2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Multi-task real-robot data with gaze attention for dual-arm fine manipulation

Heecheol Kim[1,2], Yoshiyuki Ohmura[1], Yasuo Kuniyoshi[1]

***Abstract*— Deep imitation learning is a promising approach in robotic manipulation, enabling robots to acquire versatile and adaptable skills. In such research, by learning various tasks, robots achieved generality across multiple objects. However, such multi-task robot datasets have mainly focused on single-arm tasks that are relatively imprecise and not addressed the fine-grained object manipulation that robots are expected to perform in the real world. In this study, we introduce a dataset for diverse object manipulation that includes dual-arm tasks and/or tasks that require fine manipulation. We generated a dataset containing 224k episodes (150 hours, 1,104 language instructions) that includes dual-arm fine tasks, such as bowl-moving, pencil-case opening, and banana-peeling. This dataset is publicly available [1]. Additionally, this dataset includes visual attention signals, dual-action labels that separate actions into robust reaching trajectories or precise interactions with objects, and language instructions, all aimed at achieving robust and precise object manipulation. We applied the dataset to our Dual-Action and Attention, which is a model that we designed for fine-grained dual-arm manipulation tasks that is robust to covariate shift. We tested the model in over 7k trials for real robot manipulation tasks, which demonstrated its capability to perform fine manipulation.**

***Index Terms*— Imitation Learning, Deep Learning in Grasping and Manipulation, Perception for Grasping and Manipulation**

I. Introduction

Fine manipulation, which involves the precise control of delicate objects, is an important research topic because robots are expected to replace fine motor skills traditionally performed by humans in the near future. Deep imitation learning [1], which learns state-action pairs from expert demonstration data via deep neural networks, is a promising approach for fine manipulation because it directly maps a robot's sensory observation to an action, thereby eliminating the need for engineering features and manipulation skills by robot engineers and enabling learning even in cases that involve an unknown environment or object models.

In recent studies on deep imitation learning, researchers have aimed for the generalization of robot manipulation skills by scaling up the dataset size [2], [3], [4]. These efforts have culminated in the release of [5], a comprehensive dataset which aggregates over one million episodes from these studies. This large-scale dataset helps efficient research in the robotics community by eliminating the need for individual, time-consuming data generation. However, in many of these studies, researchers primarily focused on the single-arm grasping and placing of objects (e.g., [2], [3]), which lack fine manipulation skills.

To address this issue, we generated and released a robot manipulation dataset containing 224k demonstration episodes that includes fine manipulation and dual-arm skills. This dataset includes data for fine manipulation, such as dual-arm pencil case opening and manipulation of various deformable objects (e.g., thread, banana, and handkerchief), in addition to the manipulation of small and thin objects, such as small bolts, thin coins, and plastic cards. It also includes a human expert's gaze coordinates (i.e., the pixel location of the gaze on the camera image) measured using an eye tracker during demonstrations and dual-action labels to help learning fine manipulation.

To learn fine manipulation, we used and improved the deep learning architecture by implementing Dual-Action and Attention (DAA), which was initially proposed in [6] and later enhanced in [7]. In imitation learning, the attention mechanism contributes to a policy inference that is robust to covariate shift – a scenario in which training and test domain distributions are different [8] – by extracting manipulation-related information from the robot's sensory inputs. DAA is composed of the attention mechanism and dual actions. Firstly, visual attention [9] mimics the human gaze mechanism to suppress visual disturbances caused by task-unrelated pixels. Secondly, somatosensory attention [10] is a structure designed to address covariate shift in extended robot embodiments, such as dual-arm, and is based on the Transformer architecture. Finally, dual-action, inspired by physiological studies [11] of human visuomotor control, models a global-action for robust trajectory generation against compounding errors, and a local-action for precise object manipulation, thus simultaneously achieving robustness and fine manipulation skills. This model facilitates the learning of complex manipulation tasks, such as needle-threading [6] and banana-peeling [7].

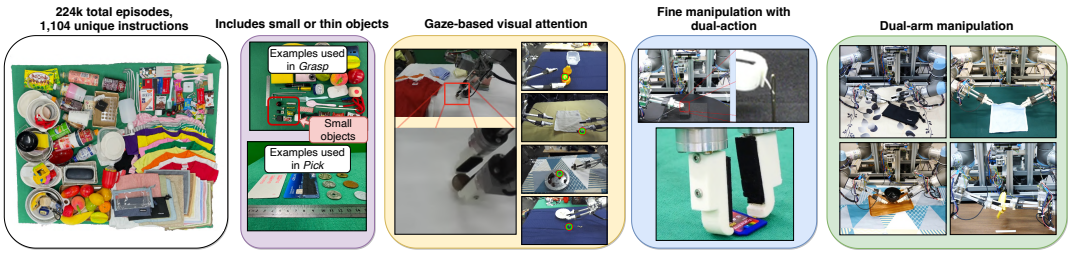
In this research, DAA shows that a variety of fine manipulation tasks are possible under various environmental conditions, thereby proving its generality. This model was validated using various tasks, such as opening pencil cases, moving large bowls with both hands, grasping various objects, picking up thin objects, such as coins, and folding towels.

[1] Laboratory for Intelligent Systems and Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan (e-mail: {h-kim, ohmura, kuniyosh}@isi.imi.i.u-tokyo.ac.jp, Fax: +81-3-5841-6314)

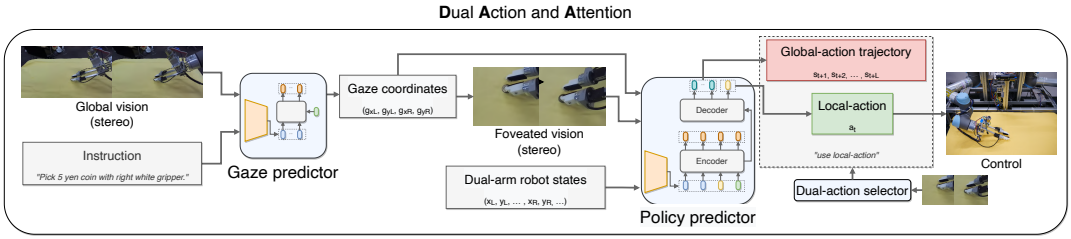
[2] Corresponding author

This study was supported in part by the Department of Social Cooperation Program "Intelligent Mobility Society Design," funded by Toyota Central R&D Labs., Inc., of the Next Generation AI Research Center, The University of Tokyo.

[1]The dataset and robot model are available at `https://sites.google.com/view/multi-task-fine`



(a) The dataset in this study is a fine-grained robot manipulation dataset that includes dual-arm tasks and a variety of objects, including deformable, small, and thin objects. This dataset includes the gaze of a human demonstrator as a visual attention signal as well as dual-action labels.



(b) Dual-Action and Attention uses a gaze predictor to efficiently focus on only high-resolution pixels related to the task, thereby enabling manipulation that is robust to task-unrelated objects and background noise. The policy predictor can perform fine-grained manipulation using a dual-action mechanism that simultaneously achieves the generation of a globally robust trajectory (global-action) and precise object manipulation (local-action).

Fig. 1: Dataset and outline of Dual-Action and Attention.

Fig. 2: Visual examples of tasks in the dataset.

The contributions of this paper are as follows: First, we create a multi-task imitation learning dataset that includes dual-arm, fine manipulation skills and gaze-based visual attention information, and make it publicly available. Second, we demonstrate that dual-action and attention mechanisms enable the multi-task dual-arm fine manipulation skills in the dataset to be performed.

II. Related Work

Deep imitation learning, particularly behavior cloning (e.g., [1]), directly maps a robot's sensory observation to an action in a model-free manner, which makes them applicable to unknown environments and object dynamics. In studies such as [1], [7], [6], researchers successfully implemented imitation learning that targets fine-grained tasks, such as opening a lid or manipulation with chopsticks. However, in these studies, training was often specialized for one task, which left the expected versatility and adaptability of future robots to new environments/tasks unassessed. Fu *et al.* [13] recently upgraded the robot system originally developed by [1] to a mobile manipulator for dual-arm fine manipulation, such as pushing chairs or cooking; however, their dataset lacks scale (50 demonstrations per task) and does not include visual attention labels, which we consider to be essential for fine manipulation tasks that require high-resolution vision.

Regarding versatility and adaptability, learning multi-task manipulation skills using a large robot dataset is a topic that has been extensively researched recently. In many studies, researchers were encouraged by the advancements in language models [14] and adopted imitation learning conditioned on natural language instructions as a breakthrough in multi-task learning. Brohan *et al.* [2] showed that using a dataset containing 130k language-instructed episodes and various tasks enabled generalization to various backgrounds and objects in object manipulation, and focused mainly on grasping[2], placing, and moving objects. In a subsequent study, Brohan *et al.* [3] further demonstrated that the use of a large vision-language model enabled the emergence of task reasoning for unseen objects. In another study, a single-arm robot was used [4], and data from 60.1k episodes across 24 environments with open vocabulary were collected and trained, which encompassed 13 skills in various scenes, such as stacking and folding clothes. Fang *et al.* [12] created a multi-robot/multi-modal dataset containing 110k episodes of contact-rich, single-arm tasks, such as cutting or plugging, accompanied by language descriptions. To the best of our knowledge, publicly available robot manipulation datasets

[2]In this research, *pick* refers to the picking up of thin objects from a table; hence, *grasp* is used to represent what is conventionally referred to as *pick*.

Dataset	# episodes	Dual-arm	Visual attention signal	Example skills
RT-1 [2]	130k	✗	✗	Grasp, place, move, knock over, open/close drawer
RH20T [12]	110k	✗	✗	Cut, plug, pour, fold
BridgeData V2 [4]	60.1k	✗	✗	Stack, fold clothes, sweep granular materials
Mobile Aloha [13]	$<$ 1k	✓	✗	Push chairs, move bowls, wipe, push buttons
DAA (ours)	224k	✓	✓	Pick up coins, open zippers, thread needles, peel bananas

TABLE I: The DAA dataset is an at-scale dataset for dual-arm fine manipulation, including signals for learning fine manipulation tasks.

at scale do not yet include examples of dual-arm fine manipulation (refer to Table I), such as the grasping of sub-centimeter-level objects, picking up thin objects, or opening zippers placed on a table, which are difficult to grasp because of their small size. Our DAA dataset, accompanied by the relatively large 224k episodes,[3] fills this gap by featuring dual-arm, fine manipulation skills, and incorporating visual attention signals and dual-action labels to help learning.

III. Dual-Action and Attention

DAA is based on the architecture proposed in [7]. In this section, we explain each element of the architecture and then clarify the modifications made in the present study.

A. Attention-Based Imitation Learning

In the context of robot manipulation, much of the information input from the robot's sensors, such as background pixels, is irrelevant to the task. Visual and somatosensory attention mechanisms have been proposed as a method to minimize the covariate shift caused by such information.

First, **visual attention** is an architecture inspired by the human retina structure and gaze that was designed to extract task-relevant information from a robot's vision. The human retina comprises a narrow central region of *foveated vision* that receives high-resolution color information and *peripheral vision* that captures low-resolution grayscale information [15]. During object manipulation, the human gaze strongly correlates with the object [16] using high-resolution information from foveated vision for manipulation [11].

Our demonstration framework includes an eye tracker mounted on the HMD that measures the x, y coordinates of where both of the human expert's eyes are looking on the screen. Based on this information, we train a gaze predictor to predict the gaze location in *global vision* [4] that a human is likely to look at while performing a task. The policy predictor receives foveated vision, that is, the high-resolution pixels around the gaze position predicted by the gaze predictor, as its vision input. Thus, the policy predictor can only receive task-related object information included in foveated vision while effectively suppressing the covariate shift caused by changes in the background or other objects in global vision. Additionally, foveated vision allows for a significant reduction in the total number of pixels ($1280 \times 720 \rightarrow 256 \times 226$) while maintaining the high resolution in the pixel area necessary for the task, thereby enabling precise object manipulation skills. By contrast, in conventional robot learning (e.g., [2]), the global vision is resized to a low resolution (300×300). Using a high-resolution global vision to maintain detail will require excessive computational resources. Because of these advantages, the visual attention mechanism has been used in multi-object manipulation [9], [17] and fine manipulation [6], [7].

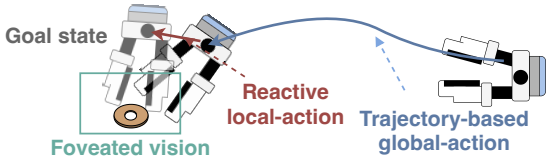
Second, **somatosensory attention**, which refers to the attention on somatosensory sensory inputs like the robot's state, was proposed for efficient information processing, particularly in dual-arm manipulation [10] and manipulation with force/torque sensory information [18]. This approach addresses covariate shifts caused by variations in a robot's limb positions between the training and testing phases. Unlike visual attention, which can be directed by the human gaze, somatosensory attention lacks a clear external signal. Therefore, in [10], the researchers used a Transformer-based self-attention mechanism to effectively process somatosensory inputs. In this research, the policy predictor's encoder corresponds to that mechanism.

B. Dual-Action

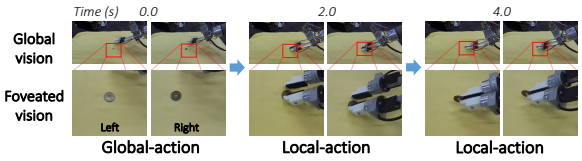
Dual-action, inspired by the human visuo-motor cognitive system [11], was proposed for precise object manipulation in [6] and modified in [7]. This human system consists of two main parts: the *movingness system* and *displacement system*. The movingness system ballistically moves the hand rapidly around the target. By contrast, the displacement system uses information from foveated vision to achieve more detailed and accurate adjustments, thereby detecting positional errors. This dual system is fundamental to human visuomotor control and can be applied to robot object manipulation skill learning. In robot learning, it is divided into global-action and local-action, where each is responsible for fast movement that approaches the vicinity of the object and precise manipulation (Fig. 3a). Specifically, in global-action, inspired by the ballistic movingness system, the trajectory from the current end-effector position to the starting point of local-action is predicted and executed. This is robust to the compounding error caused by the repetitive accumulation of small errors because the entire trajectory is predicted. Local-action is a reactive action that starts around the object and manipulates it. This allows for an immediate response to

[3]Each episode is structured to encompass only one semantic task. For instance, in a robot's grasp-and-place action, "grasp" and "place" are considered to be separate episodes. As a result, the number of episodes cannot be directly compared with those in other datasets, such as RT-1.

[4]This corresponds to human peripheral vision, but we have named it global vision because it also includes the central area (i.e., low-resolution foveated vision).



(a) Dual-action separates the entire robot trajectory in both the demonstration set and during inference by the model into trajectory-based global-action and reactive local-action.



(b) The gaze predictor uses low-resolution global vision to predict the gaze; hence, the policy predictor can use the high-resolution foveated vision of the target pixels for fine-grained manipulation. The action shifts from global to local when the end-effector is observed in foveated vision.

Fig. 3: Explanation of dual-action and visual attention.

changes in the dynamics during object manipulation, which contributes to the success of delicate object manipulation. The combination of global and local-action results in robustness against compounding errors while achieving precise object manipulation.

In this paper, we define the action shift from global-action to local-action as occurring when the end-effector is observed within foveated vision (Fig. 3b). This switching strategy is predicated on the assumption that the important interaction between the end-effector and the object, which is responsible for local-action, only occurs in foveated vision, which is selected by the gaze because the human's eye gaze is concentrated on the target object during manipulation [16].

IV. Data Generation

The dataset was generated by human demonstrators remotely operating a robot to perform various tasks. A *phase* is a cycle period from when the human teacher starts the remote operation to when they take a break. We repeated tens to hundreds of demonstration episodes per phase. Multiple tasks can exist in one phase. For example, in the phase of grasping nuts scattered on a table and placing them on a white plate, the instruction "grasp nut with right white gripper'' and "place nut on white plate with right white gripper" were repeated. In a single phase, background patterns (using tablecloths) and lighting conditions were changed multiple times. Foam floor tiles were placed on the surface of the table to prevent the robot from triggering emergency stops too frequently when it came into contact with the table, thereby enhancing safety during operation.

The dataset in this study includes, in addition to vision and robot action data, the following information. First, the **gaze** signal uses the eye tracker to measure the gaze information of both eyes of the demonstrator. This information is important because it is highly correlated with the target object in the task. Second, the **dual-action label** in the dataset is designated as follows: a value of 0 indicates a global-action, and a value of 1 indicates a local-action. To simplify dataset annotation, an automatic labeling method, similar to [7], was used to define the action shift from global to local-action when the end-effector enters foveated vision.

Finally, in this study, **language instructions** were used to provide clear directions for the tasks. Language instructions specify the target object and action, in addition to which robot arm to use: left, right, or both. Language instructions were annotated by humans for all tasks.

The dataset created in this study was divided into 11 groups according to the nature of the tasks. The task objectives and statistics of the dataset can be found in Table II. The tasks targeted by this dataset are challenging. *Grasp-and-place* mainly involves grasping small objects (e.g., bolts and candy) at the level of a few centimeters or sub-centimeters, where a few millimeters of error can lead to failure, which makes the grasping task difficult. In *pick*, picking up thin objects, such as coins (Fig. 1a), reduces the tolerance even further. In this study, the term *pick* is specifically used to denote the action of lifting thin objects off a table. Therefore, the word *grasp* is used to describe what is typically known as *pick*. Opening or closing a pencil case also involves complex and delicate dual-arm manipulation, which takes into account the position and orientation of the small zipper tip during grasping. *Needle-threading* and *banana-peeling* involve the manipulation of deformable objects, and allow for very small manipulation tolerances [5].

V. Network Architecture

The architecture of the policy predictor represents the evolution of the model proposed by [7] (Fig. 4 (a)). This neural network structure first processes foveated vision using EfficientNetV2, which is a parameter-efficient visual data processing model[19], then encodes the robot arm states and the predicted gaze position using a Transformer encoder. Subsequently, the Transformer decoder uses these encoded embeddings to output global-action and local-action. Global-action refers to a series of states up to a maximum of 20 time steps ahead, whereas local-action is the difference between the next state and the current state. Angles are represented in six dimensions for a continuous representation, as in [20], rather than as Euler angles. Both local and global-actions are trained with an ℓ_2 loss. Language instructions are first encoded using Deep Contrastive Learning for Unsupervised Textual Representations ([21]), and then used as conditioned inputs for the Transformer encoder and FiLM [22]-conditioned EfficientNet. FiLM was used, as in [2], to condition language-instructions to vision. However, unlike the study in [2] that used global vision, in the present study, foveated vision is used; thus, FiLM-based conditioning cannot choose the appropriate arm to use. For instance, for a language input such as "grasp nut with right

[5]Banana-peeling and some data from the needle-threading tasks were taken from previous studies [6], [7].

Task group	Objective	# episodes	Time (h)	# instructions
Grasp-and-place	Grasp and place various objects using one hand.	66,081	41.8	302
Pick	Pick up thin objects, such as coins and cards, using one hand.	10,936	9.09	14
Bottle	Set up or knock down a bottle (inspired by [2]).	8,114	4.70	57
Move-bowl	Move bowls or plates with both hands to a target plate.	5,775	3.33	48
Pencil-case	Open or close a pencil case using both hands.	23,650	14.6	24
Handkerchief	Fold or unfold a handkerchief using both hands.	24,671	13.3	155
T-shirt	Fold or unfold a t-shirt using both hands.	23,898	15.1	356
Lego	Grasp Lego.	12,380	9.60	75
Miscellaneous	Other tasks (e.g., rotating or stacking objects).	9,303	6.88	52
Needle-threading [6]	Pick up a thread and thread it through the needle.	9,975	8.75	3
Banana-peeling [7]	Peel a banana.	29,427	23.1	18
Sum		224,210	150	1,104

TABLE II: Task objectives and statistics.

white gripper,'' there may not be any information about the hand in the foveated image, which prevents transferring information about the hand for use in downstream neural networks. Therefore, to condition hand information, language information was added as an additional embedding in the Transformer encoder. Additionally, the encoder predicts each task's goal position (i.e., the final state of the episode), which contributes to policy stabilization by producing manipulation policies based on goals [7], which is trained using the mixture density network (MDN) loss [23] that fits the goal to the probability distribution represented by the Gaussian mixture model.

The main changes compared with the model in [7] can be summarized as follows: First, the model has been adapted to use language instructions so that, rather than having a specialized policy model for each task, a single policy model learns all tasks. Second, both global-action and local-action are inferred from the same neural network, and local-action also receives the robot's state and gaze coordinates. This is possible because of the use of the Transformer, which allows for attention to be directed to necessary somatosensory information, as shown in [10]. Finally, goal position prediction was previously performed using a unimodal model with an ℓ_2 loss, whereas DAA adopts a multimodal model using the MDN. This allows for the prediction of the probability distribution of goal positions, which enables the accurate prediction of target states, even in the presence of multiple objects.

The gaze predictor uses the cross-attention of the Transformer (Fig. 4 (b)). Previous gaze prediction models (e.g., [9]) predicted the gaze coordinates as a probability distribution using the MDN. However, training gaze prediction with the MDN was challenging under conditions involving multiple object scenes, which is a frequent occurrence in the training dataset. Therefore, in this study, left global vision is divided into an 18×32 grid, and the left gaze coordinate is transformed into a one-hot vector within this 18×32 grid of the left global vision, thereby solving it as a classification problem. Left global vision 144×256 serves as the input to EfficientNetV2, and the processed 18×32 visual features undergo cross-attention with language, as proposed in [24], to condition language on vision. The output of cross-attention was trained using cross-entropy loss to predict the left gaze coordinate. Predicting both gazes independently (e.g., [7]) often led to predictions of different objects in multi-object environments. Therefore, in this study, the right gaze is calculated inversely using trigonometry. This is based on the predicted left gaze coordinates and depth information computed from the stereo camera.

VI. Experiment

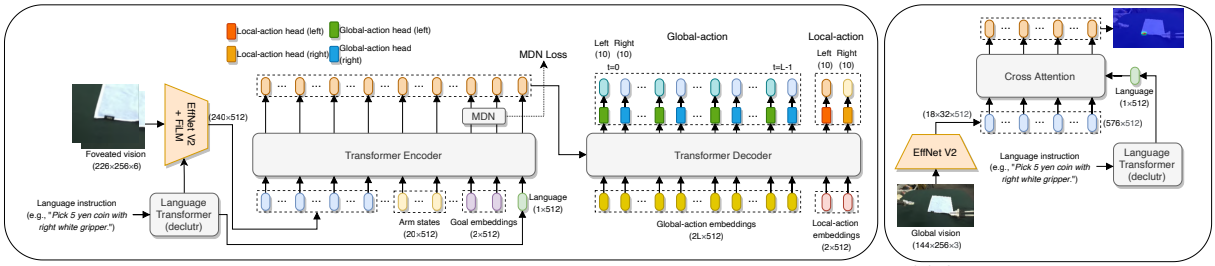
We validated the generalization and robustness of multi-task trained DAA in a total of 7,815 trials for a real robot.

A. Multi-Task Performance of DAA

We evaluated the generalization ability across various tasks and novel objects. To achieve this, 103 tasks with objects included in the training set and 39 tasks involving objects not included in the training set were attempted 9 to 36 times each. The test task setups are explained in Table III. *Pick* tests, which require picking up very thin objects, such as coins or plastic cards on a table, are differentiated from *grasp* tests because the former tests demand distinct manipulation skills. For the sake of experimental fairness, the background was uniformly set to green. The multi-task trained model was trained on the task groups *grasp-and-place*, *pick*, *bottle*, *move-bowl*, *pencil-case*, *handkerchief*, *t-shirt*, and *miscellaneous* from the dataset[6]. Task-specific trained models were trained separately for each object in each task group. In the testing of the proposed DAA, results for *t-shirt* and *miscellaneous* were excluded. The *t-shirt* task failed because the t-shirt, which was large, occupied the edges of the camera vision, which led to significant errors in the eye tracker. Thus, the gaze predictor failed to predict the correct gaze coordinate, which resulted in the failure of the folding t-shirt task.

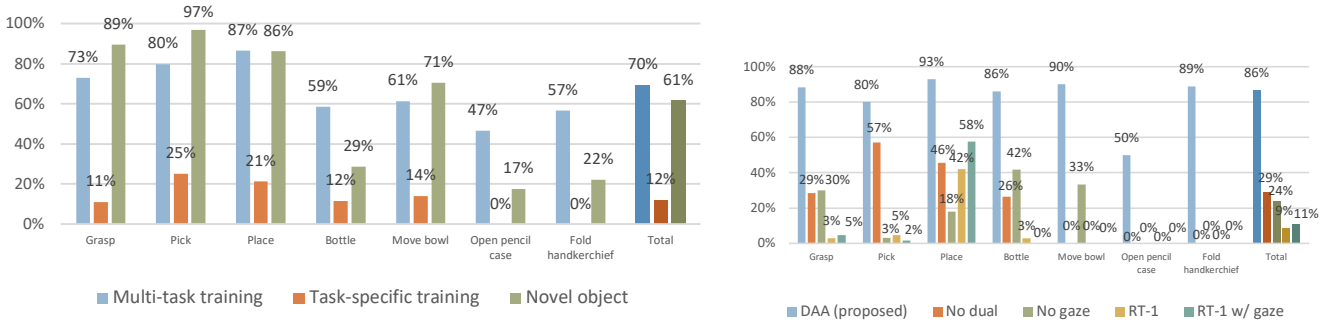
The results for the multi-task trained model and task-specific trained model are shown in Fig. 5a. The proposed multi-task model achieved a 69.6% success rate across all tasks, whereas the task-specific model achieved a 12.2% success rate, which indicates that multi-task learning significantly improved the success rate. Additionally, the multi-task model recorded a 61.5% success rate on new objects, which suggests that the multi-task model is capable of generalizing its manipulation skills to new objects.

[6]*Lego* was excluded because of its complexity, and *needle-threading* and *banana-peeling* were experimented on additionally in VI-E.



(a) Policy predictor (b) Gaze predictor

Fig. 4: Neural network architectures. (a) The policy predictor outputs global-action and local-action through a Transformer encoder-decoder structure. It processes inputs such as high-resolution foveated vision attended by the gaze predictor, robot arm states, gaze coordinates, and language instructions. (b) The gaze predictor generates output by applying cross-attention to visual embeddings of global vision and language embeddings. The final gaze position is sampled based on the probability distribution of this output.

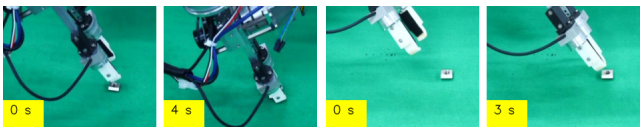


(a) Manipulation results of multi-task and task-specific trained models. (b) Results of the ablation studies.

Fig. 5: Experimental results of DAA.

Test	Objective	# of Trials
Grasp	Grasp a target object.	16-32
Pick	Lift a thin target object off a table.	21
Place	Place a bolt in a bowl or plate.	25
Bottle	Knock down a bottle or set it upright.	9
Move-bowl	Support both sides of a bowl and place it on a plate.	10
Open-pencil-case	Grasp the tip and zipper of a pencil case and open it.	30
Fold-handkerchief	Grasp the left and right parts of a handkerchief and fold it.	9

TABLE III: Test setups.



(a) DAA (proposed) succeeded in grasping the target object. (b) RT-1 (with gaze) failed to grasp the small target object.

Fig. 6: Comparison of DAA and RT-1. RT-1 (with gaze) moved the end-effector close to the target object, but it failed to achieve successful manipulation.

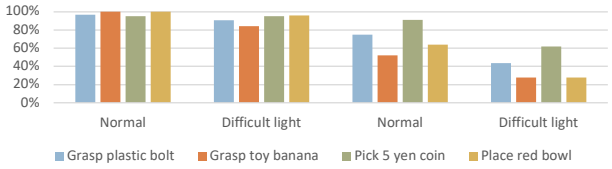
B. Ablation Studies

DAA includes visual attention and dual-action. To verify the importance of these elements in multi-task learning, in Fig. 5b, DAA is compared across 26 tasks with a *No gaze* model, which lacks the visual attention mechanism, but uses global vision instead of foveated vision, and a *No dual* model, which lacks dual-action, and thus does not distinguish dual-action and considers all actions as reactive. All of these models have the same number of policy predictor parameters. Additionally, comparisons were made with RT-1 [2], and a modified version of RT-1 using foveated vision and somatosensory information (RT-1 with gaze, Appendix A). As a result, both the *No dual* and *No gaze* models achieved lower success rates, which indicates the necessity of visual attention and dual-action. Furthermore, the RT-1s, as shown in Fig. 6, were not suitable for datasets that include tasks that require fine manipulation skills, although they achieved some success in *place* tasks that did not require precision. This implies that the manipulation skill properties of our dataset are significantly different from those of RT-1, which is focused on generalizing to new objects or environments using a large dataset that mainly comprises pick-and-place tasks.

C. Robustness Against Light Condition Variations

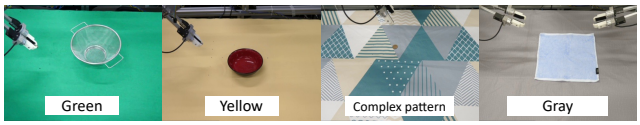
Changes in lighting conditions, such as luminance and shadow direction, alter various visual elements, and the DAA is required to adapt to these changes robustly. To verify the

(a) Different light conditions.

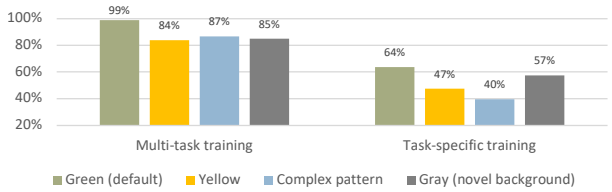


(b) Comparison of multi-task and task-specific trained models for various light conditions.

Fig. 7: Test results for various light conditions.



(a) Various backgrounds.



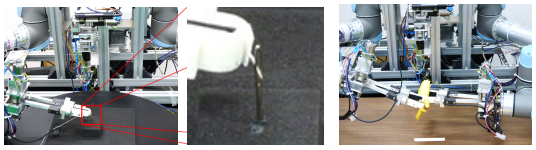
(b) Comparison of multi-task and task-specific trained models on various backgrounds.

Fig. 8: Test results for background changes.

robustness of the DAA, the degree of lighting change was controlled as *Easy*, *Medium*, and *Difficult* (Fig. 7a), and the robot was tested on three grasping tasks. In Fig. 7b, the results of the proposed multi-task trained model and task-specific models are compared for four tasks. The results showed that the multi-task model recorded only a slight decrease in the success rate, even under *Difficult* conditions, whereas the task-specific trained models experienced a significant drop in the success rate.

D. Robustness Against Background Variations

The robustness of the multi-task trained DAA to changes in the robot's background was tested on four types of tablecloth backgrounds (Fig. 8a): *Green* (default background), *Yellow* (another color included in the training dataset), *Complex pattern* (a complex pattern included in the training data), and *Gray* (a color not included in the training data). In Fig. 8b, the proposed multi-task trained model is compared with task-specific trained models on three tasks. These tasks were selected because the task-specific models achieved relatively high performance in them. As a result, it was observed that the success rate of the task-specific trained models, particularly in the case of complex patterned backgrounds, was significantly lower than in the case of the default background.



(a) Needle-threading. (b) Banana-peeling.

Fig. 9: DAA executes challenging manipulation tasks.

By contrast, the multi-task trained model maintained a more consistent success rate across various backgrounds.

E. Additional Results

DAA was trained on challenging manipulation tasks (Fig. 9) that were previously presented in [6] and [7]. First, for needle-threading, DAA achieved manipulation accuracy comparable with the results reported in [6] (Table IV)[7]. It is noteworthy that, compared with a model trained from scratch, there was no performance improvement in the model that started training from the multi-task trained model in Section VI. This result suggests that multi-task training may not be effective for learning completely novel skills.

Previous work [6]	DAA (scratch)	DAA (pretrained)
81.25%	85.71%	85.19%

TABLE IV: Needle-threading results.

The result of DAA for the banana-peeling task is presented in Table V, which demonstrates performance comparable with previous the study [7]. Because there was no performance improvement with DAA (pretrained) in the needle-threading task, only DAA (scratch) was tested.

Previous work [7]	DAA (scratch)
0.870	0.900

TABLE V: Banana-peeling results.

VII. Conclusions and Discussion

We addressed the challenge of imitation learning for dual-arm fine manipulation using a multi-task, at-scale dataset and DAA. We created and released a robot demonstration dataset containing 224k episodes for imitation learning, and implemented a multi-task learning agent by applying the dataset to DAA. This dataset is novel in that it includes visual attention signals and dual-action labels, and it includes fine manipulation skills with deformable objects and/or with dual arms. We successfully extended DAA for learning dual-arm fine manipulation to multi-task learning using this dataset and observed an improvement in manipulation performance through multi-task learning. Furthermore, the effectiveness of gaze-based visual attention was confirmed for fine object manipulation in multi-task learning. This architecture selectively

[7]These results used different amounts of data (8.75 hours in the current dataset vs. 3.60 hours in [6]). Because the environment including the lighting condition changed from [6], these results cannot be directly compared.

extracts high-resolution pixels related to the target object in a scene. Additionally, the dual-action approach, which distinguishes between ballistic global-action for approaching the vicinity of the object and reactive local-action for precise manipulation, was also important.

The limitations of this study include, first, the use of only one type of robot framework. Considering that in studies such as [5], researchers produced learning data for various types of robots, our framework's adaptability to novel robot frameworks can be regarded as relatively lower. However, we believe that the usability of our robot data in the robot research community can be maximized for the following reasons. First, by making the design of the robot public, we have enabled reproducibility. Second, we expect that the gaze-based visual attention mechanism, which removes the robot arm's image from policy inference and only includes the end-effector's visual appearance, will enhance usability because the end-effector is relatively easy to replicate. Another limitation is that, in this study, we focused on the generalization of fine dual-arm manipulation skills and did not address the acquisition of semantic reasoning. This would require a combination of DAA dataset with large vision-language models, as discussed in [3]; however, exploring the emergence of such features for fine manipulation tasks is challenging and may require large language-action pairs and computational power.

APPENDIX

A. Model Details

The DAA's architecture comprises 12 layers each for the encoder and decoder, which have an embedding size of 768 and feedforward network dimension of 3072 within the Transformer layers. The total number of parameters is 350M.

The training of RT-1 was based on `https://github.com/lucidrains/robotic-transformer-pytorch`. Additionally, the model was modified to use images at the current time step rather than accumulating six images as proposed in [2] because accumulation did not work in our environment and to reduce the training time, which was roughly half the speed of DAA, even without accumulation. The number of parameters of the RT-1s was 198M. Global vision to 224×224 and an output dimension of $10 + 10$ were used to represent dual-arm actions. RT-1 with gaze inputs dual-arm states and gaze coordinates into Conv as additional channel data, in addition to the resized foveated vision (224×224). This structure was designed to avoid making too many changes to the original RT-1 structure.

REFERENCES

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of the Robotics: Science and Systems (RSS) 2023*, 2023.

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *arXiv preprint arXiv:2307.15818*, 2023.

[4] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[5] O. X.-E. Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," https://arxiv.org/abs/2310.08864, 2023.

[6] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Gaze-based dual resolution deep imitation learning for high-precision dexterous robot manipulation," *Robotics and Automation Letters*, vol. 6, no. 2, pp. 1630–1637, 2021.

[7] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Robot peels banana with goal-conditioned dual-action deep imitation learning," *arXiv preprint arXiv:2203.09749*, 2022.

[8] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[9] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Using human gaze to improve robustness against irrelevant objects in robot manipulation tasks," *Robotics and Automation Letters*, vol. 5, no. 3, pp. 4415–4422, 2020.

[10] ——, "Transformer-based deep imitation learning for dual-arm robot manipulation," in *International Conference on Intelligent Robots and Systems*, 2021.

[11] J. Paillard, "Fast and slow feedback loops for the visual correction of spatial errors in a pointing task: a reappraisal," *Canadian Journal of Physiology and Pharmacology*, vol. 74, no. 4, pp. 401–417, 1996.

[12] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "Rh20t: A robotic dataset for learning diverse skills in one-shot," in *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[13] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[15] H. Kolb, "Simple anatomy of the retina," *Webvision: The Organization of the Retina and Visual System [Internet]*., pp. 13–36, 1995. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK11533/

[16] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, pp. 188–94, 2005.

[17] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Memory-based gaze prediction in deep imitation learning for robot manipulation," in *2022 International Conference on Robotics and Automation*, 2022, pp. 2427–2433.

[18] H. Kim, Y. Ohmura, A. Nagakubo, and Y. Kuniyoshi, "Training robots without robots: Deep imitation learning for master-to-robot policy transfer," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2906–2913, 2023.

[19] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[20] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.

[21] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "Declutr: Deep contrastive learning for unsupervised textual representations," *arXiv preprint arXiv:2006.03659*, 2020.

[22] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] C. M. Bishop, "Mixture density networks," Neural Computing Research Group, Aston University, Tech. Rep., 1994.

[24] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, *et al.*, "Interactive language: Talking to robots in real time," *IEEE Robotics and Automation Letters*, 2023.