and Shigeki Sugano are with Department of Modern Mechanical Engineering, School of Creative Science and Engineering, Waseda University, Okubo 3-4-1, Shinjuku, Tokyo 169-8050, Japan kawata.riko@ruri.waseda.jp, sugano@waseda.jp

[2]Hyogo Hiruma, Hiroshi Ito and Tetsuya Ogata are with the Department of Intermedia Art and Science, Waseda University, Okubo 3-4-1, Shinjuku, Tokyo 169-8050, Japan

[3]Hyogo Hiruma is with the Research and Development Group, Hitachi, Ltd., Ibaraki 319-1292, Japan

[4]Tetsuya Ogata is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 169-8050, Japan ogata@waseda.jp

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

model to achieve both robust perception and adaptive motion generation under real-world uncertainty.

II. Related works

A. Motion Generation with Uncertainty Awareness

Methods using recurrent architecture have been used to resolve temporal uncertainty. Notable examples include Stochastic RNN [13], PV-RNN [14], and UF-RNN [15]. Stochastic RNN generate probabilistic behaviors by predicting both the mean and variance of future states. PV-RNN extends this capability by dynamically modulating the stochasticity of latent variables, allowing it to replicate the randomness inherent in training sequences. UF-RNN further advances uncertainty handling by internally simulating multiple future trajectories and guiding decisions toward those with minimal future variance. This mechanism enables exploratory actions under ambiguous conditions.

Despite these advances, none of the above models incorporate explicit attention mechanisms for perception alongside uncertainty modeling. Such integration is crucial for executing adaptive behaviors in complex environments.

In parallel, Diffusion Policy has been proposed to address multimodal action distributions [16]. It employs a denoising diffusion model that allows stochastic action sampling during inference. Attention-based architectures such as the Action Chunking Transformer (ACT) [17] have demonstrated high recognition performance using Vision Transformer (ViT) [18]. While effective, such models typically demand a vast amount of pretraining data, making them less suitable for tasks with limited demonstrations.

In contrast, our approach builds on deep predictive learning, which enables effective learning from small datasets while simultaneously addressing both uncertainty and attention-guided perception. This makes it particularly suitable for real-time applications in unstructured environments.

B. Motion Generation Using Stereo Images

Acquiring depth information is essential for robotic manipulation. Approaches to obtaining depth information include using RGB-D images [19], depth estimation on RGB images [20], and employing LiDAR sensors [21]. While all of these methods have demonstrated improvements in manipulation accuracy, they typically require pre-training depth estimation modules for a specific target domain.

In contrast, methods have been proposed that learn depth information from stereo images using weight-shared convolutional neural network (CNN) [22]. This approach eliminates the need for task-specific labeling or custom-designed modules, enabling the use of a generic model to handle a wide range of scenarios. While the use of stereo vision improves depth perception and enhances task performance, existing studies have not thoroughly evaluated its effectiveness under uncertain or dynamically changing environments.

Therefore, the novelty of our model lies in adapting stereo vision to operate robustly in uncertain environments.

III. Method

A. Model Learning with Variance Prediction

The Stochastic SARNN (S-SARNN) model proposed in this study is based on the framework of deep predictive learning. During model training, the robot's sensor time-series data, provided through demonstrations, are learned using a Long Short Term Memory (LSTM). Specifically, given sensor inputs at time step t, the model predicts the sensor values at time step $t+1$. The model utilizes images and joint angles as input modalities. For the image modality, it predicts attention points, as detailed in Section III-B. The model is trained to minimize the error between the predicted values and the ground truth at time $t+1$. During inference on a physical robot, the model predicts the next-step sensor values from the observed sensor inputs. The predicted joint angles are output as command values to the robot motors.

The proposed model, illustrated in Fig.1, extends the original SARNN architecture by incorporating stochastic properties inspired by S-CTRNN [13] and introducing a hierarchical LSTM structure [23]. Each component of the model is described in detail in the following subsections.

B. SARNN

The SARNN model is a model that identifies salient locations in visual features extracted by a CNN. Specifically, it outputs the spatial coordinates corresponding to the neuron with the highest activation per channel, treating them as attention points. These coordinates are then fed into an RNN model with current robot states, such as arm joint angles, to predict their future values. This enables robots to operate in accordance with the environment. Based on the predicted attention points, the system reconstructs the next image frame, thereby performing temporal prediction of visual information.

In the proposed model, stereo images were input into the SARNN model. The weights of the Convolutional Neural Network (CNN) layers in the SARNN modules that process the left and right images are shared. This enables the extraction of consistent features from both images while also reducing the number of trainable parameters, thereby making the learning process more lightweight [22].

C. Stochasticity

To introduce stochasticity, our model predicts both the mean and variance of future joint angles, following S-CTRNN. This is implemented using a Gaussian negative log-likelihood loss function. By placing the predicted variance in the denominator of the loss expression, the model naturally assigns lower weights to samples that are more difficult to predict, enabling uncertainty-aware learning.

On the other hand, variance prediction is not applied to image prediction. Introducing variance modeling into image prediction often makes training convergence more difficult. This is because the spatial attention points, although they highlight task-relevant visual information, do not necessarily correspond to fixed physical objects and

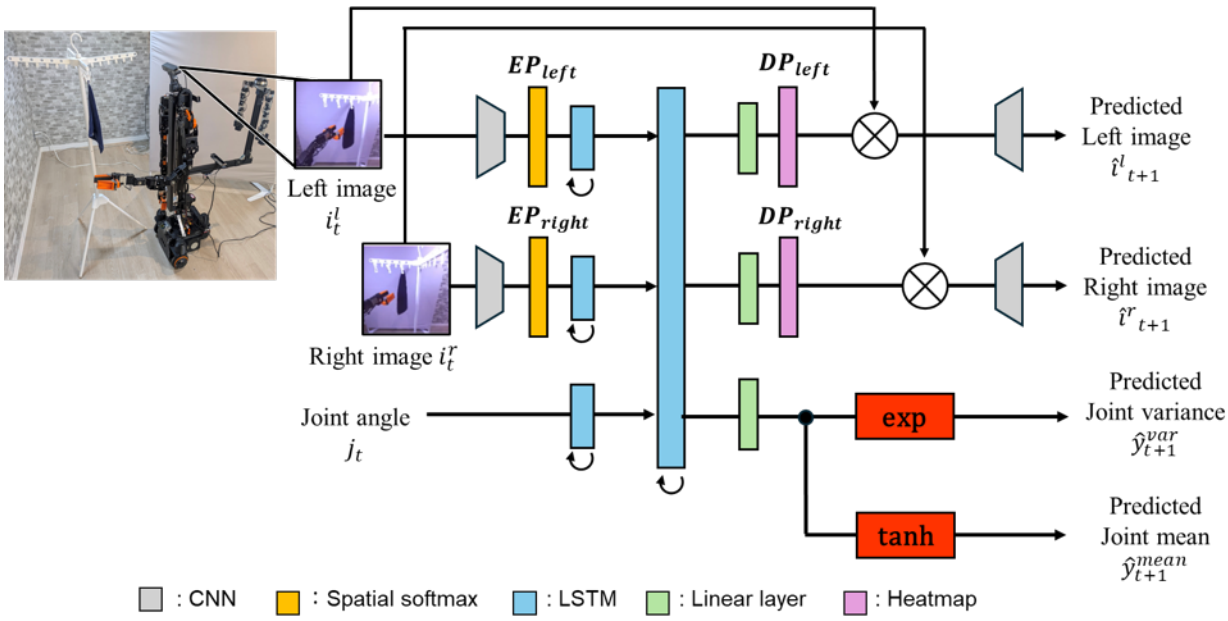

Fig. 1. The model structure of the proposed Stochastic SARNN. The model takes the input of two camera streams mounted on the head of the AIREC-Basic robot and the robot's joint angles. The attention points extracted from the images and the joint angles are each fed into separate lower-layer LSTM networks, and their outputs are then passed into a higher-layer LSTM. The outputs from the image pathway are referred to as Decoder points (DP). The DP is further processed with transposed CNNs to predict future images, while the outputs from the joint angle pathway are processed using a hyperbolic tangent function to obtain the mean prediction and an exponential function to obtain the variance prediction.

may vary dynamically across time steps, introducing instability in the predicted image features. In SARNN, the image processing module incorporates a spatial attention mechanism that outputs an attention map representing salient image features and predicts the next-step image features. To mitigate convergence issues, variance prediction is applied only to joint angles, not to image outputs.

D. Hierarchical LSTM

The hierarchical LSTM module is inspired by Multiple Timescales RNN (MTRNN) [24] and is designed to handle multi-timescale sensory information. It consists of a lower layer that directly processes high-frequency sensor signals and an upper layer that encodes slower, more abstract dynamics. Each modality—such as joint angles and visual features—has its own dedicated weights, and bidirectional information exchange occurs between layers. This design enables accurate modality-specific predictions while maintaining coherent integration across modalities. One of the key benefits of this structure is its resistance to training imbalance across different modalities.

E. Loss function

The basic loss function is defined as the sum of the following modality-specific losses. For the image data, we compute the reconstruction loss ($loss_I$) via mean squared error (MSE) between the demonstrated image at time $t+1$ $(y_{I_L}^{t+1}, y_{I_R}^{t+1})$ and the predicted image at time t $(\hat{y}_{I_L}^t, \hat{y}_{I_R}^t)$ (Equation 1). In addition, the MSE between the attention points extracted from the demonstrated image (Encoder points, EP) at time t and those predicted from the generated image (Decorder points, DP) at time t (Equation 2, 3). This loss enables the model to maintain consistency in the predicted attention points, leading to more reliable and robust predictions.

For the joint angle data, the loss is defined as the negative log-likelihood based on the demonstrated values $(y_J^t)$and the predicted mean $(\hat{y}_{Jmean}^t)$ and variance $(\hat{y}_{Jvar}^t)$ (Equation 4).

The total loss is then obtained by summing these individual losses (Equation 5). To balance the contribution of each modality during training, each loss term is weighted by a corresponding hyperparameter α. T denotes the batch size, and N denotes the number of steps within a single episode.

$$loss_I = \left(\mathrm{MSE}\left(y_{I_L}^{t+1}, \hat{y}_{I_L}^t\right) + \mathrm{MSE}\left(y_{I_R}^{t+1}, \hat{y}_{I_R}^t\right)\right) \times \alpha_I \quad (1)$$

$$loss_{P_L} = \mathrm{MSE}\left(\mathrm{Left}DP^t, \mathrm{Left}EP^{t+1}\right) \times \alpha_P \quad (2)$$

$$loss_{P_R} = \mathrm{MSE}\left(\mathrm{Right}DP^t, \mathrm{Right}EP^{t+1}\right) \times \alpha_P \quad (3)$$

$$loss_J = \left(\frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} -\frac{\ln(2\pi\hat{y}_{Jvar}^t)}{2} - \frac{(y_J^t - \hat{y}_{Jmean}^t)^2}{2\hat{y}_{Jvar}^t}\right) \times \alpha_J \quad (4)$$

$$loss_{basis} = loss_I + loss_{P_L} + loss_{P_R} + loss_J \quad (5)$$

F. Disparity Learning from Stereo Images

In the proposed model, stereo images are used as input. Since objects can undergo changes in three-dimensional space, depth information must be appropriately represented. To address this, disparity between RGB image pairs is leveraged to represent depth. The model incorporates two key mechanisms to achieve effective disparity-aware processing.

First, attention points are constrained to appear only within the overlapping regions where the same object is visible in

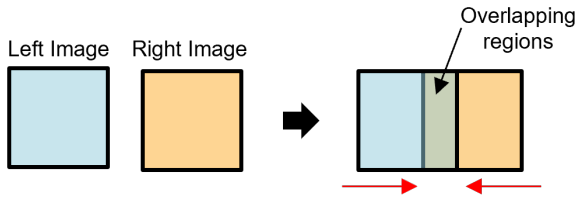

Fig. 2. Overlap the left and right images and calculate the MSE for the overlapping region.

both left and right images. In the camera images used in our experiments, 41 pixels from the right side of the left image and 41 pixels from the left side of the right image are utilized. The method for finding this region is shown in the Fig.2. This region was determined by sliding the left and right images over each other by one pixel at a time and identifying the displacement that minimizes the mean squared error between overlapping regions.

Second, a loss is introduced to encourage attention points to align at the same physical location for corresponding objects in the stereo image pair. Specifically, the MSE is computed between the X coordinates (EP_x, DP_x)or between the Y coordinates (EP_y, DP_y) of the attention points in the left and right images. (Equation 6-9). This loss is calculated for both the Encorder points and the Decorder points, and added to the basic loss function (Equation 10, 11).

$$loss_{DP_y} = \mathrm{MSE}\left(LeftDP_y, RightDP_y\right) \times \alpha_y \quad (6)$$

$$loss_{EP_y} = \mathrm{MSE}\left(LeftEP_y, RightEP_y\right) \times \alpha_y \quad (7)$$

$$loss_{DP_x} = \mathrm{MSE}\left(LeftDP_x, RightDP_x\right) \times \alpha_x \quad (8)$$

$$loss_{EP_x} = \mathrm{MSE}\left(LeftEP_x, RightEP_x\right) \times \alpha_x \quad (9)$$

$$loss_{stereo} = loss_{DP_y} + loss_{EP_y} + loss_{DP_x} + loss_{EP_x} \quad (10)$$

$$loss_{total} = loss_{basis} + loss_{stereo} \quad (11)$$

IV. Experiment

To evaluate the performance of the proposed model, we conducted real-world experiments on a laundry collection task. Two verifications were performed. In both verifications, the training data consisted of demonstrations in which laundry was grasped at three fixed positions. Data on moving objects is not provided.

First, we evaluated the model's positional generalization under static conditions. For physical robot execution, we evaluated the task success rate at five positions, including two unseen (untrained) locations, to assess generalization performance. Fig.3(a) shows the hanging positions of the laundry object.

Second, we examined the model's ability to adapt to previously unseen object motion. During physical execution, the object was moved during the task, and we evaluated the model's success rate under this dynamic condition. In this setup, the model is required to adapt its motion in response to environmental changes involving object displacement. Fig.3(b) illustrates the movement of the laundry object.

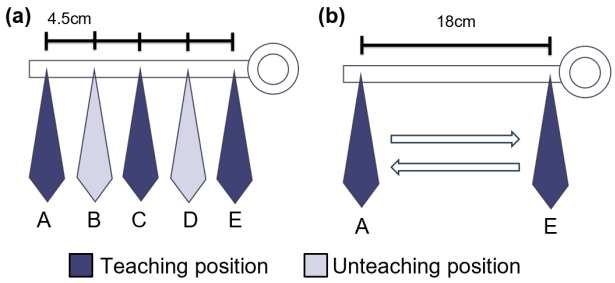

Fig. 3. Experimental Setup. (a) Setup for the static condition experiment, showing the hanging positions of the laundry labeled A–E from left to right. (b) Setup for the dynamic condition experiment.

A. Setup

For the experiments, we used the AIREC-Basic, a humanoid robot developed by Hitachi, Ltd. AIREC-Basic [25] is equipped with dual 8-DoF arms with grippers attached at the end-effectors, and a ZED 2 stereo camera (by Stereolabs) mounted on its head. Demonstration data were collected via teleoperation using a leader robot mounted behind the main robot.

In this experiment, we used 9-dimensional joint angle data (covering the left arm and gripper) and stereo camera images, each of resolution $64 \times 64 \times 3$ pixel . We collected 10 teleoperated demonstrations for each of the three hanging positions (positions A, C, and E in Fig.3(a)), resulting in a total of 30 trajectories. Each trajectory consisted of 200 timesteps recorded at 10 Hz. The trajectory generation rate and the actual position control frequency of the robot were also set to 10 Hz.

B. Training Procedures

The model was trained for 4000 epochs using the Adam optimizer with a learning rate of 0.0001 and a batch size of 7.

C. Comparative Baselines

To compare with the proposed S-SARNN model, we prepared five baseline models as listed in Table.I. Each model differs in terms of the number of input cameras, the presence or absence of variance prediction, and whether disparity constraints are applied.

TABLE I
List of Comparison Models

Model	Camera	Variance prediction	Disparity constraints
1 (SARNN)[9]	Monocular	×	-
2	Monocular	✓	-
3	Stereo	×	×
4	Stereo	✓	×
5	Stereo	×	✓
6 (proposed)	Stereo	✓	✓

V. Results and Discussion

A. Comparison of success rates

We first evaluate the proposed model by comparing task success rates. The success rates under static conditions and dynamic conditions – where the object is replaced during task execution –, are shown in Table II and Table III respectively. In the static condition, the experiment was repeated 20 times for each object hanging position. In the dynamic condition, 20 trials were conducted for each direction of object movement as described in Fig.3(b). A trial was considered successful if the robot successfully pinched and lifted the cloth between its grippers.

First, we compared our proposed model (Model 6) with the based SARNN model (Model 1). Under static conditions, both models performed evenly, achieving high success rates at all positions. However, under dynamic conditions, the proposed model significantly outperformed the baseline model.

To investigate the cause of difference in performance, we conducted an ablation study involving comparison models 2 through 5, as described in Table I. These models vary in model structure and training configurations, specifically (1) the use of stereo images, (2) predictive variance, and (3) disparity-based constraints.

The results of the ablation study are as follows:

- Model 2 : Stochastic SARNN, Monocular images
  In this model, the attention points (both EP and DP) appeared unstable. The positions of the attention points fluctuated significantly at each time step, resulting in a low task success rate. This suggests that the model failed to properly recognize the relevant visual features required for the task.
- Model 3 : SARNN, Stereo images
  This model struggled to learn meaningful attention points, leading to their concentration at a fixed location. As a result, it overfitted to joint angles and produced nearly identical motion trajectories regardless of the towel's position.
- Model 4 : Stochastic SARNN, Stereo images
  In this model, the attention points were stably expressed. The robot was able to adjust its reaching motion according to the towel's position; however, it failed to close the gripper successfully.
- Model 5 : SARNN, Stereo images, w/ Disparity Loss
  In this model, the attention points remained stable. While it achieved a high success rate under static conditions, its performance dropped significantly when the towel was moving.

From these findings, we conclude the following contributions of each model component:

- Number of cameras : Improves inference accuracy.
- Disparity constraint: Enhances the learnability and stability of stereo attention.
- Variance prediction: Enables robust adaptation to dynamic environmental changes.

These results demonstrate the effectiveness of the proposed model in both static and dynamic manipulation tasks.

TABLE II
Results of grasping motion for static objects

Model	Number of Successes					Success rate
	A	B	C	D	E	%
1	18	19	19	18	12	86
2	2	1	0	1	3	7
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	16	18	15	15	13	77
6	**17**	**19**	**19**	**18**	**12**	**86**

Fig. 4. The visualization of the predicted attention points at different time steps in model 2. Blue markers indicate the EP, and red markers represent the DP.

B. Analysis of RNN Hidden States

To examine the effect of variance prediction, we compared Model 5 and the proposed model (Model 6) under a dynamic condition where the object is in motion. The PCA results for each module (i.e., lower-layer RNN) are shown in Fig. 5.

Fig. 5 illustrates the temporal evolution of the RNN hidden states for each model. Specifically, Principal Component Analysis (PCA) was applied to the hidden states of the lower-layer RNNs during timesteps 60 to 110 (when reaching for a towel) of the predicted motions, and the results are visualized.

For Model 5, the PCA trajectory of the joint-angle module followed a similar path regardless of the object's movement direction, while the PCA trajectory of the image module appeared disordered and unstable, indicating a failure in visual recognition. Due to this recognition failure, the robot executed the gripper's closing actions at positions where it could not grasp the towel during the reach operation.

In contrast, for the proposed model, the joint-angle hidden states trajectories clearly branched depending on the object's movement direction, and the image hidden state maintained a consistent and coherent trajectory without significant fluctuation.

These results suggest that the proposed model can adapt to object movement through the effect of variance prediction. By predicting variance, the model is able to tolerate discrepancies between the lower-level RNNs and still output appropriate actions.

C. Stability Analysis of Attention Points

To evaluate the effectiveness of the disparity constraint, we measured the average distance between attention point coordinates in the left and right images. For each timestep, we identified the closest pairs of attention points between the stereo image pair—specifically, the 10 pairs with the smallest Euclidean distances—and computed the mean distance across these 10 pairs.

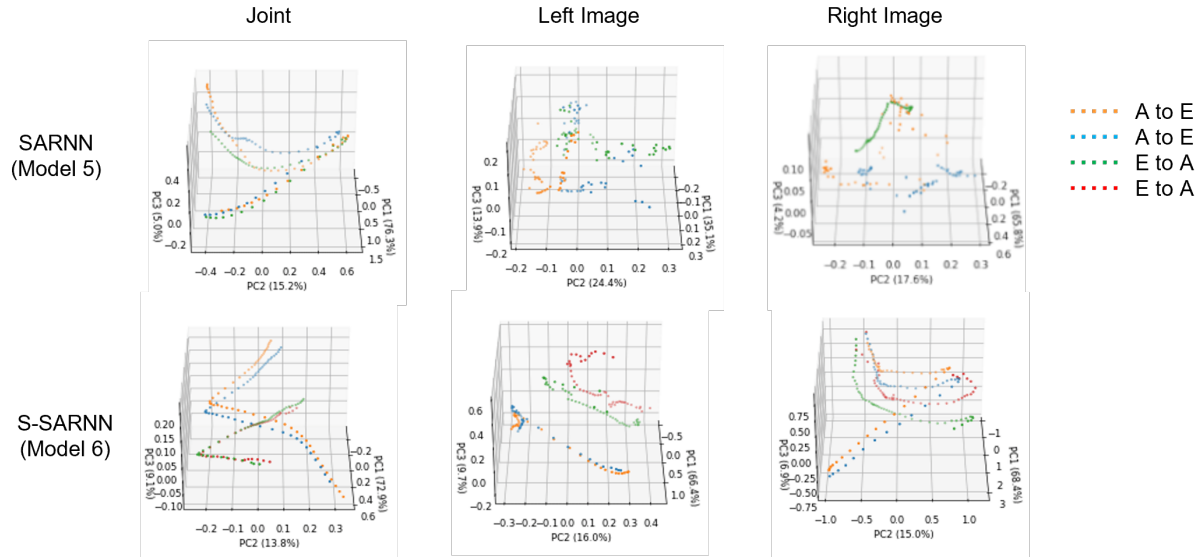

Fig. 5. We performed Principal Component Analysis (PCA) on the hidden states of the lower-layer RNNs in Model 5 and Model 6. This analysis aimed to investigate the impact of incorporating variance predictione by comparing how the internal states evolved over time.

TABLE III

RESULTS OF GRASPING MOTION FOR DYNAMIC OBJECTS

Model	A to E		E to A		Success rate
	success	failure	success	failure	%
1	5	15	1	19	15.0
2	0	20	3	17	7.5
3	0	0	0	0	0
4	0	0	0	0	0
5	7	13	0	20	17.5
6	**15**	**5**	**13**	**7**	**70.0**

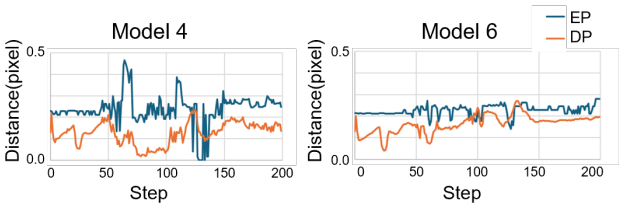

Fig. 6. Temporal variation of the distances between attention points on stereo images. A comparison between Model 4 and Model 6. The orange line denotes EP, and the blue line denotes DP.

The temporal progression of this average distance is shown in Fig.6, and the overall average across the entire task duration is summarized in Table IV.

For EP, no significant difference was observed between the two models. However, for DP, the proposed model (Model 6) achieved smaller distances between attention points, indicating better consistency. DP represents the predicted outcomes of the RNN. A high degree of deviation in the RNN's predictions indicates that the overall task dynamics are not being accurately captured. Model 4 exhibited a sharp increase in attention point distances around the 70th and 110th timestep, which corresponded to critical decision points such as object replacement and the initiation of grasping actions.

Based on the above, it can be concluded that incorporating the disparity constraint stabilizes the distance between decoder attention points. Without the disparity constraint, the distance between attention points increases sharply at the moment the gripper opens and closes. A large attention-point distance indicates a failure to correctly recognize the same object in both images, suggesting that maintaining a consistently small distance between attention points contributes to task success. Therefore, introducing the disparity constraint is essential for stabilizing attention and improving task performance.

TABLE IV

MEAN DISTANCE BETWEEN ATTENTION POINTS

Model	Encorder point	Decorder point
4	0.2567	0.2764
6	0.2469	0.1574

VI. CONCLUSION

In this study, we proposed a novel model S-SARNN, which is characterized by its ability to predict not only the mean but also the variance during joint angle prediction. This model is also characterized by its use of stereo images and the application of disparity constraints.

We evaluated the proposed model on a real-world laundry-grasping task. Experiment results demonstrated that, unlike conventional deterministic RNN models, S-SARNN was capable of performing the task with high success rates even in previously unseen situations, such as when the object was moving. These results confirm that predicting variance allows the model to tolerate uncertainties such as object motion and to appropriately switch behaviors.

As of limitations, the proposed model still suffers in handling situations in which the object moves randomly, such as

hanging clothes that are swaying by the wind. This is because this model passively decides on actions after a change in the situation occurs. In future work, we are planning to address this by extending our model to proactively predict changes.

ACKNOWLEDGEMENT

This work was supported by JST [Moonshot R&D] Grant-Number [JPMJMS2031].

REFERENCES

[1] W. J. Wilson, C. C. Williams Hulls, and G. S. Bell, "Relative end-effector control using Cartesian position based visual servoing," IEEE Transactions on Robotics and Automation, vol. 12, no. 5, pp. 684-696, 1996.

[2] Ken Ichiryu, Haruo Watanabe, Tadahiko Nogami, Ichiro Nakamura, and Masakatsu Fujie, REALIZATION OF BIPED ROBOT BY HYDRAULIC DRIVE, Proceedings of the JFPS International Symposium on Fluid Power, vol.1989, no.1, pp.421-428, 1989.

[3] Yang Pin-Chu, Sasaki Kazuma, Suzuki Kanata, Kase Kei, Sugano Shigeki, and Ogata Tetsuya, Repeatable Folding Task by Humanoid Robot Worker Using Deep Learning, IEEE Robotics and Automation Letters, vol.2, no.2, pp.397-403, 2017.

[4] Ito Hiroshi, Yamamoto Kenjiro, Mori Hiroki, and Ogata Tetsuya, Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control, Science Robotics, vol.7, 65:eaax8177, 2022.

[5] Ichiwara Hideyuki, Ito Hiroshi, Yamamoto Kenjiro, Mori Hiroki, and Ogata, Tetsuya, Contact-Rich Manipulation of a Flexible Object based on Deep Predictive Learning using Vision and Tactility, 2022 International Conference on Robotics and Automation (ICRA), pp.5375-5381, 2022.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail and Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang and Ury Zhilinsky, π_0: A Vision-Language-Action Flow Model for General Robot Control, 2024.

[7] Hiroshi Ito, Kenjiro Yamamoto, Hiroki Mori, and Tetsuya Ogata, Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control, Science Robotics, 2022.

[8] Kanata Suzuki, Hiroshi Ito, Tatsuro Yamada, Kei Kase, and Tetsuya Ogata, Deep Predictive Learning: Motion Learning Concept inspired by Cognitive Robotics, arXiv preprint arXiv:2306.14714, 2024.

[9] Hideyuki Ichiwara, Hiroshi Ito, Kenjiro Yamamoto, Hiroki Mori, and Tetsuya Ogata, Spatial Attention Point Network for Deep-learning-based Robust Autonomous Robot Motion Generation, arXiv preprint arXiv:2103.01598, 2021.

[10] Finn Chelsea, Tan Xin Yu, Duan Yan, Darrell Trevor, Levine Sergey, and Abbeel Pieter, Deep spatial autoencoders for visuomotor learning, 2016 IEEE International Conference on Robotics and Automation (ICRA), 512-519, 2016.

[11] Hyogo Hiruma, Hiroshi Ito, Hiroki Mori, and Tetsuya Ogata, Deep Active Visual Attention for Real-time Robot Motion Generation: Emergence of Tool-body Assimilation and Adaptive Tool-use, 2022 IEEE/RAS International Conference on Intelligent Robots and Systems (IROS 2022), Kyoto, Japan, October 23-27, 2022.

[12] Hyogo Hiruma, Hiroshi Ito, Hiroki Mori, and Tetsuya Ogata: A3RNN: Bi-directional Fusion of Bottom-up and Top-down Process for Developmental Visual Attention in Robots, 13th IEEE International Conference on Development and Learning 2025 (ICDL 2025), accepted for Oral Presentation, Prague, Czech Republic, September 16-19, 2025.

[13] Murata Shingo, Namikawa Jun, Arie Hiroaki, Sugano Shigeki, and Tani Jun, "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.

[14] Ahmadi Ahmadreza, and Tani Jun, A novel predictive-coding-inspired variational RNN model for online prediction and recognition, Neural computation, vol.31, no.11, pp.2025-2074, 2019.

[15] Hyogo Hiruma, Hiroshi Ito, and Tetsuya Ogata: UF-RNN: Real-Time Adaptive Motion Generation Using Uncertainty-Driven Foresight Prediction, 2025 IEEE/RAS International Conference on Intelligent Robots and Systems (IROS 2025), Hangzhou, China, October 19-25, 2025.

[16] Cheng Chi and Zhenjia Xu and Siyuan Feng and Eric Cousineau and Yilun Du and Benjamin Burchfiel and Russ Tedrake and Shuran Song, Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, 2024.

[17] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn, Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, arXiv preprint arXiv:2304.13705, 2023.

[18] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, and others, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

[19] Zhang Yinda, and Funkhouser Thomas, Deep depth completion of a single rgb-d image, Proceedings of the IEEE conference on computer vision and pattern recognition, 175-185, 2018.

[20] Chen Zhao, Badrinarayanan Vijay, Drozdov Gilad, and Rabinovich Andrew, Estimating depth from rgb and sparse sensing, Proceedings of the European conference on computer vision (ECCV), 167-182, 2018.

[21] Park Kihong, Kim Seungryong, and Sohn Kwanghoon, High-Precision Depth Estimation with the 3D LiDAR and Stereo Fusion, 2018 IEEE International Conference on Robotics and Automation (ICRA), 2156-2163, 2018.

[22] Xianbo Cai, Hiroshi Ito, Hyogo Hiruma, and Tetsuya Ogata, "3D Space Perception via Disparity Learning Using Stereo Images and an Attention Mechanism: Real-Time Grasping Motion Generation for Transparent Objects," IEEE Robotics and Automation Letters, vol. 9, no. 12, pp. 11857-11864, Dec. 2024.

[23] Hyogo Hiruma, Hiroshi Ito, Hiroki Mori, and Tetsuya Ogata, "Deep Active Visual Attention for Real-time Robot Motion Generation: Emergence of Tool-body Assimilation and Adaptive Tool-use," IEEE Robotics and Automation Letters. Preprint Version, June, 2022.

[24] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment", PLoS Computational Biology, 4, 11, pp.e000220-1–e1000220-18, 2008.

[25] Ito Hiroshi, Kanai Yoshiki, Kanazawa Akira, Ichiwara Hideyuki, Yoshida Takahiro, Noguchi Naoaki and Ogata Tetsuya "AIREC-Basic: Consistent Demonstration Data Collection for Imitation Learning with Redundant Robot Arms." 2026 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2026.