IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 August 2023, accepted 28 October 2023, date of publication 8 November 2023, date of current version 15 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331373

RESEARCH ARTICLE

Policy Optimization for Waste Crane Automation From Human Preferences

YUHWAN KWON[1], (Member, IEEE), HIKARU SASAKI[1], (Member, IEEE), TERUSHI HIRABAYASHI[2], KAORU KAWABATA[2], AND TAKAMITSU MATSUBARA[1], (Member, IEEE)

[1]Robot Learning Laboratory, Division of Information Sciences, Nara Institute of Science and Technology, Nara 630-0192, Japan
[2]Hitachi Zosen Corporation, Osaka 559-8559, Japan

Corresponding author: Yuhwan Kwon (y-kwon@is.naist.jp)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Nara Institute of Science and Technology.

ABSTRACT This research introduces a novel approach to optimizing control policies for waste cranes operating at waste-to-energy plants. Although previous methods forced people to define evaluation functions for automation, such design works in actual environments can often be challenging due to limited sensors and design difficulties. This paper aims to establish a methodology that achieves automation by having people respond to interactive pairwise comparison queries, which is relatively simple compared to design work. On the other hand, considering such automation, it becomes imperative to address the increased sample cost associated with slow crane operation and the complexities of decision-making due to waste inhomogeneity. Our proposed Preferential Bayesian Policy Optimization (PBPO) optimizes control policies with a small number of queries using Preference-based Bayesian optimization (PbBO) and mitigates the difficulty of decision-making by providing human evaluators who have an option to skip queries. We also incorporate a query synthesis mechanism to enhance query efficiency that generates a new preference relation from the skipped queries. PBPO's effectiveness was validated with a scattering task employed in previous studies. Experimental results with simulated evaluators show the effectiveness of the PBPO and query synthesis. Furthermore, results with actual human evaluators indicate that our proposed method performs as well as the Bayesian optimization (BO) method, which requires an evaluation function.

INDEX TERMS Automation, Bayesian methods, cranes, humans in the loop, interactive systems, optimization methods, waste handling.

I. INTRODUCTION

Waste-to-energy plants typically have pits that temporarily store incoming waste, and large cranes maneuver and agitate it to facilitate stable combustion. Even though a portion of such crane operations is automated, adjusting and fine-tuning the controllers by crane operators are often essential. Such adjustments are caused by the diversity of the incoming waste: its shape, size, weight, hardness, and flammability. Therefore, the operators, who are familiar with the behavior of the cranes and the plant conditions, design and adjust their controllers based on their experience to ensure optimal performance. However, designing and adjusting controllers by operators is problematic because their training is costly and time-consuming.

A promising approach to reduce reliance on operators is a trial-and-error method of adjusting the controller to minimize a predefined evaluation function. Mackin et al. defined such a function based on the number of agitations and proposed a method to optimize the crane operation schedule, defined by a sequential list of commands, using a genetic algorithm [1]. Although they trained their scheduler on a simulator, conducting trial-and-error in a real environment is essential to account for waste diversities. Sasaki et al. proposed an

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiwu Li.

automation method based on Bayesian optimization [2] and policy optimization [3] to consider the high data collection costs due to a crane's slow motion. They also defined an evaluation function for the weight of the waste grasped by the crane.

Although previous studies have achieved some success, preparing evaluation functions can be difficult for several reasons. First, manually designing them in the real world is exhausting due to such issues as negative side effects and hacking [4], [5]. Although step-by-step modification of an evaluation function may solve these problems, this is a heavily operator-dependent process. Second, as a unique challenge in waste crane automation, limited sensors complicate designing an evaluation function due to waste's diversity. In typical waste-to-energy plants, we can only access limited sensor information: the values of load cells installed on the cranes, the pseudo-heights, and the number of agitations in a delimited area in the pit. Indeed, the evaluation functions in previous studies [1], [2], [3] are well-designed, leaving the more complex tasks that consider waste's diversity as a challenge.

Our idea for eliminating the need to design an evaluation function requires humans to choose between the desirability of two different crane behaviors. We optimize the control policy by interactively presenting the behavior generated by two different policies to a human evaluator and ask her to choose the option that better achieves the present task objective. Since it is easier to judge whether a task is good or bad based on shared criteria and objectives than to design and adjust control policies and evaluation functions, we argue that this query-based approach relaxes the requirement that evaluators possess operational skills and domain knowledge. As a result, this approach allows people without such skills or knowledge to become evaluators. However, in adopting this approach, we must address the following technical difficulties: 1) how to learn with fewer queries to reduce the high trial-and-error cost of actual heavy machinery and 2) how to deal with situations where making clear decisions is complicated due to the characteristics of waste.

In this paper, we propose Preferential Bayesian Policy Optimization (PBPO) as a solution to the problems mentioned earlier. As illustrated in Fig. 1, PBPO repeatedly presents two crane operations to human evaluators who select the one that better matches the task purpose, thus obtaining an optimal control policy. To address problems 1) and 2), PBPO uses Preference-based Bayesian optimization (PbBO) [6], [7] for the query selection and provides a skip option to the evaluators for difficult-to-judge queries. Furthermore, to reduce the number of necessary queries, PBPO introduces a query synthesis mechanism that generates new preference relations from queries skipped by evaluators and their past choices.

The following are the key contributions of this paper:

1) To eliminate the challenging work of designing evaluation functions for automating heavy machinery, we proposed a method that optimizes the control policies through interactive pairwise comparison queries to human evaluators.
2) We validated its effectiveness using several implemented *simulated evaluators* with varying response uncertainties.
3) With the help of human evaluators, we validated our query synthesis mechanism and compared its acquisition performance with previous methods that explicitly require an evaluation function.

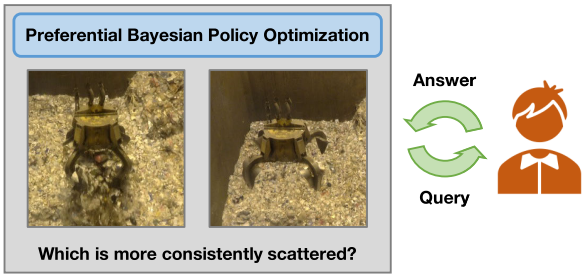


FIGURE 1. Applying proposed Preferential Bayesian Policy Optimization (PBPO) to a waste crane operation: PBPO optimizes control policy by repeatedly asking human evaluators which of two videos with different control strategies is working better.

The remainder of this paper is organized as follows. Section II summarizes related studies. Section III provides the PbBO details used in the proposed method, and. Section IV outlines our proposed method, PBPO. Then we describe the tasks used in the experiments in Section V. Sections VI and VII present the experiments with *simulated evaluators* and actual human evaluators. Sections VIII and IX present the discussion and conclusion.

II. RELATED WORK

A. HEAVY MACHINERY AUTOMATION BY DATA-DRIVEN APPROACH

The automation of construction equipment, especially excavators, is an area of intense research [8], [9], [10], [11], [12]. As a data-driven automation approach, Egli et al. proposed a method for automating the arm of a hydraulic excavator using a data-driven actuator model and a control policy trained by reinforcement learning on a simulator and tested it on a grading task with actual equipment [13]. Tahara et al. focused on the variations in the quality of demonstrations and the lack of diversity, both of which are problems in automating excavation through imitation learning, and proposed an efficient learning method that explicitly uses task accomplishment [14].

There has also been extensive research on automation for various types of cranes [15], including tower [16], gantry [17], [18], and overhead cranes [19]. Chun et al. introduced a method that integrates deep reinforcement learning with an algorithm for identifying static initial equilibrium states to automate the lifting of large blocks by cranes, a typical operator-dependent task [20]. Cho et al. described their strategy for automating tower-crane-lifting operations and

estimating the lifting times at construction sites. They utilized agents trained by reinforcement learning on a dynamic simulator [21].

Only a few studies, however, have focused on the automation of cranes used in waste-to-energy plants. These studies aim to effectively manage the waste in pits to ensure stable combustion. Mackin et al. utilized genetic algorithms (GAs) and tackled the challenge of automating the scheduling of action sequences, such as where in the pit to collect the waste and where to transport it [1]. GA's fitness was defined in terms of the degree of trash agitation and the flatness of the pit planes. Concentrating on the task of scattering the collected waste to agitate it, Sasaki et al. proposed a Bayesian optimization (BO)-based method [2] that accounts for robustness to waste inhomogeneity and query efficiency and a self-triggered policy search method [3], where predefined control strategies and their durations were employed as control policies. Both studies assessed the policy performance using a weight-based evaluation function.

Our proposed method diverges from prior work by optimizing the control policies from human preferences without predefined evaluation functions.

B. PREFERENCE-BASED POLICY OPTIMIZATION

Much research has been devoted to preference-based deep reinforcement learning in recent years. Christiano et al. proposed a deep reinforcement learning framework that learns a reward function from human responses to pairwise comparison queries [22]. Since their framework requires hundreds to thousands of human feedback, subsequent studies [23], [24], [25], [26] have addressed improving feedback efficiency, which is essential for learning reward functions from human preferences. On the other hand, those studies are based on model-free reinforcement learning, which is unsuitable for automating targets that require a lot of trial-and-error time, such as heavy equipment.

Several studies have attempted to optimize policies from dozens of queries based on approaches that do not rely on deep learning. Sadigh et al. proposed a method for learning the weights of a reward function expressed as a weighted sum of predefined features in an active learning framework [27]. Biyik et al. proposed a learning method that captures nontrivial nonlinearities in reward functions by modeling them as a Gaussian process [28]. Basu et al. developed a model that captures human reward dynamics as they change in response to environmental interactions and proposed a learning method using hierarchical queries [29].

Compared to policy optimization based on learned reward functions, a framework that directly optimizes policies has the potential to reduce the number of queries. Tucker et al. proposed a Bayesian dueling bandit approach to optimize the gait parameters of the lower body exoskeleton, which incorporates cooperative feedback to allow the selection of a preferred action between two presented alternatives [30]. Tucker et al. also extended their algorithm to capture preferences on higher dimensional parameter spaces by iteratively exploring random one-dimensional subspaces [31]. Additionally, some studies used GLISp [32], a pairwise preference-based optimization algorithm, to tune the parameters of model predictive control [33] or a path-based velocity planner with fuzzy logic [34].

Our research is unique because it proposes policy optimization using preference-based Bayesian optimization to improve the query efficiency for automating heavy machinery.

III. PRELIMINARIES

A. BAYESIAN OPTIMIZATION

We consider a situation where we aim to find optimal parameter $\mathbf{w}^* = \operatorname{argmax}_{\mathbf{w}} f(\mathbf{w})$ for objective function $f(\mathbf{w})$ of parameter $\mathbf{w} \in \mathcal{W}$, even though its optimization is analytically difficult. Bayesian optimization (BO) [35] is a sequential design strategy for such optimization.

First, BO approximates the objective function with a surrogate model that is relatively easy to evaluate. We use a Gaussian process (GP) [36] as a surrogate model and regress the relationship between parameter w_n and corresponding objective function value e_n:

$$e_n = f_n + \varepsilon_n, \tag{1}$$

where $\varepsilon_n \sim \mathcal{N}(0, \beta)$ is the Gaussian noise and $f_n = f(\mathbf{w}_n)$ with simplified notation. Let $\mathbf{W} := [\mathbf{w}_1, \cdots, \mathbf{w}_N]^\intercal$ be the parameters already evaluated by the objective function, and let $E := [e_1, \cdots, e_N]^\intercal$ be the corresponding evaluation values. GP regresses the objective function on the predictive distribution as follows:

$$\begin{aligned} &p(f(\mathbf{w}) \mid \mathbf{w}, E, \mathbf{W}) \\ &\quad = \mathcal{N}(f(\mathbf{w}) \mid \mu(\mathbf{w}), \sigma^2(\mathbf{w})), &(2)\\ &\mu(\mathbf{w}) = \mathbf{k}_{\mathbf{W},*}^\intercal (\mathbf{K}_\mathbf{W} + \beta \mathbf{I})^{-1} E, &(3)\\ &\sigma^2(\mathbf{w}) = \mathrm{k}(\mathbf{w}, \mathbf{w}) - \mathbf{k}_{\mathbf{W},*}^\intercal (\mathbf{K}_\mathbf{W} + \beta \mathbf{I})^{-1} \mathbf{k}_{\mathbf{W},*}, &(4) \end{aligned}$$

where $\mathbf{K}_\mathbf{W}$ is a gram matrix with $[\mathbf{K}_\mathbf{W}]_{ij} = \mathrm{k}(\mathbf{w}_i, \mathbf{w}_j)$, $\mathrm{k}(\cdot, \cdot)$ is a kernel function with kernel parameter $\boldsymbol{\theta}_k$, $\mathbf{k}_{\mathbf{W},*}$ is a kernel vector with $[\mathbf{k}_{\mathbf{W},*}]_i = \mathrm{k}(\mathbf{w}_i, \mathbf{w})$, and $\mathbf{I}$ is a unit matrix. Moreover, mean function $\mu(\mathbf{w})$ and variance function $\sigma^2(\mathbf{w})$ represent the mean and variance of the predictive distribution, and the value of $\sigma^2(\mathbf{w})$ tends to increase in regions with insufficient data.

Then BO generates queries $\mathbf{w}'$ using predictive distribution. We introduce an acquisition function $\alpha(\cdot)$ that considers the trade-off between exploration and exploitation and generates a query:

$$\mathbf{w}' \leftarrow \operatorname*{argmax}_{\mathbf{w}} \alpha(\mathbf{w}). \tag{5}$$

This query $\mathbf{w}'$ is then evaluated by the objective function, and BO updates the surrogate model using $\mathbf{w}'$ and evaluation value $f(\mathbf{w}')$. These steps are repeated until the query parameter converges to $\mathbf{w}^*$.

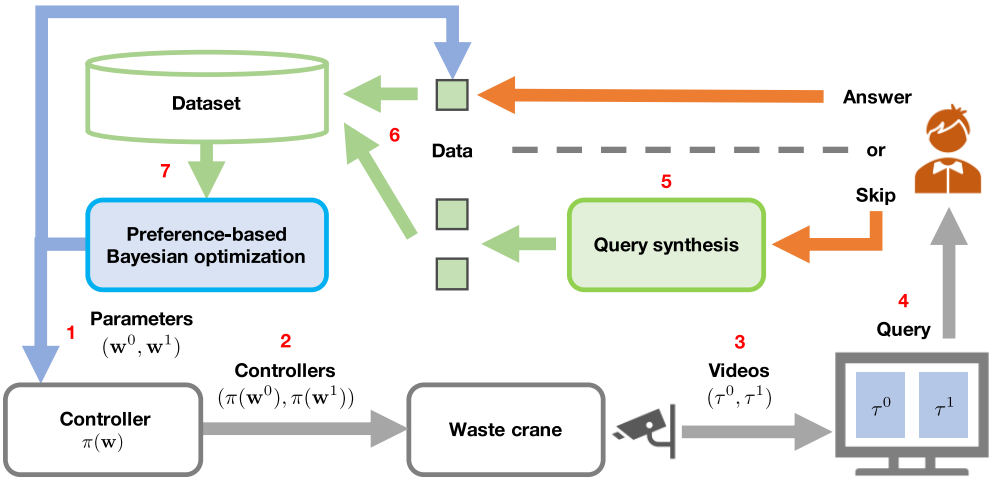


FIGURE 2. Overview of parameter optimization of a waste crane controller by proposed Preferential Bayesian Policy Optimization (PBPO).

B. PREFERENCE-BASED BAYESIAN OPTIMIZATION

Unlike BO, we now consider a case where no evaluation value e is given directly, but instead a preference relation is given for a query that bundles two parameters $(\mathbf{w}^0, \mathbf{w}^1)$. In this paper, we refer to BO for such problem settings as a Preference-based BO (PbBO) [6], [7]. Assume that the response to query y is given as $y = 0$ when $\mathbf{w}^0$ is preferred over $\mathbf{w}^1$, and $y = 1$ when $\mathbf{w}^1$ is preferred over $\mathbf{w}^0$. We now consider that y is given based on latent evaluation function $f(\cdot)$:

$$y = \begin{cases} 0, & \text{if } f(\mathbf{w}^0) \geq f(\mathbf{w}^1) \\ 1, & \text{if } f(\mathbf{w}^0) < f(\mathbf{w}^1) \end{cases} \tag{6}$$

To simplify the notation, $f(\mathbf{w}^0)$ and $f(\mathbf{w}^1)$ are referred to as f^0 and f^1.

PbBO treats the preference relation as a probability distribution, and the distribution of the latent evaluation function is approximated by variational inference using the responses to the queries. First, let $\boldsymbol{\theta}_k \in \Theta$ be a kernel parameter, and let $\mathbf{W} := \{\mathbf{w}_i^0, \mathbf{w}_i^1\}_{i=1}^N$ be previous queries; we denote prior distribution $p(\mathbf{f}|\boldsymbol{\theta}_k)$ by GP with mean 0 and covariance matrix $\mathbf{K}$, where $\mathbf{f}$ is a simplified notation of $f(\mathbf{W})$. Furthermore, assuming that the preference relation is polluted by Gaussian noise, we define the likelihood of the preference relation [37]:

$$p(y|f^0, f^1) = \int_{-\infty}^{\frac{f^1 - f^0}{\sqrt{2}\epsilon}} \mathcal{N}(\gamma|0, 1) d\gamma, \tag{7}$$

where ϵ is a hyperparameter. The distribution of answers $\mathbf{Y} := \{y_i\}_{i=1}^N$ to previous queries $\mathbf{W}$ is described:

$$p(\mathbf{Y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i^0, f_i^1). \tag{8}$$

Therefore, from Bayes' theorem, the posterior distribution is described:

$$p(\mathbf{f}|\mathbf{Y}, \boldsymbol{\theta}_k) = \frac{p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\boldsymbol{\theta}_k)}{\int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\boldsymbol{\theta}_k) d\mathbf{f}}. \tag{9}$$

Now, since the likelihood given by (7) is only obtained numerically, we cannot analytically obtain the posterior distribution. Therefore, we approximate $p(\mathbf{f}|\mathbf{Y}, \boldsymbol{\theta}_k)$ by the Variational Bayesian method [38]. Let $q(\mathbf{f})$ be the variational distribution, and we approximate $p(\mathbf{f}|\mathbf{Y}, \boldsymbol{\theta}_k)$ by maximizing log marginal likelihood $\log p(\mathbf{Y}|\boldsymbol{\theta}_k)$. Using Jensen's inequality, maximizing the log marginal likelihood is replaced by maximizing the Evidence Lower Bound (ELBO):

$$\begin{aligned} &\log p(\mathbf{Y}|\boldsymbol{\theta}_k) \\ &= \log \int \frac{q(\mathbf{f}) p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\boldsymbol{\theta}_k)}{q(\mathbf{f})} d\mathbf{f} \\ &\geq \int q(\mathbf{f}) \log p(\mathbf{Y}|\mathbf{f}) d\mathbf{f} - \int q(\mathbf{f}) \log \frac{q(\mathbf{f})}{p(\mathbf{f}|\boldsymbol{\theta}_k)} d\mathbf{f} \\ &= \underbrace{\mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{Y}|\mathbf{f})\right] - \mathrm{KL}(q(\mathbf{f})||p(\mathbf{f}|\boldsymbol{\theta}_k))}_{\text{ELBO}}. \end{aligned} \tag{10}$$

We maximize ELBO instead of the log marginal likelihood. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a multivariate Gaussian distribution where $q(\mathbf{f}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and then ELBO is rearranged:

$$\begin{aligned} \text{ELBO} = &\sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{f})}\left[\log p(y_i|f_i^0, f_i^1)\right] - \frac{1}{2}\mathrm{tr}\{\mathbf{K}^{-1}\boldsymbol{\Sigma}\} \\ &- \frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\mathbf{K}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}. \end{aligned} \tag{11}$$

We alternately optimize variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and kernel parameter $\boldsymbol{\theta}_k$ using automatic differentiation.

Finally, we generate a new two-choice query $(\mathbf{w}_0', \mathbf{w}_1')$. Now, the inference of f' for parameter candidate $\mathbf{w}'$ for the

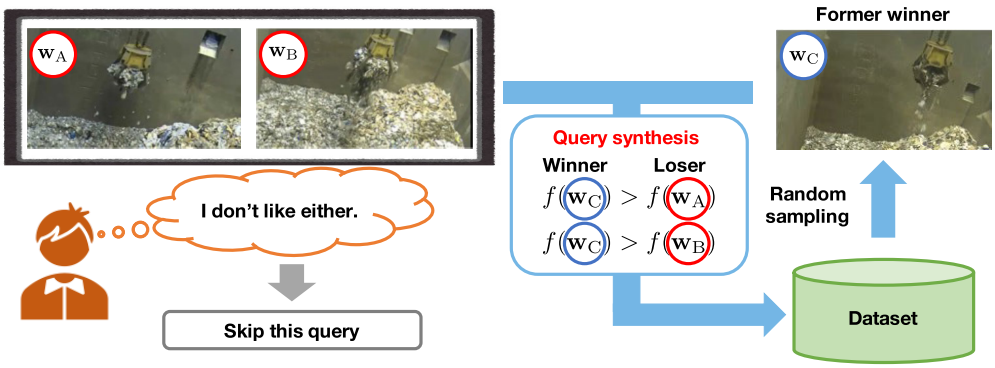


FIGURE 3. Overview of query synthesis: When evaluator skips answering a query, PBPO synthesizes new queries using parameters previously selected by evaluator and parameters that comprise skipped query.

new query is given:

$$p(f'|\mathbf{Y}, \mathbf{w}') = \int p(f'|\mathbf{Y}, \mathbf{f}, \mathbf{w}')q(\mathbf{f})d\mathbf{f}. \tag{12}$$

From (12) and the acquisition function, we can select parameter candidates $\mathbf{w}'_0, \mathbf{w}'_1$ for the new query. This study uses Thompson Sampling (TS) as its acquisition function. TS samples functions over $\mathcal{W}$ and selects the largest $\mathbf{w}'$ in the sampled functions as a candidate, and $\mathbf{w}'_0, \mathbf{w}'_1$ is acquired by repeating this operation twice.

IV. PREFERENTIAL BAYESIAN POLICY OPTIMIZATION

A. POLICY EVALUATION SYSTEM

We consider a controller with parameter $\mathbf{w} \in \mathcal{W}$ as a control policy for the waste cranes and optimize their operation by optimizing $\mathbf{w}$. Our proposed Preferential Bayesian Policy Optimization (PBPO) estimates latent evaluation function $f(\cdot)$ and obtains optimal parameter $\mathbf{w}^*$ by repeating a procedure that presents two crane behaviors to an evaluator and asks her to select the operation that better meets the objective. For example, let τ' be the crane operation by controller $\pi(\cdot)$ with $\mathbf{w}'$. We estimate $f(\cdot)$ from the answers to the two-choice query of τ, assuming that the preference relation between τ and $\mathbf{w}$ coincides. We record the operation as a video and present it to evaluators so that the two operations in the query can be compared simultaneously. Therefore, we consider τ to be a video.

Fig. 2 shows the PBPO's process flow for optimizing the controller of the waste cranes:

1) Generate a new two-choice query $(\mathbf{w}^0, \mathbf{w}^1)$ with PbBO.
2) Set each parameter $\mathbf{w}$ that comprises the query to controller $\pi(\cdot)$.
3) Execute the task with the waste crane controlled by $\pi(\mathbf{w})$. The operation is recorded by a camera as a pair of videos (τ^0, τ^1).
4) Present (τ^0, τ^1) corresponding to $(\mathbf{w}^0, \mathbf{w}^1)$ to the evaluator.
5) If the evaluator skips an answer, perform query synthesis (Section IV-B).
6) Add a query and answer pair to the dataset.

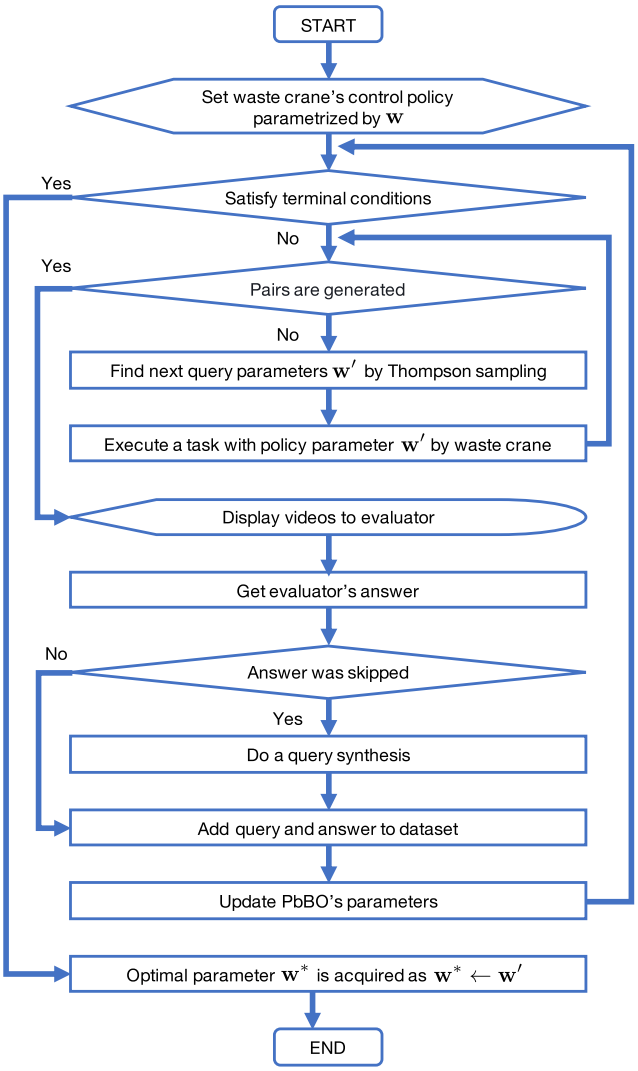


FIGURE 4. PBPO flowchart.

7) Update PbBO parameters with the dataset.

Here PBPO estimates $f(\cdot)$ and obtains w^* by repeating steps 1) through 7) several times. Note that 2) and 3) are

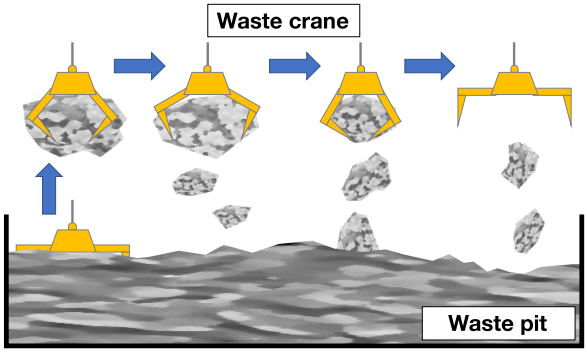


FIGURE 5. Illustration of waste-scattering task by a waste crane, where blue arrows indicate task flow: Yellow crane first grabs the waste in pit and opens and closes its claws to scatter it in small quantities while moving.

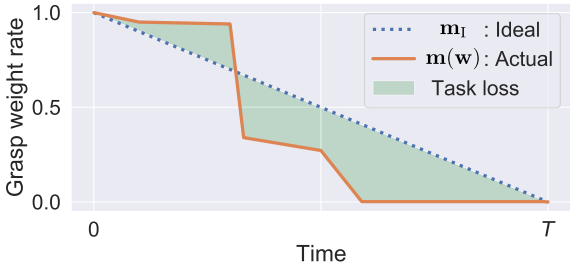


FIGURE 6. Overview of evaluation function for scattering task: We define green-shaded region between ideal grasp weight transition (dashed blue line) and actual grasp weight transition (solid orange line) as loss.

executed twice in each loop. In addition, we assume that an RGB camera that records τ is installed in appropriate locations to monitor the work. Indeed, our targeted plant, tested in a previous work [2], installed a camera in its control room overlooking the pit.

B. QUERY SKIPPING AND QUERY SYNTHESIS

Query synthesis is a method of improving query efficiency while reducing the burden on evaluators by allowing them to skip answering a difficult-to-judge query only if "both alternatives in a query are undesirable." Avoiding the direct use of potentially erroneous feedback in estimations can contribute to maintaining the algorithm's reliability. Fig. 3 shows an overview of query synthesis. Assume a situation where PBPO presents crane operations corresponding to $\mathbf{w}_A$ and $\mathbf{w}_B$; however, an evaluator judges that neither is desirable and skips answering. When an answer is skipped, PBPO randomly extracts parameters corresponding to the operation that the evaluator answered as desirable from the past two-choice queries in the dataset. Let $\mathbf{w}_C$ be a randomly extracted parameter; query synthesis generates pairs $(\mathbf{w}_C, \mathbf{w}_A)$ and $(\mathbf{w}_C, \mathbf{w}_B)$, both of which are labeled with $y = 0$, and adds them to the dataset. If the selected parameter does not yet exist, queries are accumulated until the evaluator chooses between two options.

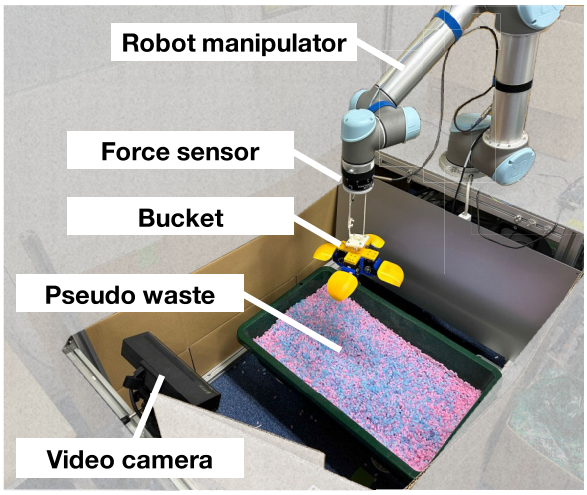


FIGURE 7. Constructed robotic waste crane system and pseudo waste.

Finally, we show a PBPO flowchart, including query synthesis, in Fig. 4.

V. WASTE-SCATTERING TASK

We conducted validation experiments on a scattering task for which an evaluation function was previously defined [2]. As shown in Fig. 5, the scattering task homogenizes the waste in a pit by first grasping a sufficient amount with the crane's bucket and dropping it at a certain rate. Note that our PBPO does not require a predefined evaluation function; we have chosen tasks for which we can define an evaluation function to quantitatively evaluate the performance. The following subsection describes the predefined evaluation function, the controller used in the task, the constructed experimental environment, and the task settings.

A. PREDEFINED EVALUATION FUNCTION

We define the ideal control policy for waste-scattering tasks as continually dropping a precise amount of waste while moving, such that all the grasped waste is completely gone by the end of the movement. As in a previous study [2], the task's achievement is evaluated from the time series of the crane's grasped waste weight. Fig. 6 shows the normalized grasp weight transition corresponding to the crane's moving time. We define ideal weight transition $\mathbf{m}_\mathrm{I}$ that results from the ideal policy as one that decreases at a constant pace over time T, plotted by the blue dashed line in Fig. 6. Additionally, we plotted actual weight transition $\mathbf{m}(\mathbf{w})$ when the task is performed under $\pi(\mathbf{w})$ as a solid orange line and define the task loss as the green-shaded region. Therefore, we define the evaluation function for a given weight value series $\mathbf{m}(\mathbf{w})$ as follows:

$$g(\mathbf{m}(\mathbf{w})) = -\mathrm{RMS}(\mathbf{m}(\mathbf{w}) - \mathbf{m}_\mathrm{I}), \tag{13}$$

where RMS is the root mean square.

Although (13) provides the evaluation value of the weight transitions, the waste inhomogeneity complicates directly

evaluating parameters $\mathbf{w}$ using (13). Consequently, different weight transitions can result from inhomogeneity, even when tasks are performed by the controller with identical parameters. Hence, we evaluate the parameters by the mean of the evaluation values obtained from multiple runs of the task. Specifically, consider a situation where $\pi(\mathbf{w})$ executes a task K times, and let $\mathbf{m}^{(k)}(\mathbf{w})$ denote the weight value series obtained for the kth time. Using the task evaluation function (13), we define the evaluation function for parameter $\mathbf{w}$:

$$f(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} g\left(\mathbf{m}^{(k)}(\mathbf{w})\right). \tag{14}$$

Note that (14) is used only when evaluating the parameters obtained by each estimation method. In the estimation phase, each method aims to obtain optimal parameter $\mathbf{w}^*$ from the evaluation values obtained by (13).

B. CRANE CONTROLLER

We consider controller $\pi(\cdot)$ for a scattering task with parameter $\mathbf{w} = [w_0, w_1]$. Parameters w_0 and w_1 correspond to the opening and closing times of the bucket's claws. Here the waste crane moves at a constant speed from a predefined starting point to an ending point while scattering the waste by repeatedly opening and closing its bucket's claws. Since the ideal movement in the scattering task is to continuously drop waste, we need to minimize the time the claws stop during the opening and closing actions. Hence, we restrict the space of the parameters we address to region $w_0 < w_1$ where the claws' opening time exceeds the closing time.

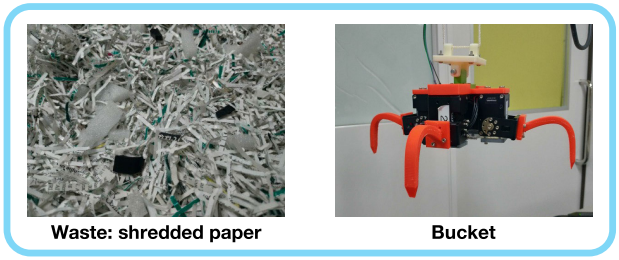
C. EXPERIMENT ENVIRONMENT

1) ROBOTIC WASTE CRANE SYSTEM

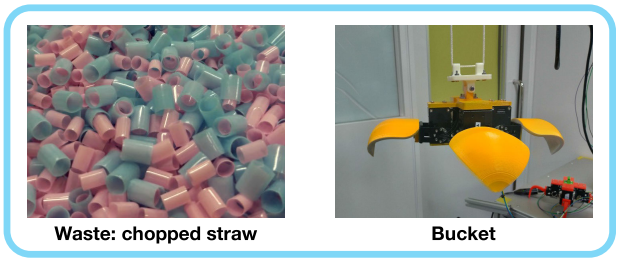
We developed a robotic waste crane system based on a previous study [2], as shown in Fig. 7. Our system uses a robot manipulator (Universal Robots UR5) for the crane and a force sensor (Robotiq FT 300) for the weight measurement. The bucket's four claws are each driven by a servo motor. A video camera (Microsoft Kinect v2) was also positioned in full view of the crane and the pseudo waste to record video for two-choice queries.

2) PSEUDO WASTE ENVIRONMENT

To validate the effectiveness of the proposed method in various environments, we prepared two environments with different characteristics (Fig. 8). The pseudo waste in Environment 1 (Env1) consists of a mixture of shredded paper and chopped packing material, characterized by its tendency to fall in clumps due to entanglement. Therefore, the bucket in Env1 has a claw shape suitable for grasping such waste. The pseudo waste in Environment 2 (Env2) consists of chopped straw, which resembles grains of sand that do not get tangled together. Hence, the bucket in Env2 is shaped to prevent the waste from spilling between the claws.

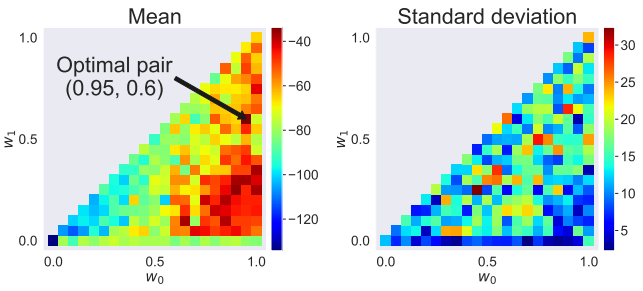


(a) Environment 1 (Env1)

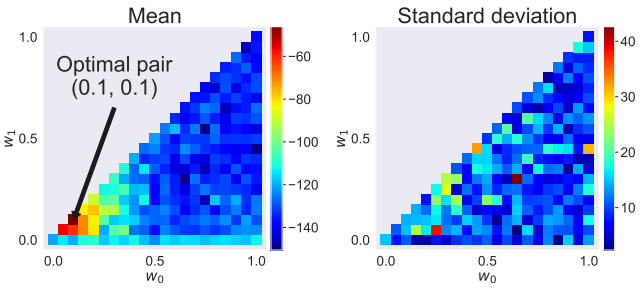


(b) Environment 2 (Env2)

FIGURE 8. Two pseudo waste environments.



(a) Env1: shredded paper



(b) Env2: chopped straw

FIGURE 9. Comparison of evaluation function values $g(\mathbf{m}(\mathbf{w}))$ in each pseudo waste environment: Mean and Standard Deviation are mean and standard deviation of five trials. We treat each Mean as evaluation function $f(\mathbf{w})$ value for each environment.

D. TASK SETTINGS

In the following experiments with both *simulated evaluators* and actual human evaluators, we limited the opening and closing times of the crane's claws as follows: $0\,\text{s} \leq w_0, w_1 \leq 1\,\text{s}$. The crane moves from the starting point to the ending point in 10 s, repeatedly opening and closing its claws. Moreover, we discretized w_0 and w_1 with 21 points each and restricted the parameter space (as noted in Section V-B), and so the number of combinations to explore is 231.

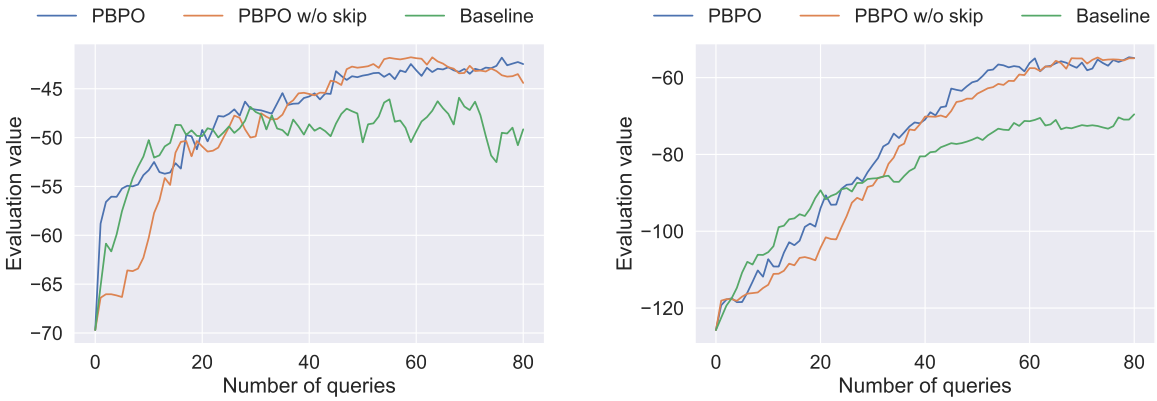


FIGURE 10. Transition of evaluation function $f(\mathbf{w})$ values of parameters acquired by `PBPO` and comparison methods from responses of *simulated evaluators* that return accurate answers: Horizontal axis represents number of two-choice queries, and vertical axis represents $f(\mathbf{w})$ value. $\mathbf{w}$ acquired by `PBPO` and `PBPO w/o skip` correspond to point of maximum posterior mean, and $\mathbf{w}$ acquired by `Baseline` corresponds to $\mathbf{w}^0$. Each solid line is drawn from average of 50 trials.

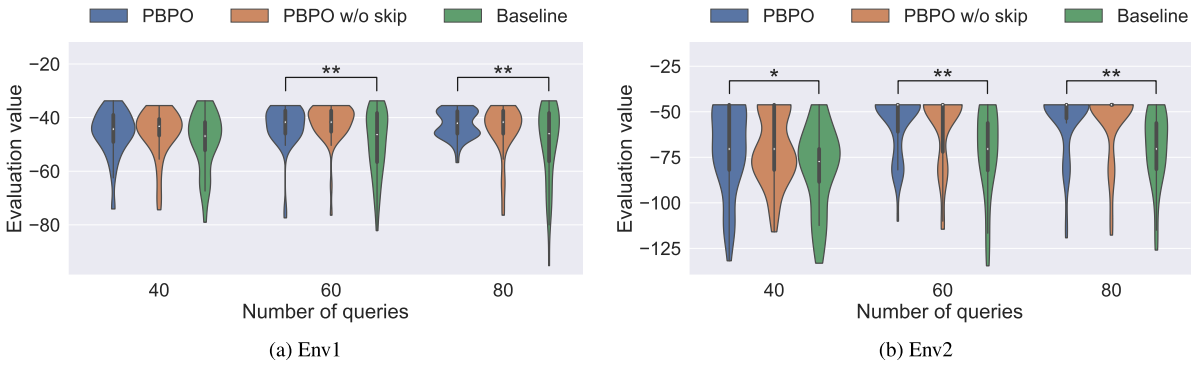


FIGURE 11. Violin plot of evaluation function $f(\mathbf{w})$ values according to parameters acquired at 40, 60, and 80 two-choice queries to *simulated evaluator* that returns accurate answers: Number of trials for each is 50. White dots indicate median, and asterisks indicate statistically significant differences (∗: $p < .05$, ∗∗: $p < .01$).

To minimize the time required for the human evaluators to participate in the experiment, we pre-collected the crane motions to be presented to them. Before the experiment, the crane was operated five times in each environment for each parameter to collect weight series and video clips. Then in the experiment, one of the five candidate motions, all corresponding to the same parameter, was randomly selected and presented to them. In all the experiments described in this paper, the crane motion corresponding to a parameter was randomly extracted from these collected sequences. We also obtained evaluation function $f(\mathbf{w})$ value from these collected series, assuming $K = 5$. In Fig. 9, the Mean and Standard Deviation represent the mean and standard deviation of $g(\mathbf{m}(\mathbf{w}))$ for five trials, and the color of the heatmap changes from blue to red as the values increase. Here Mean represents $f(\mathbf{w})$. Furthermore, $(0.95, 0.6)$ and $(0.1, 0.1)$, indicated by the black arrows in Mean, represent optimal parameter $\mathbf{w}^*$ for Env1 and Env2, and the respective optimal evaluation function values $f(\mathbf{w}^*)$ are -33.7 and -46.2.

Comparing the Means for each environment, we found that Env1 has high values over a wide area, and Env2 has such values only near $\mathbf{w}^*$. Comparing the Standard Deviations shows that Env1 tends to have higher values than Env2. PBPO seems to struggle more to acquire optimal parameters for Env2 than Env1 because it is more likely to encounter queries with choices that have both low evaluation values in Env2. Finally, the mean and standard deviations of $g(\mathbf{m}(\mathbf{w}))$ across $\mathcal{W}$ for each environment were -69.7 and 23.4 for Env1 and -125.7 and 20.0 for Env2.

VI. EXPERIMENTS WITH SIMULATED EVALUATORS

A. SETTINGS

Prior to experiments with actual human evaluators, we verified the effectiveness of our proposed method using a program called *simulated evaluators*, which returns a preference decision for a two-choice query. In this experiment, we did 1) a performance validation of the control parameters acquired by PBPO from *simulated evaluator* responses and

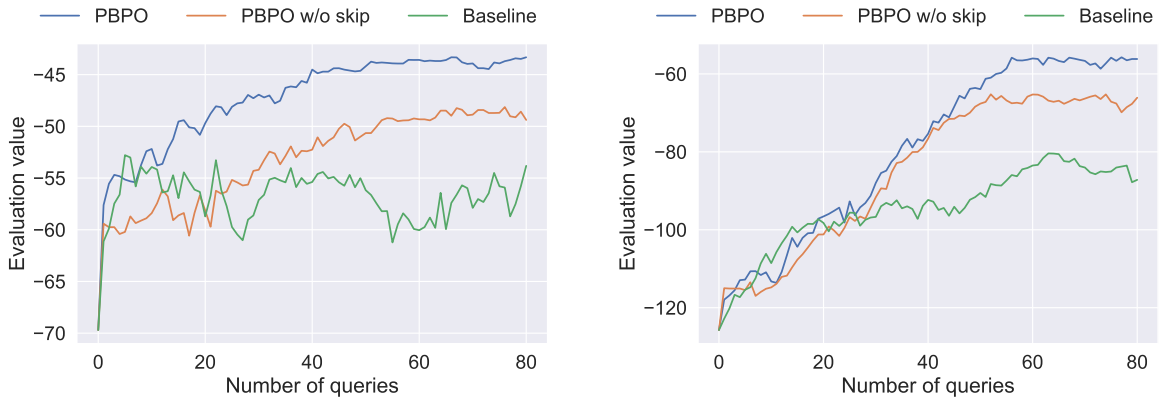


FIGURE 12. Transition of evaluation function $f(\mathbf{w})$ values of parameters acquired by `PBPO` and comparison methods from responses of *simulated evaluators* that probabilistically return an inaccurate answer: Horizontal axis represents number of two-choice queries, and vertical axis represents $f(\mathbf{w})$ value. $\mathbf{w}$ acquired by `PBPO` and `PBPO w/o skip` correspond to point of maximum posterior mean, and $\mathbf{w}$ acquired by `Baseline` corresponds to $\mathbf{w}^0$. Each solid line is drawn from average of 50 trials.

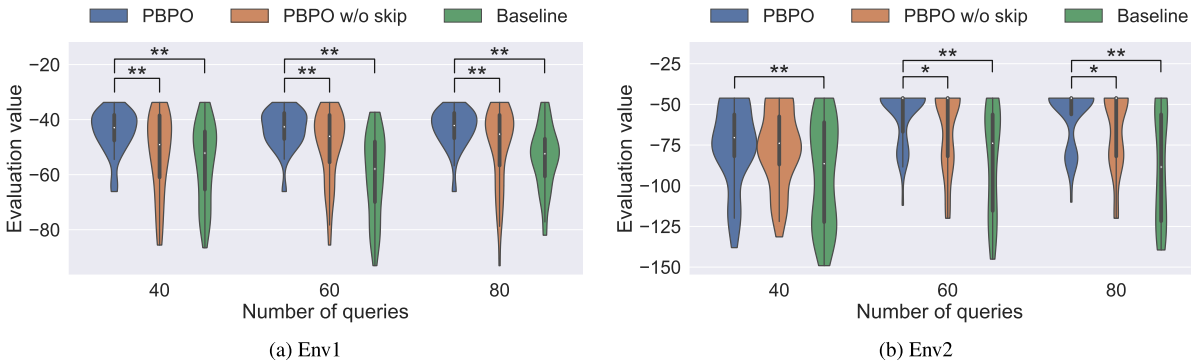


FIGURE 13. Violin plot of evaluation function $f(\mathbf{w})$ values based on parameters acquired at 40, 60, and 80 two-choice queries to *simulated evaluators* that probabilistically return inaccurate answers: Number of trials for each is 50. White dots indicate median, and asterisks indicate statistically significant differences (*: $p < .05$, **: $p < .01$).

2) an ablation study to verify the query synthesis's effectiveness when the *simulated evaluators* return uncertain answers. In the following, we describe the details of the *simulated evaluators* and the comparison methods.

1) SIMULATED EVALUATORS

We implemented *simulated evaluators* (introduced as a simulated user in Kwon et al. [39]) that answer preferences for two-choice queries not from videos but directly from evaluation values $g(\mathbf{m}(\mathbf{w}))$. We investigated the effect of uncertainty in human responses [40] on acquisition performance by preparing the following two *simulated evaluators*:

- `Certain`: evaluators that return the following accurate answers to a two-choice query $(\mathbf{w}^0, \mathbf{w}^1)$:

$$y = \begin{cases} 0, & \text{if } g(\mathbf{m}(\mathbf{w}^0)) \geq g(\mathbf{m}(\mathbf{w}^1)) \\ 1, & \text{if } g(\mathbf{m}(\mathbf{w}^0)) < g(\mathbf{m}(\mathbf{w}^1)) \end{cases}, \tag{15}$$

- `Uncertain`: evaluators that often return inaccurate answers based on the Bradley-Terry model [41]. The probability of responding with $y = 0$ to $(\mathbf{w}^0, \mathbf{w}^1)$ is defined as follows:

$$p(y=0) = \frac{\exp\left(\eta g(\mathbf{m}(\mathbf{w}^0))\right)}{\exp\left(\eta g(\mathbf{m}(\mathbf{w}^0))\right) + \exp\left(\eta g(\mathbf{m}(\mathbf{w}^1))\right)}, \tag{16}$$

where η is a positive constant that adjusts the uncertainty of the responses, and we set η to 0.08.

2) COMPARISON METHODS

We validated PBPO's effectiveness and the query synthesis itself by comparing its acquisition performance to the following two methods:

- `Baseline`: A query generation method based on a knockout algorithm (Appendix A) and
- `PBPO w/o skip`: a PBPO that does not give the evaluators the option to skip answers.

For consistency of notation, we also refer to PBPO using the query synthesis as `PBPO` within the experimental section.

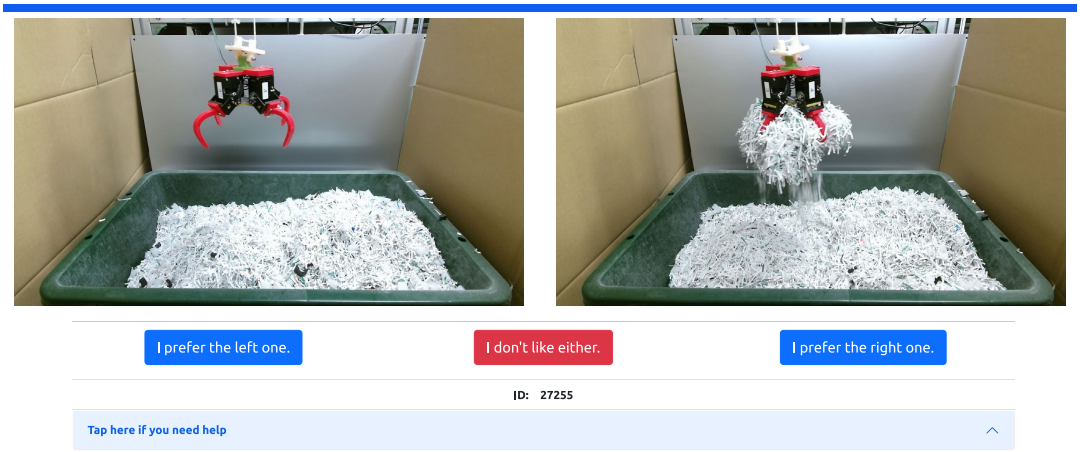


FIGURE 14. User interface for experiments with human evaluators: They compared two videos and then pressed blue button below the video they thought accomplished the task better. If they are allowed to skip an answer, they can also skip a query by pressing red button.

Here for both PBPO w/o skip and PBPO, we used an RBF kernel with length scale l of 0.01 as the GP kernel and set the ϵ of the likelihood (7) to 0.01 for Env1 and 0.0001 for Env2, respectively. In addition, until five unskipped responses have been accumulated, each method generates a two-choice query that compares randomly selected parameters.

Finally, we denote the parameter that was acquired by each method as $\mathbf{w}_{\text{acq}}$ in every step of the experiment. We define $\mathbf{w}_{\text{acq}}$ for PBPO w/o skip and PBPO as the parameter that maximizes the mean of the posterior (9) at each step. Also we define $\mathbf{w}_{\text{acq}}$ for the Baseline as the parameter selected by the evaluator in the previous query. Note that since $\mathcal{W}$ is discretized in this experiment, $\mathbf{w}_{\text{acq}}$ is one of the discretized candidate points.

B. RESULTS

1) RESULTS WITH CERTAIN

We review the results when Certain is used as the *simulated evaluator* whose response contains no uncertainty. We collected data for 50 trials for each PBPO, PBPO w/o skip, and Baseline.

a: TRANSITION OF EVALUATION VALUES

Fig. 10 shows the transition of evaluation function $f(\mathbf{w})$ values of the parameters acquired by each method based on the number of queries, and each solid line represents the mean of 50 trials. According to Fig. 10, ttPBPO obtained parameters with higher evaluated values than Baseline after the 30th query in both environments. On the other hand, we found no difference between PBPO and PBPO w/o skip.

b: COMPARISON OF VIOLIN PLOTS

Fig. 11 shows a violin plot representing $f(\mathbf{w})$ values of the parameters acquired by each method at the 40th, 60th, and 80th queries. Since the interquartile range (IQR) of the Baseline is wider than the others in Fig. 11 (a), ttBaseline acquired a relatively large number of parameters with low evaluation values in Env1. Furthermore, Fig. 11 (b), which is the result of Env2, shows that the median of PBPO and PBPO w/o skip reached optimal evaluation function value $f(\mathbf{w}^*)$ at the 60th and 80th times; the median of Baseline is relatively low, and the IQR is wider. Here the t-test results show significant differences between PBPO and Baseline at the 60th ($p = 0.0029 < 0.01$) and 80th ($p = 0.0023 < 0.01$) times for Env1 and at the 40th ($p = 0.043 < 0.05$), 60th ($p = 0.00027 < 0.01$), and 80th ($p = 0.00018 < 0.01$) times for Env2.

These results indicate that PBPO has the potential to acquire parameters with high evaluation values from accurate responses.

2) RESULTS WITH UNCERTAIN

We review the results when Uncertain is used as the *simulated evaluators* whose responses contain probabilistic inaccuracies. We collected data for 50 trials for each method, as in Section VI-B1.

a: TRANSITION OF EVALUATION VALUES

Fig. 12 shows the $f(\mathbf{w})$ values of the acquired parameters against the number of queries. According to Fig. 12, there is a difference in the evaluation values between PBPO and PBPO w/o skip in both environments when the *simulated evaluators* probabilistically return inaccurate answers, differing from the case when the response is certain.

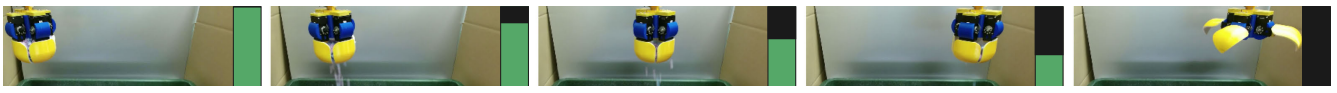
b: COMPARISON OF VIOLIN PLOTS

Fig. 13 compares the $f(\mathbf{w})$ values of the parameters acquired by each method at the 40th, 60th, and 80th queries. According to Fig. 13, ttPBPO has a narrower IQR than PBPO w/o skip at the 60th and 80th queries in both environments. Also, Fig. 13(a) shows a difference in the median value

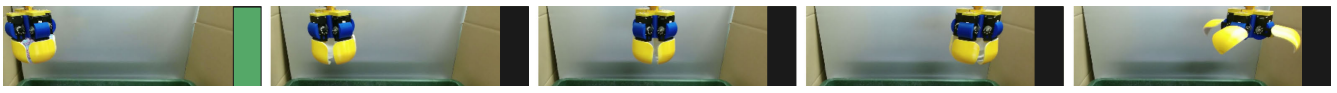
(a) Scattering behavior with a parameter acquired by PBPO ($\mathbf{w} = (0.9, 0.25)$)

(b) Scattering behavior with an initial parameter ($\mathbf{w} = (1.0, 0.05)$)

FIGURE 15. Comparison of scattering behavior based on a parameter acquired by PBPO (upper) and an initial parameter (lower) in Env1: Each behavior is represented by series of five photos equally spaced in time. To display weight transition of grasped waste, normalized weight is shown to right of each image. Green bar corresponds to grasped weight at each time point.



(a) Scattering behavior with parameters acquired by PBPO ($\mathbf{w} = (0.1, 0.1)$)



(b) Scattering behavior with initial parameter ($\mathbf{w} = (0.75, 0.2)$)

FIGURE 16. Comparison of scattering behavior based on a parameter acquired by PBPO (upper) and an initial parameter (lower) in Env2: Each behavior is represented by series of five photos equally spaced in time. To display weight transition of grasped waste, normalized weight is shown to right of each image. Green bar corresponds to grasped weight at each time point.

between PBPO and PBPO w/o skip in Env1. The t-test results show significant differences between PBPO and PBPO w/o skip at the 40th ($p = 0.0011 < 0.01$), 60th ($p = 0.0061 < 0.01$), and 80th ($p = 0.0073 < 0.01$) times for Env1 and at the 60th ($p = 0.029 < 0.05$) and 80th ($p = 0.019 < 0.05$) times for Env2.

These results indicate that the proposed method executes stable estimates even in situations where the evaluators frequently provide contradictory answers by mitigating the impact of these through query skipping and synthesis.

VII. EXPERIMENTS WITH HUMAN EVALUATORS

A. SETTINGS

We verified the effectiveness of our proposed method with actual human evaluators. All the experiments were conducted under the approval of the Ethics Committee of Nara Institute of Science and Technology. In this experiment, we conducted the following two validations: 1) an ablation study to verify the effectiveness of the query synthesis on actual human evaluators and 2) a comparison of acquisition performance between PBPO and methods that explicitly use evaluation functions [2]. The following sections describe the flow of the experiment and the details of each validation.

1) EXPERIMENTAL PROCEDURE

This experiment consists of the following two steps:

1) **Instructions to human evaluators**: We instructed the human evaluators about the scattering task, its purpose, and the characteristics of the waste. The text we distributed to them is found in Appendix B. In addition, we randomly selected 100 videos for each environment from those collected in Section V-C and requested that the evaluators watch them to gain a better understanding of the simulated environment and the task. We especially asked them to focus on the characteristics of waste in each environment and the type of crane movement that causes it to fall.
2) **Interactive two-choice query**: We instructed the human evaluators to answer two-choice queries in the user interface (UI) shown in Fig. 14. They reviewed the two videos presented and answered a two-choice query by pressing the blue button below the video where they relatively accomplished the task. The red button between the blue buttons allows the evaluators to skip answering a query. We set the number of two-choice queries per trial to 30 and included a two-minute break after each trial.

2) ABLATION STUDIES

We conducted an ablation study to inspect the effectiveness of the answer skipping and query synthesis mechanisms. It compared the following three methods:

- PBPO: Proposed method: We instructed the actual human evaluators to use the skip button only if both the task achievements of the two videos were undesirable. We also changed the button's label on the UI to "I don't like either."

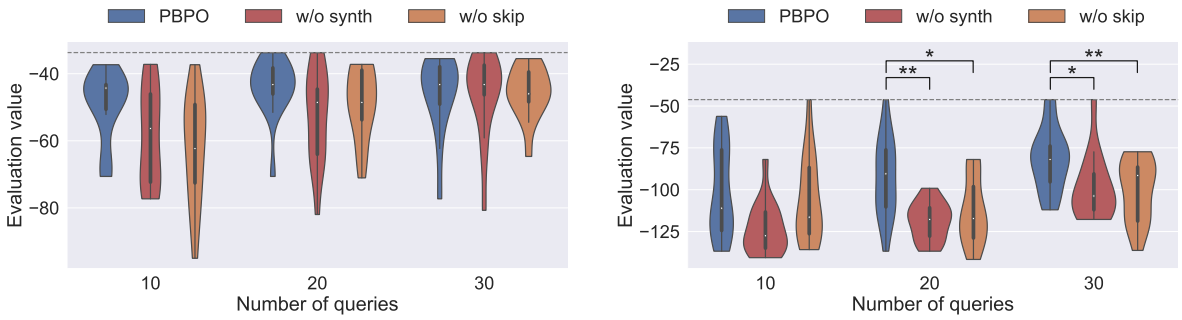


FIGURE 17. Violin plots illustrating evaluation function $f(\mathbf{w})$ values corresponding to parameters acquired by `PBPO`, `PBPO w/o synth`, and `PBPO w/o skip` using up to 10th, 20th, and 30th answers to two-choice queries: Each violin plot is drawn from results of 15 trials, where black dashed line represents evaluation function value with optimal parameter $f(\mathbf{w}^*)$. Asterisks indicate statistically significant differences (∗: $p < .05$, ∗∗: $p < .01$).

- `PBPO w/o synth`: A PBPO that removes the query synthesis mechanism: Although the evaluators can skip queries, the skipped queries are not synthesized. We instructed the evaluators to use the skip button only if the task achievements of the two videos were equal. We also changed the button's label on the UI to "It's hard to choose."
- `PBPO w/o skip`: A PBPO that gives the evaluators no option to skip answers: We also removed the red skip button from the UI.

We performed a three-trial experiment for each method on five evaluators (ages: 22–29), and so the total number of trials for each method is 15. The hyperparameters for each method were set as in Section VI-A.

3) COMPARISON WITH BOS THAT RELY ON EVALUATION FUNCTIONS

We compared our `PBPO` to previous methods that required a predefined evaluation function as a performance baseline. We again note that our `PBPO` method achieves automation without such a predefined evaluation function. We prepared a robust BO (RBO [2]) and a BO as our comparison methods. Concerning the details of the comparison method, we chose Upper Confidence Bound (UCB) [42] as an acquisition function and a GP with a RBF kernel as the surrogate model. We set hyperparameters κ of UCB and length scale l of the GP as follows: $\kappa = 2.0, l = 0.1$ for BO in Env1, $\kappa = 0.1, l = 0.01$ for RBO in Env1, $\kappa = 4.0, l = 0.0001$ for BO in Env2, and $\kappa = 4.0, l = 0.001$ for RBO in Env2. We also randomly selected the first five queries of the trials and set the number of trials of both methods in each environment to 30.

B. RESULTS

1) ABLATION STUDIES

a: CRANE BEHAVIOR WITH AN ACQUIRED PARAMETER

We first compared the scattering behavior with parameters acquired by `PBPO` and the behavior with the unoptimized initial parameters (i.e., random parameters). Figs. 15 and 16 compare the behavior in Env1 and Env2. To clearly express the change in waste weight, we show the current grasped weight as a green bar graph on the right side of each image. These bars are normalized by the weight of the initial grasp. Comparing the scattering behavior, the parameters acquired by the `PBPO` produce a desirable scattering behavior that causes the waste to drop gradually. In contrast, the initial parameters produce an undesirable behavior that causes most waste to drop in the early stages of the behavior. The parameters acquired in Env1 and Env2 with the highest evaluation function $f(\mathbf{w})$ values were $(0.9, 0.25)$ and $(0.1, 0.1)$, with corresponding $f(\mathbf{w})$ values of -35.5 and -46.2. Similarly, the acquired parameters with the lowest $f(\mathbf{w})$ values were $(1.0, 0.5)$ and $(0.25, 0.25)$, and the corresponding $f(\mathbf{w})$ values were -77.3 and -112.0. For reference, optimal parameters $\mathbf{w}^*$ for Env1 and Env2 were $(0.95, 0.6)$ and $(0.1, 0.1)$ and $f(\mathbf{w}^*)$ were -33.7 and -46.2.

b: PERFORMANCE COMPARISON OF FINALLY ACQUIRED PARAMETERS

We compared the $f(\mathbf{w})$ values that correspond to the parameters acquired by `PBPO`, `PBPO w/o synth`, and `PBPO w/o skip` to validate the effectiveness of the query synthesis. Fig. 17 shows the violin plots of $f(\mathbf{w})$ values corresponding to the parameters acquired from using the answers up to the 10th, 20th, and 30th queries in each environment. Each violin plot is drawn from the results of 15 trials. First, we confirmed the results of the parameters acquired using up to the 30th answers. Fig. 17 (a) shows no significant difference between the methods in Env1, although Fig. 17 (b) shows a difference between `PBPO` and the other two methods in Env2. The t-test results in Env2 show significant differences between `PBPO` and `PBPO w/o synth` ($p = 0.047 < 0.05$) and between `PBPO` and `PBPO w/o skip` ($p = 0.004 < 0.01$). These results indicate that `PBPO` can acquire parameters with high evaluation values through query synthesis even when obtaining optimal parameters is difficult due to the low evaluation values in most parameter spaces.

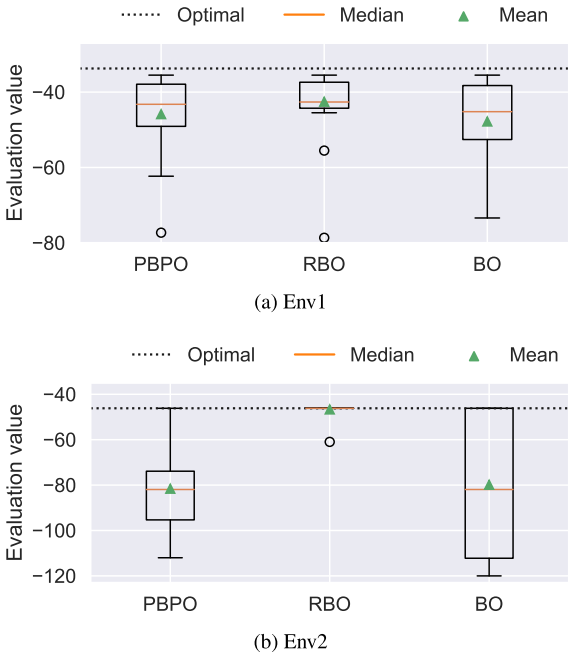


FIGURE 18. Box-and-whisker diagram showing evaluation function $f(\mathbf{w})$ values of parameters acquired by PBPO, RBO, and BO: Number of trials for PBPO is 15, and number of trials for RBO and BO is 30. PBPO results are based on parameters acquired at 60 crane operations (30 queries), and RBO and BO results are based on parameters after convergence. Green triangles represent mean values, solid orange lines represent median values, and black dashed lines represent $f(\mathbf{w}^*)$.

c: PERFORMANCE COMPARISON OF PARAMETERS ACQUIRED BY FEWER QUERIES

Finally, we compare the $f(\mathbf{w})$ values of the parameters acquired from the answers up to the 10th with those obtained from answers up to the 20th time. The shape of the violin plot in Fig. 17 (a) (the Env1 results) indicates that PBPO acquires more parameters with higher evaluation values than the other two methods. Subsequently, according to Fig. 17 (b) (the Env2 results), the parameters acquired by the PBPO are more widely distributed than the parameters by the other two methods, and the difference is more pronounced when compared to PBPO w/o synth. The t-test results in Env2 also indicate significant differences between PBPO and PBPO w/o synth ($p = 0.008 < 0.01$) and between PBPO and PBPO w/o skip ($p = 0.019 < 0.05$) when using up to the 20th response. These results indicate that using query synthesis allows for acquiring highly evaluated parameters with fewer queries than in the absence of query synthesis and skips.

All the results on the validation of query synthesis show that query synthesis works effectively and provides stable estimation, even when optimizing from actual human responses, which are often inconsistent.

2) COMPARISON WITH BOS THAT RELY ON EVALUATION FUNCTIONS

We compared evaluation function $f(\mathbf{w})$ values of the parameters acquired by PBPO, which optimizes the policy from the human evaluator's response, with RBO and BO, which explicitly require an evaluation function (13). We compared the three methods by the number of crane operations rather than the number of queries because the RBO and BO queries are composed of a single crane movement, while the two-choice query used for PBPO is composed of two of them.

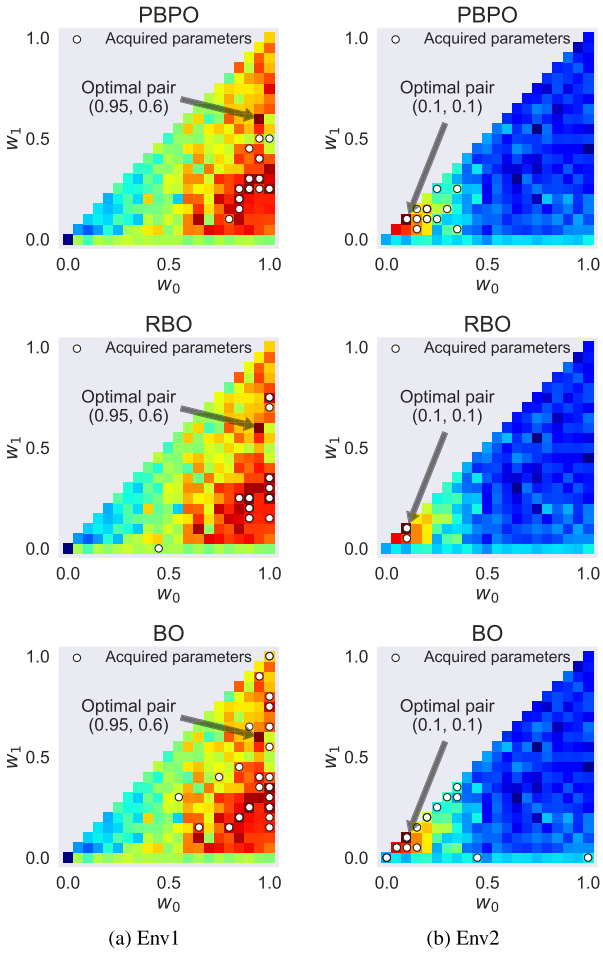
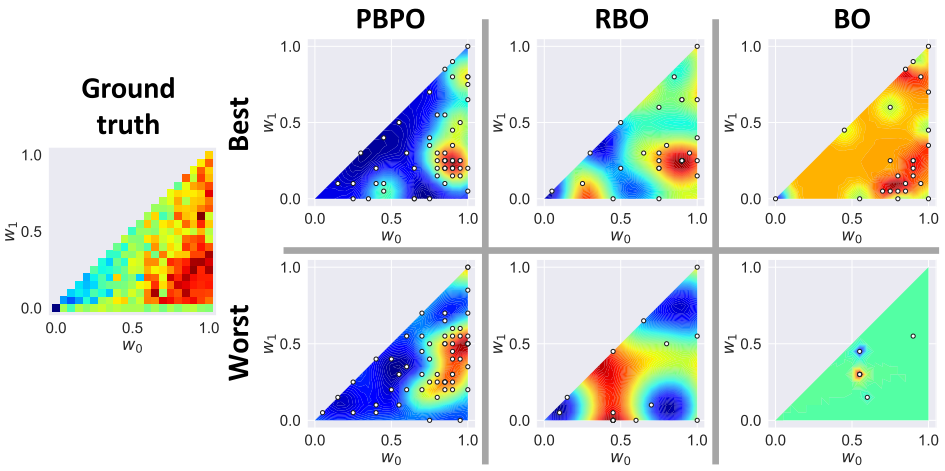


FIGURE 19. Scatter plots of parameters acquired by PBPO, RBO, and BO in each environment: Heatmap represents evaluation function $f(\mathbf{w})$ value, and white dots represent acquired parameter pairs. Number of trials for PBPO is 15, and number of trials for RBO and BO is 30.

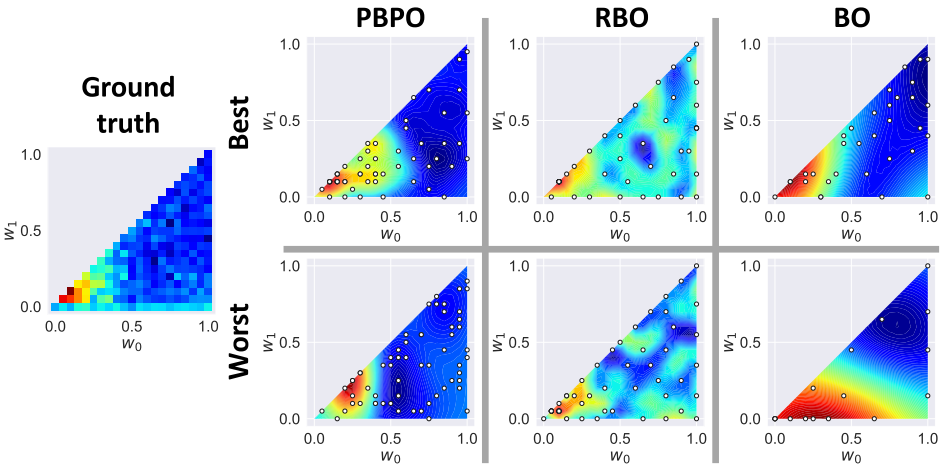
a: PERFORMANCE COMPARISON OF FINALLY ACQUIRED PARAMETERS

Fig. 18 shows a box-and-whisker diagram of the $f(\mathbf{w})$ values corresponding to the parameters acquired by each method. According to Fig. 18 (a) (the Env1 results), the mean (green triangles) and median (solid orange lines) of the $f(\mathbf{w})$ values of RBO are slightly higher than those of PBPO and BO. In contrast, the maximum does not differ significantly among the methods. The mean values for PBPO, RBO, and BO in Env1 are respectively −45.8, −41.6, and −47.8.

Next Fig. 18 (b), which represents the results of Env2, a more challenging environment for PBPO, shows that although there is a significant difference between PBPO and



(a) Env1



(b) Env2

FIGURE 20. Comparison of posterior distribution of GPs used in surrogate models for PBPO, RBO, and BO with Ground truth $f(\mathbf{w})$: Upper (Best) and lower (Worst) figures on right-hand side of each figure show highest and lowest $f(\mathbf{w})$ values among parameters acquired by each method. White dots indicate query points in each trial.

RBO, whose mean and median are located on the black dashed line representing optimal evaluation value $f(\mathbf{w}^*)$, the average is equivalent to BO. The fact that the maximum of PBPO reaches $f(\mathbf{w}^*)$ indicates that there were trials in which the optimal parameters were acquired. The means of PBPO, RBO, and BO in Env2 are respectively -81.6, -46.7, and -79.8.

b: COMPARISON OF ACQUIRED PARAMETER DISTRIBUTIONS

Fig. 19 shows the parameters acquired by each method on parameter space $\mathcal{W}$. The heatmaps in Figs. Fig. 19 (a) and Fig. 19 (b) correspond to evaluation function $f(\mathbf{w})$ values in Env1 and Env2, and the "Optimal pair" in the figures point to optimal parameter $\mathbf{w}^*$. Since we discretized $\mathcal{W}$ to generate candidates, as described in Section V-C, some of the white dots in each figure representing the acquired parameters are shown as overlapping dots. However, 15 dots were originally plotted for PBPO and 30 dots for the other two methods. Fig. 19 (a) shows that all the methods failed to acquire $\mathbf{w}^*$ in Env1, and the parameters acquired by PBPO and RBO are concentrated in the region of higher evaluation values compared to those acquired by BO. Furthermore, Fig. 19 (b) (the Env2 results) indicates that the parameters acquired by RBO are concentrated in the region of high evaluation values. In fact, RBO acquired optimal parameter $\mathbf{w}^*$ in 29 of 30 trials. A comparison of the parameters acquired with PBPO and BO shows that some were acquired where BO converged on the candidate points with low evaluation values, such as $(w_0, w_1) = (1.0, 0.0)$. The parameters acquired with PBPO are concentrated in regions with relatively high evaluation values around $\mathbf{w}^*$.

c: COMPARISON OF POSTERIORS

At the end of the results of experiments with actual human evaluators, we compared the posterior distribution of the GP used in the surrogate models of the three BO methods with

evaluation function Ground truth (GT) $f(\mathbf{w})$. Fig. 20 shows the GT and the posterior mean for each method at 60 crane operations; the color of the heatmap changes from blue to red as the mean value increases. The ''Best'' in the figure represents the posterior distribution in which the parameter was acquired with the highest evaluation value. Similarly, the ''Worst'' represents the posterior distribution in which the parameter was acquired with the lowest evaluation value. As shown in Fig. 20 (a), in Env1, both the ''Best'' and ''Worst'' of BO differ significantly from GT, and the ''Worst'' of RBO does not correctly represent the lower right region, which has high GT values. On the other hand, both PBPO's ''Best'' and ''Worst'' generally well represent regions with high evaluation values on GT. According to Fig. 20 (b) (the Env2 results), all the posterior distributions for PBPO and RBO have higher means only around the optimal parameter, although the ''Worst'' for BO has higher values over a wider region.

These results indicate that PBPO's acquisition performance is comparable to methods that explicitly require a predefined evaluation function.

VIII. DISCUSSION

We attempted to solve the following problems to achieve preference-based waste crane automation: 1) how to acquire an optimal control policy with a small number of queries and 2) how to deal with cases where making a clear decision is difficult due to the characteristics of the waste.

We solved 1) by optimizing the controller parameters using a GP-based PbBO. Experimental results show that our PBPO can acquire parameters with high evaluation values from just a few dozen two-choice queries. Since similar deep learning-based approaches [22], [23] require hundreds to thousands of queries, our proposed method is relatively query efficient.

In addition, we used query skipping and synthesis mechanisms to solve 2). Experimental results with *simulated evaluators* and human evaluators demonstrate the effectiveness of query synthesis under the realistic assumption that responses contain uncertainty. On the other hand, depending on the task settings, the possibility that a human evaluator will skip answering all the queries increases. Generating early queries considering state entropy [23] rather than randomly generating them may alleviate this problem. Another approach to 2) is to present carefully selected queries to the evaluators, a direction discussed by several studies [43], [44]. An approach that integrates the responses of multiple evaluators [45], [46] may also be effective for this problem.

We also discuss the performance differences in Env2 presented in VII-B2. In Env2, PBPO and BO show significantly lower acquisition performance than RBO, results that suggest that considering the robustness in Env2 is effective. Fig. 9 shows that only a narrower region has higher values in Env2 compared to Env1. Therefore, the performance degradation is more severe in Env2 when the parameters near the optimal ones are incorrectly acquired due to the variance of the evaluation values caused by the waste diversity. Since RBO is less sensitive to variations in the evaluation values, we assume that optimal parameters can be acquired. Part of our future work will add robustness against outliers to our proposed method.

IX. CONCLUSION

We proposed Preferential Bayesian Policy Optimization (PBPO) as a novel heavy equipment automation method without an evaluation function design. PBPO obtains optimal control policies from an evaluator's responses to interactive two-choice queries. The proposed method generates queries by a sample-efficient Preference-based Bayesian optimization (PbBO) to cope with the high trial-and-error cost of heavy equipment. Furthermore, to deal with difficult-to-judge queries for evaluators, PBPO provides the option of skipping answers and employs a query synthesis mechanism that synthesizes new preference relations using the skipped queries.

We also conducted experiments with *simulated evaluators* and actual human evaluators in a scattering task where the evaluation function is explicitly defined to verify the proposed method's effectiveness. Experimental results with *simulated evaluators* indicate that the query synthesis mechanism is effective when their answers contain probabilistic mistakes. The results with actual human evaluators show that the control performance of the parameters acquired by PBPO is comparable to RBO and BO in the first environment and comparable to BO in the second environment, which is more difficult for PBPO.

APPENDIX A
BASELINE METHOD: KNOCKOUT PAIRWISE COMPARISON

The comparison method used in the experiments with *simulated evaluators* in Section VI follows Algorithm 1. For each two-choice query except the first one, the parameter selected by the evaluator in the previous query was compared with a randomly selected parameter. For the first two-choice query, two randomly selected parameters were compared.

APPENDIX B
INSTRUCTIONS PROVIDED TO ACTUAL HUMAN EVALUATORS

Background:

- Waste-to-energy plants are expected to burn waste at a constant temperature, although the diversity of the transported waste complicates that goal.
- To address this problem, the plant uses cranes equipped with buckets to agitate the waste in a pit where it is temporarily accumulated.
- Waste is mixed by opening and closing the crane's claws as it moves to scatter the grabbed waste. Optimal claw action allows waste to fall steadily during the crane's movement.
- Note that due to waste uncertainty (i.e., waste diversity), an identical claw action may not cause identical scatter.

Algorithm 1 Knockout Pairwise Comparison

input: Parameter space $\mathcal{W}$, controller $\pi(\cdot)$, query Q
output: Acquired parameter $\mathbf{w}_{\text{acq}}$

for each iteration **do**
  if iteration $==0$ **then**
    Sample $\mathbf{w}^0$ from $\mathcal{W}$ uniformly
  else
    $\mathbf{w}^0 \leftarrow \mathbf{w}_{\text{acq}}$
  end if
  Sample $\mathbf{w}^1$ from $\mathcal{W}$ uniformly
  $Q \leftarrow \phi$
  for $i = 0$ to 2 **do**
    Set $\mathbf{w}^i$ for $\pi(\cdot)$
    Perform task with $\pi(\mathbf{w}^i)$ to generate τ^i
    $Q \leftarrow Q \cap \tau^i$
  end for
  Present Q to evaluator
  Get answer y from evaluator
  if $y==0$ **then**
    $\mathbf{w}_{\text{acq}} \leftarrow \mathbf{w}^0$
  else
    $\mathbf{w}_{\text{acq}} \leftarrow \mathbf{w}^1$
  end if
end for

This experiment asks you to do two tasks:

1) Watch videos of a waste crane to check the waste and crane characteristics.
2) Answer repeatedly presented two-choice queries.

Here are some additional aspects to bear in mind when answering the queries:

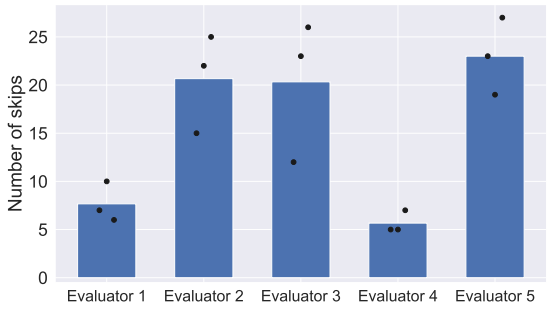
- All of the grabbed waste should drop when the crane finishes moving.
- Avoid situations where the waste falls together at the beginning or end of scattering.
- Pay attention not only to the falling waste but also to the claw movements. Note which claw movement is more likely to cause the waste to drop.
- The amount of waste initially grabbed by the crane or the surface geometry of the waste in the pit should not affect your evaluation.

In some experiments, a button appears that allows you to skip answers. The two types of skip buttons have different usage conditions:

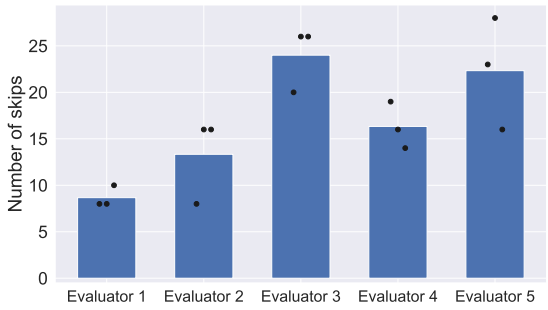
- "I don't like either.": when both behaviors in a query are undesirable.
- "It's hard to choose.": when prioritizing behaviors in a query is complicated.

APPENDIX C
COMPARISON OF NUMBER OF SKIPS

Fig. 21 shows the number of skipped answers per human evaluator in the experiment using PBPO. The results show that the number of query skips by some evaluators is not limited to a few but reaches dozens. The results identify individual differences in the criteria for query skipping. For example, Evaluator 5 often skips; Evaluator 1 is reluctant.



(a) Env1



(b) Env2

FIGURE 21. Comparison of number of skipped answers per actual human evaluators in each environment: Number of two-choice queries in each trial is 30, and black dots represent number of skips in each trial.

REFERENCES

[1] K. J. Mackin and M. Fujiyoshi, "Intelligent waste crane scheduling using evolutionary computation," in *Proc. Joint 10th Int. Conf. Soft Comput. Intell. Syst. (SCIS), 19th Int. Symp. Adv. Intell. Syst. (ISIS)*, Dec. 2018, pp. 689–692.
[2] H. Sasaki, T. Hirabayashi, K. Kawabata, Y. Onuki, and T. Matsubara, "Bayesian policy optimization for waste crane with garbage inhomogeneity," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4533–4540, Jul. 2020.
[3] H. Sasaki, T. Hirabayashi, K. Kawabata, and T. Matsubara, "Gaussian process self-triggered policy search in weakly observable environments," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 5946–5952.
[4] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," 2016, *arXiv:1606.06565*.
[5] H. Zhu, J. Yu, A. Gupta, D. Shah, K. Hartikainen, A. Singh, V. Kumar, and S. Levine, "The ingredients of real world robotic reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–20.
[6] E. Siivola, A. K. Dhaka, M. R. Andersen, J. González, P. G. Moreno, and A. Vehtari, "Preferential batch Bayesian optimization," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2021, pp. 1–6.
[7] J. González, Z. Dai, A. Damianou, and N. D. Lawrence, "Preferential Bayesian optimization," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1282–1291.
[8] J. S. Lee, Y. Ham, H. Park, and J. Kim, "Challenges, tasks, and opportunities in teleoperation of excavator toward human-in-the-loop construction automation," *Autom. Construct.*, vol. 135, pp. 104–119, Mar. 2022.
[9] S. Dadhich, U. Bodin, and U. Andersson, "Key challenges in automation of earth-moving machines," *Autom. Construct.*, vol. 68, pp. 212–222, Aug. 2016.

[10] A. A. Dobson, J. A. Marshall, and J. Larsson, "Admittance control for robotic loading: Design and experiments with a 1-tonne loader and a 14-tonne load-haul-dump machine," *J. Field Robot.*, vol. 34, no. 1, pp. 123–150, Jan. 2017.

[11] R. J. Sandzimier and H. H. Asada, "A data-driven approach to prediction and optimal bucket-filling control for autonomous excavators," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2682–2689, Apr. 2020.

[12] H. Maske, E. Kieson, G. Chowdhary, and C. Abramson, "Learning task-based instructional policy for excavator-like robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1962–1969.

[13] P. Egli and M. Hutter, "A general approach for the automation of hydraulic excavator arms using reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5679–5686, Apr. 2022.

[14] H. Tahara, H. Sasaki, H. Oh, B. Michael, and T. Matsubara, "Disturbance-injected robust imitation learning with task achievement," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2466–2472.

[15] L. Ramli, Z. Mohamed, A. M. Abdullahi, H. I. Jaafar, and I. M. Lazim, "Control strategies for crane systems: A comprehensive review," *Mech. Syst. Signal Process.*, vol. 95, pp. 1–23, Oct. 2017.

[16] S. Kang and E. Miranda, "Planning and visualization for automated robotic crane erection processes in construction," *Autom. Construct.*, vol. 15, no. 4, pp. 398–414, Jul. 2006.

[17] O. Sawodny, H. Aschemann, and S. Lahres, "An automated gantry crane as a large workspace robot," *Control Eng. Pract.*, vol. 10, no. 12, pp. 1323–1338, Dec. 2002.

[18] M. I. Solihin and A. Legowo, "Fuzzy-tuned PID anti-swing control of automatic gantry crane," *J. Vibrat. Control*, vol. 16, no. 1, pp. 127–145, Jan. 2010.

[19] H.-H. Lee, "A new approach for the anti-swing control of overhead cranes with high-speed load hoisting," *Int. J. Control*, vol. 76, no. 15, pp. 1493–1499, Jan. 2003.

[20] D.-H. Chun, M.-I. Roh, and H.-W. Lee, "Automation of crane control for block lifting based on deep reinforcement learning," *J. Comput. Des. Eng.*, vol. 9, no. 4, pp. 1430–1448, Aug. 2022.

[21] S. Cho and S. Han, "Reinforcement learning-based simulation and automation for tower crane 3D lift planning," *Autom. Construct.*, vol. 144, Dec. 2022, Art. no. 104620.

[22] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4302–4310.

[23] K. Lee, L. M. Smith, and P. Abbeel, "PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 6152–6163.

[24] Z. Cao, K. Wong, and C.-T. Lin, "Weak human preference supervision for deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5369–5378, Dec. 2021.

[25] J. Park, Y. Seo, J. Shin, H. Lee, P. Abbeel, and K. Lee, "SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–17.

[26] D. J. Hejna III and D. Sadigh, "Few-shot preference learning for human-in-the-loop RL," in *Proc. The 6th Conf. Robot Learn. (CoRL)*, vol. 205, 2023, pp. 2014–2025.

[27] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia, "Active preference-based learning of reward functions," *Proc. Robot., Sci. Syst. (RSS)*, vol. 13, Jul. 2017.

[28] E. Biyik, N. Huynh, M. J. Kochenderfer, and D. Sadigh, "Active preference-based Gaussian process regression for reward learning," *Robot., Sci. Syst. XVI*, vol. 16, May 2020.

[29] C. Basu, E. Biyik, Z. He, M. Singhal, and D. Sadigh, "Active learning of reward dynamics from hierarchical queries," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 120–127.

[30] M. Tucker, E. Novoseller, C. Kann, Y. Sui, Y. Yue, J. W. Burdick, and A. D. Ames, "Preference-based learning for exoskeleton gait optimization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 2351–2357.

[31] M. Tucker, M. Cheng, E. Novoseller, R. Cheng, Y. Yue, J. W. Burdick, and A. D. Ames, "Human preference-based learning for high-dimensional optimization of exoskeleton walking gaits," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 3423–3430.

[32] A. Bemporad and D. Piga, "Global optimization based on active preference learning with radial basis functions," *Mach. Learn.*, vol. 110, no. 2, pp. 417–448, Feb. 2021.

[33] M. Zhu, A. Bemporad, and D. Piga, "Preference-based MPC calibration," in *Proc. Eur. Control Conf. (ECC)*, Jun. 2021, pp. 638–645.

[34] L. Roveda, B. Maggioni, E. Marescotti, A. A. Shahid, A. M. Zanchettin, A. Bemporad, and D. Piga, "Pairwise preferences-based optimization of a path-based velocity planner in robotic sealing tasks," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6632–6639, Oct. 2021.

[35] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2, 2012, pp. 2951–2959.

[36] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2006.

[37] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 137–144.

[38] M. Opper and C. Archambeau, "The variational Gaussian approximation revisited," *Neural Comput.*, vol. 21, no. 3, pp. 786–792, Mar. 2009.

[39] Y. Kwon, Y. Tsurumine, T. Shimmura, S. Kawamura, and T. Matsubara, "Physically consistent preferential Bayesian optimization for food arrangement," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11863–11870, Oct. 2022.

[40] S. E. F. Chipman, *The Oxford Handbook of Cognitive Science*. Oxford, U.K.: Oxford Univ. Press, Oct. 2017.

[41] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

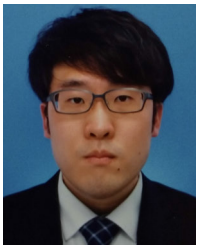
[42] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *J. Mach. Learn. Res.*, vol. 3, pp. 397–422, Nov. 2003.

[43] G. Canal, A. Massimino, M. Davenport, and C. Rozell, "Active embedding search via noisy paired comparisons," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, 2019, pp. 902–911.

[44] E. Biyik, M. Palan, N. C. Landolfi, D. P. Losey, and D. Sadigh, "Asking easy questions: A user-friendly approach to active reward learning," in *Proc. Conf. Robot Learn. (CoRL)*, vol. 100, 2020, pp. 1177–1190.

[45] E. Simpson and I. Gurevych, "Scalable Bayesian preference learning for crowds," *Mach. Learn.*, vol. 109, no. 4, pp. 689–718, Apr. 2020.

[46] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 193–202.



YUHWAN KWON (Member, IEEE) received the Ph.D. degree in information science from the Division of Information Science, Nara Institute of Science and Technology, Nara, Japan, in 2023. He is currently a specially-appointed Assistant Professor with the Robot Learning Laboratory, Nara Institute of Science and Technology. His research interests include machine learning, preference learning, and control theory for robotics.

HIKARU SASAKI (Member, IEEE) received the Ph.D. degree in information science from the Division of Information Science, Nara Institute of Science and Technology, Nara, Japan, in 2021. He is currently an Assistant Professor with the Robot Learning Laboratory, Nara Institute of Science and Technology. His research interests include machine learning, reinforcement learning, imitation learning, Bayesian statistics, and control theory for robotics.

TERUSHI HIRABAYASHI joined Hitachi Zosen Corporation, in 1993, and worked in the software design of electronic control systems. He worked on many projects in the development of a system for waste-to-energy plants: designing automation systems for the plants, in 1997, and the development of a remote monitoring system (remon) for the plants and setting up a remote monitoring center (remon center), in 2001. Since 2018, he has been the Development-Promoting Manager of an automated waste crane control system and has managed its improvement.

KAORU KAWABATA joined Hitachi Zosen Corporation, in 1990, and worked in electrical instrumentation design. He worked on many projects in the environmental business: long-term management of environmental facilities, in 2003, setting up and management of the Remote Operations Center (ROC) for waste incineration plants, in 2011, the development of a 24-hour remote monitoring system from the ROC, in 2012, the development of an image recognition system called CosMoS, in 2013, the development of a large-capacity distributed system for collection and monitoring plant data with JASRI, in 2015, and the development of an automatic combustion advanced system for waste incineration called Comp. ACC, in 2018. Since 2020, he has been an Engineering Leader of the Environmental Business Headquarters and Development Center, where he manages the development of new products and technologies.

TAKAMITSU MATSUBARA (Member, IEEE) received the Ph.D. degree in information science from the Nara Institute of Science and Technology, Nara, Japan, in 2007. From 2005 to 2007, he was a Research Fellow (DC1) of the Japan Society for the Promotion of Science. From 2013 to 2014, he was a Visiting Researcher with the Donders Institute for Brain Cognition and Behavior, Radboud University Nijmegen, Nijmegen, The Netherlands. He is currently a Professor with the Nara Institute of Science and Technology and a Visiting Researcher with ATR, Kyoto, Japan, and the AIST AI Center, Tokyo, Japan. His research interests include machine learning and control theory for robotics.