

Preference Aligned Visuomotor Diffusion Policies for Deformable Object Manipulation

Anonymous - Double Blind Submission

Abstract— Humans naturally develop preferences for how manipulation tasks should be performed, which are often subtle, personal, and difficult to articulate. Although it is important for robots to account for these preferences to increase personalization and user satisfaction, they remain largely underexplored in robotic manipulation, particularly in the context of deformable objects like garments and fabrics. In this work, we study how to adapt pretrained visuomotor diffusion policies to reflect preferred behaviors using limited demonstrations. We introduce RKO, a novel preference-alignment method that combines the benefits of two recent frameworks: RPO and KTO. We evaluate RKO against common preference learning frameworks, including these two, as well as a baseline vanilla diffusion policy, on real-world cloth-folding tasks spanning multiple garments and preference settings. We show that preference-aligned policies (particularly RKO) achieve superior performance and sample efficiency compared to standard diffusion policy finetuning. These results highlight the importance and feasibility of structured preference learning for scaling personalized robot behavior in complex deformable object manipulation tasks.

I. INTRODUCTION

As robots become more affordable and integrated into daily life, there is a growing need for adaptive behaviors that reflect individual user preferences [1]. Personalization often requires robots to learn directly from interaction [2], with users providing demonstrations of desired behaviors. In this context, a particularly relevant yet underexplored domain is Deformable Object Manipulation (DOM), involving everyday items like clothing, textiles, and food. Enabling robots to handle such objects with human-like dexterity while following human preferences would support applications in laundry folding, assisted dressing, feeding, and healthcare [3]. Personalization in DOM may involve different folding styles, dressing strategies tailored to mobility constraints, or household routines reflecting different preferences between users, as illustrated in Fig.1.

Deformable objects pose unique challenges compared to rigid ones due to their complex dynamics, high-dimensional state spaces, and varied physical properties [4]. These complexities make both manipulation and the expression of user preferences more difficult. While strategies for rigid objects can often be described verbally, the subtleties of tasks like garment folding are harder to articulate with words, making physical demonstrations a more natural modality for expressing preferences [5]. However, collecting such demonstrations is costly and time-consuming [6], highlighting the need for sample-efficient adaptation frameworks. Recent advances in data-driven learning have shown great progress in manipulation, including DOM, with robotic foundation models [7] and visuomotor diffusion policies [8]

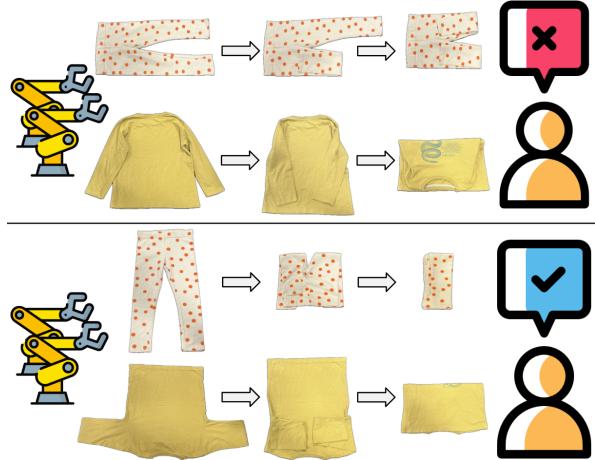


Fig. 1: Two different user preferences for folding the same garments demonstrate how variations in execution can reflect personal styles or practical needs. Capturing and aligning with such preferences is essential for enabling robots to perform personalized and user-aligned behaviors in deformable object manipulation tasks like cloth folding.

leveraging large-scale demonstration datasets to generalize across tasks. However, adapting these pretrained models to reflect user-specific manipulation preferences has received scarce attention. The challenge lies in aligning them to new behaviors without forgetting prior knowledge or requiring many new demonstrations, a common scenario where existing task demonstrations are available and could be leveraged to support the adaptation process to a new preferred behavior.

In this work, we address the problem of accounting for user preferences in deformable object manipulation. Specifically, we study how to align a pretrained visuomotor diffusion policy to new demonstrations reflecting a user’s preferred garment-folding strategy. While preference optimization techniques have seen success in domains like text-to-image generation [9], their application to robotic DOM remains limited. We investigate direct preference optimization (DPO) [10], along with two recent variants: relative preference optimization (RPO) [11], which leverages observation similarity for contrastive weighting, and Kahneman-Tversky Optimization (KTO) [12], which enables training from per-sample binary feedback instead of preference pairs. Building on these, we propose RKO, a novel method combining the strengths of KTO and RPO.

We evaluate RKO against standard preference optimization frameworks and vanilla diffusion policies. Results show that

preference-based methods, particularly RKO, achieve better alignment to user-preferred behaviors with fewer demonstrations, highlighting their sample efficiency and ability to retain knowledge from pretrained models.

The contributions of this paper are as follows:

- We introduce RKO, a novel method that combines the sample efficiency of KTO with the context-aware weighting of RPO.
- We present the first systematic comparison of preference optimization frameworks (DPO, RPO, KTO) for aligning pretrained diffusion policies to user-specific strategies in deformable object manipulation.
- We conduct extensive real-world evaluations on three garment types (trousers, sleeves, tshirt), each with multiple folding preference tasks, demonstrating the effectiveness and sample efficiency of preference-based alignment over vanilla diffusion policies.

II. RELATED WORK

We structure the related work along the areas of deformable object manipulation, visuomotor diffusion policies, as well as preference alignment frameworks and their applications in robotics.

A. Learning Deformable Object Manipulation Skills

Deformable object manipulation (DOM) poses unique challenges due to the high-dimensional, nonlinear dynamics of garments, ropes, and tissues [13]. Their virtually infinite degrees of freedom complicate perception and state estimation, while analytical deformation models are often computationally prohibitive [14]. Data-driven methods have therefore become increasingly prominent. Learning from demonstrations offers a scalable way to acquire complex skills without explicit physics-based modeling [15], [16], [17]. In this context, diffusion models [18] have shown strong ability to capture multimodal distributions across robotics tasks, including navigation [19], planning [20], and both rigid and deformable manipulation [21], [22]. Visuomotor diffusion policies learn a stochastic transport map from a simple prior (e.g., Gaussian noise) to a target distribution of action sequences conditioned on observations, enabling better generalization than traditional discriminative models [8], [23]. Their ability to capture structured, high-dimensional behaviors directly from demonstrations makes them well suited for DOM, where dynamics are complex and difficult to model explicitly. However, many real-world DOM applications (such as folding clothes, dressing patients, or household routines) are shaped by user-specific preferences, but remain understudied [5]. Addressing this requires sample-efficient alignment of pretrained policies to diverse demonstrations, which motivates the present work.

B. Preference Alignment in Robotics

Adapting robot behavior to user-specific needs is often studied through preference learning [24], [25], where predictive models of preferences are inferred from human feedback, typically in the form of pairwise comparisons between task

executions (e.g., a user indicating which of two executions they prefer) [26]. A common approach is reinforcement learning from human feedback (RLHF) [27], [28], which first trains a reward model on preference data and then fine-tunes a policy to maximize this learned reward [29], [2], [26]. RLHF has been pivotal in aligning large language models with human intent, but its dependence on reward modeling and RL optimization limits scalability in robotics [30], [31].

To overcome the limitations of RLHF, direct preference optimization (DPO) [10] and related methods bypass reward modeling by directly contrasting preferred and dispreferred behaviors, though they still require paired feedback. Relative preference optimization (RPO) [29] extends this idea by exploiting similarities across demonstrations, weighting all win-lose pairs within a batch to improve alignment. Kahneman-Tversky Optimization (KTO) [12] further reduces data requirements by using per-sample binary labels instead of pairwise comparisons. Diffusion-based preference optimization has shown strong results in text-to-image alignment [32], [11], [12], demonstrating scalability to high-dimensional generative models. However, its role in robotic manipulation (especially DOM) remains limited. Prior work has largely addressed rigid-object tasks such as rearrangement with language-based preferences [33], or narrow DOM scenarios like assisted dressing for improved comfort and safety [34], [1], but without leveraging the capabilities of diffusion models. Other studies have examined preference learning in handovers or coercive settings [35], [36], but these remain sparse and often limited to simple tasks. More recently, diffusion policies combined with DPO have been used to improve data efficiency in long-horizon planning [37] and even in DOM [38], though not for preference alignment. Bridging this gap by systematically comparing DPO, RPO, and KTO with diffusion-based visuomotor policies for DOM is one of the central focuses of this work.

III. METHOD

We begin with a brief overview of diffusion models and the preference alignment frameworks used in this work, followed by our proposed optimization method (RKO) and details on system design.

A. Preliminaries: Visuomotor diffusion models

The DDPM formulation defines a forward diffusion process (noise addition) and a backward denoising process. Starting from a noise sample x_K drawn from a Gaussian, the model iteratively denoises through K steps toward $x_0 = \epsilon_\theta(x_k, k)$ (a data sample), using a learned noise-prediction (score) network ϵ_θ . For visuomotor control, the formulation is adapted in [18] with two key modifications: (i) the output x becomes an action sequence, so diffusion operates in action space A_t ; (ii) the score network is conditioned on observations O_t (e.g., recent visual inputs), modeling the conditional distribution $p(A_t | O_t)$. Observation features are extracted once per time step and reused across denoising steps for efficiency and stability.

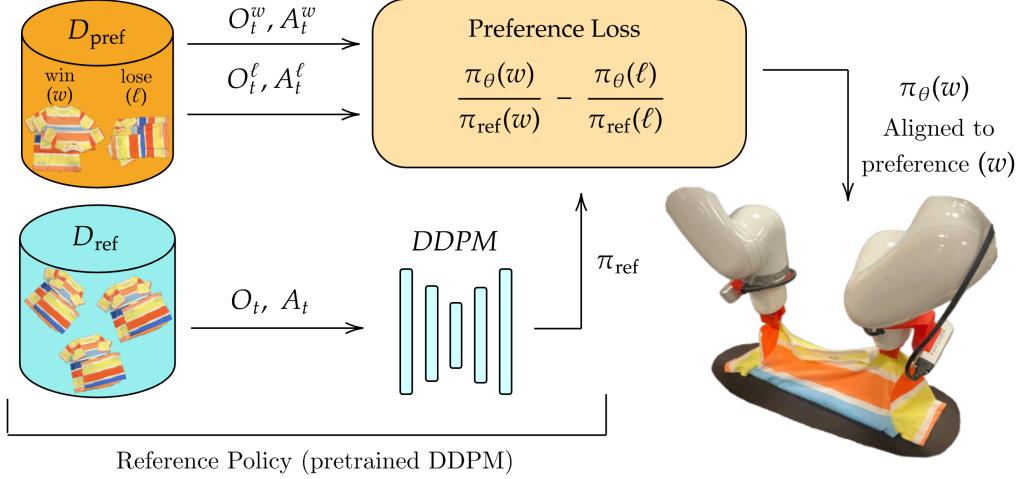


Fig. 2: General preference alignment framework used in this work. A reference model is first trained on a large set of demonstrations (D_{ref}) for a given task. To align it to a user’s preferred strategy, a new set of *winning* demonstrations is collected and combined into D_{pref} along with *losing* demonstrations, i.e., examples of alternative, dispreferred behaviors (which may also come from D_{ref}). The preference loss then aligns the new policy π_θ to the preferred behavior by explicitly contrasting it with the *losing* demonstrations. This enables more effective and sample-efficient alignment than training a diffusion policy solely on the *winning* demonstrations.

The denoising step is:

$$A_t^{k-1} = \alpha_k \left(A_t^k - \gamma_k \varepsilon_\theta(O_t, A_t^k, k) \right) + \mathcal{N}(0, \sigma_k^2 I), \quad (1)$$

with $\alpha_k, \gamma_k, \sigma_k$ forming the noise schedule. Training samples real actions A_0 , adds noise at a random step k , and trains ε_θ to predict ε_k via an MSE loss, corresponding to minimizing a variational lower bound:

$$\mathcal{L}_{\text{DDPM}}(\theta) = \text{MSE}(\varepsilon_k, \varepsilon_\theta(O_t, A_t^0 + \varepsilon_k, k)) \quad (2)$$

B. Preference Alignment frameworks

In this section, we detail how a pretrained diffusion model can be aligned to a new preferred behavior, expressed through novel preferred demonstrations, using diffusion-based preference alignment methods (DPO, RPO, KTO, RKO (ours)), highlighting their key similarities and differences. We first outline the common problem setting and how these frameworks are adapted to visuomotor policies.

Problem setting: for each preference optimization framework, we consider a dataset $D_{\text{pref}} = ((O^w, A^w), (O^\ell, A^\ell))$, where each example consists of preferred (*win*) demonstrations (O^w, A^w) and dispreferred (*lose*) ones (O^ℓ, A^ℓ) , with labels $w = 1, \ell = -1$. Alongside this, we use a pretrained reference diffusion model π_{ref} , trained on a larger dataset D_{ref} of preference-free demonstrations. These reference demonstrations represent a default or neutral strategy for performing the task (e.g., folding the same garment in a different but acceptable way). The reference model serves both as a reference for evaluating relative output quality and as initialization for the preference-aligned policy π_θ , which is then fine-tuned using the preference losses of each framework. In contrast, the lose demonstrations in D_{pref} are not neutral but explicitly represent a behavior the user does

not want (e.g., an undesired folding style) and are included to teach the model to avoid that behavior while learning from the *winning* demonstrations. The goal is to align π_θ with the *win* behaviour while using the same *winning* demonstrations across frameworks, enabling a fair comparison of sample efficiency. The preference loss amplifies the distinction between *winning* and *losing* behaviors, guiding the policy towards the preferred strategy while discouraging the dispreferred one. For clarity, we omit the denoising step k and denote ℓ for *lose* and w for *win*. The overall pipeline is illustrated in Fig. 2. Preference-free datasets D_{ref} are first used to train reference diffusion policies conditioned on observation-action sequences (O_t, A_t) . Each reference policy π_{ref} is then used in combination with a novel preference dataset D_{pref} for preference alignment. For evaluation, we compare the policies trained with our proposed framework (π_{RKO}) against the main classes of preference-optimization frameworks ($\pi_{\text{DPO}}, \pi_{\text{RPO}}, \pi_{\text{KTO}}$) and a vanilla DDPM trained on preference demonstrations (π_{DDPM}).

1) *Diffusion-DPO*: Unlike RLHF, Diffusion-DPO [32] directly aligns a diffusion model with human preferences, avoiding both reward modeling and reinforcement learning. The reference model serves as a baseline policy, anchoring what is considered “default” behavior and enabling relative comparisons between *win* and *lose* samples.

The objective is defined as the following pairwise loss:

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-DPO}}(\theta) &= \mathbb{E}_{(A_0^w, A_0^\ell) \sim D_{\text{pref}}, t \sim \mathcal{U}(0, T), \varepsilon \sim \mathcal{N}(0, I)} \\ &\quad \beta \log \sigma \left[(\|\varepsilon - \varepsilon_\theta(O_t^w, A_t^w, t)\|_2^2 - \|\varepsilon - \varepsilon_{\text{ref}}(O_t^w, A_t^w, t)\|_2^2) \right. \\ &\quad \left. - (\|\varepsilon - \varepsilon_\theta(O_t^\ell, A_t^\ell, t)\|_2^2 - \|\varepsilon - \varepsilon_{\text{ref}}(O_t^\ell, A_t^\ell, t)\|_2^2) \right] \end{aligned} \quad (3)$$

where $A_t^w \sim q(A_t^w | A_0^w)$ and $A_t^\ell \sim q(A_t^\ell | A_0^\ell)$ denote noisy samples at step t obtained from the forward diffusion process,

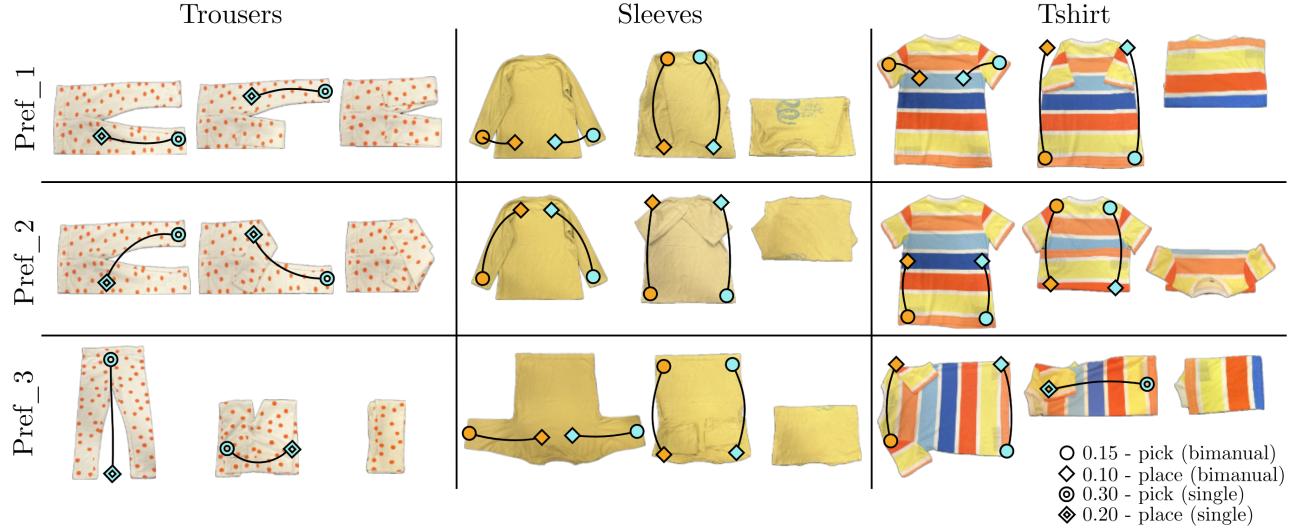


Fig. 3: Illustration of three garment-folding preferences for each garment type. Each panel shows the pick (circle) and place (diamond) positions for the left (orange) and right (light blue) arms. The scores are visible on the bottom right: bimanual actions are executed synchronously, and their scores are normalized so that each individual action contributes equally. The total score for a complete, correct folding sequence is 1.

$\varepsilon \sim \mathcal{N}(0, I)$ is sampled noise, and β is a scaling factor.

This loss measures the relative advantage of the current noise prediction network ε_θ over the reference one ε_{ref} when comparing preferred and dispreferred demonstrations.

2) *Diffusion-RPO*: While Diffusion-DPO relies on single win–lose pairs, it does not exploit the similarity and structure shared across demonstrations. Diffusion-RPO [29] extends this formulation with a context-aware reweighting scheme: each winning sample is compared against all losers in the batch, with importance weights assigned based on semantic similarity. The loss is defined over all pairs in the batch:

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-RPO}}(\theta) &= \mathbb{E}_{(A_0^w, A_0^\ell) \sim D_{\text{pref}}, t \sim \mathcal{U}(0, T), \varepsilon \sim \mathcal{N}(0, I)} \\ \omega_{i,j} \cdot \beta \log \sigma &\left[(\|\varepsilon - \varepsilon_\theta(O_{t,i}^w, A_{t,i}^w, t)\|_2^2 - \|\varepsilon - \varepsilon_{\text{ref}}(O_{t,i}^w, A_{t,i}^w, t)\|_2^2) \right. \\ &\quad \left. - (\|\varepsilon - \varepsilon_\theta(O_{t,j}^\ell, A_{t,j}^\ell, t)\|_2^2 - \|\varepsilon - \varepsilon_{\text{ref}}(O_{t,j}^\ell, A_{t,j}^\ell, t)\|_2^2) \right] \quad (4) \end{aligned}$$

where i indexes winning samples and j indexes losing samples within the batch. The key addition in RPO is the contextual reweighting factor $\omega_{i,j}$, which measures the similarity between each win–lose pair, where each winner is contrasted against all losers in the batch:

$$\omega_{i,j} = \frac{\exp\left(-\frac{1-\cos(f(o_i^w), f(o_j^\ell))}{\tau}\right)}{\sum_{j'} \exp\left(-\frac{1-\cos(f(o_i^w), f(o_{j'}^\ell))}{\tau}\right)} \quad (5)$$

where $f(o^*)$ is an encoder extracting image embeddings and τ a temperature hyperparameter. This soft alignment focuses the model on semantically similar pairs, ensuring each winner distributes its weight across all losers while giving greater emphasis to losers that are close in semantic space [29].

3) *Diffusion-KTO*: Diffusion-KTO [12], inspired by Kahneman-Tversky Optimization (KTO), aligns diffusion models using only per-sample binary (*win* or *lose*) feedback, rather than paired comparisons. Unlike DPO and RPO, which require preference pairs, KTO incorporates the relative value of each sample with respect to the reference policy, enabling training even with batches containing only winners or losers.

Each sample in D_{pref} is labeled as winner ($q = 1$) or loser ($q = -1$) and at each sampling step $t \sim \mathcal{U}(0, T)$ the deviation of the current policy from the reference policy is computed through:

$$\begin{aligned} \mathcal{L}_{\text{Diffusion-KTO}}(\theta) &= -\mathbb{E}_{A_0 \sim \mathcal{D}_{\text{pref}}, t \sim \mathcal{U}(0, T), \varepsilon \sim \mathcal{N}(0, I)} \\ \sigma(\beta \cdot q \cdot [\|\varepsilon - \varepsilon_\theta(O_t, A_t, t)\|_2^2 - \|\varepsilon - \varepsilon_{\text{ref}}(O_t, A_t, t)\|_2^2] - Q_{\text{ref}}) \end{aligned} \quad (6)$$

where ε and ε_{ref} are as in DPO, $\sigma(\cdot)$ is a sigmoid utility function and β controls the scale of the reward signal as in DPO and RPO. The sigmoid utility function acts as a smooth preference indicator for each sample, while Q_{ref} stabilizes training by centering the reward and is constrained to be non-negative as it approximates a KL-divergence term (see Eq. 7 in [12]).

C. Diffusion-RKO

We propose Diffusion-RKO, which unifies the benefits of Diffusion-KTO and Diffusion-RPO: like KTO, Diffusion-RKO uses binary preference labels $q \in \{+1, -1\}$ and does not require explicit win–lose pairs; at the same time, it integrates RPO’s similarity-based batch reweighting to emphasize hard negatives and ambiguous winners. Given a mini-batch of B preference-labeled samples $\{(O_0^b, A_0^b, q)\}_{b=1}^B$, ε_θ and ε_{ref} are evaluated on noised inputs (O_t^b, A_t^b) , with timestep $t \sim \mathcal{U}(0, T)$ and noise $\varepsilon \sim \mathcal{N}(0, I)$. The per-sample

reward advantage is:

$$A_b = \beta \left[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(O_t^b, A_t^b, t)\|_2^2 - \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_{\text{ref}}(O_t^b, A_t^b, t)\|_2^2 \right] - Q_{\text{ref}}$$

The utility of each sample is computed using a sigmoid function:

$$U_b = \sigma(q_b \cdot A_b)$$

To amplify the training signal around semantically ambiguous regions, RKO applies a contextual weight s_b to each sample, based on similarity between winners and losers in the batch. Following RPO, a similarity matrix $\omega_{i,j}$ is computed between winner and loser embeddings following Equation 5.

These weights are then aggregated into per-sample scalars:

$$\begin{aligned} s_i^{\text{pos}} &= 1 + \max_j \omega_{i,j} && \text{(for each winner } i) \\ s_j^{\text{neg}} &= \sum_i \omega_{i,j} && \text{(for each loser } j) \end{aligned}$$

After normalization to mean 1, we define s_b for each sample b depending on whether it is a winner or a loser.

The weighted KTO loss is then given by:

$$\mathcal{L}_{\text{Diffusion-RKO}}(\theta) = -\frac{1}{\sum_{b=1}^B s_b} \sum_{b=1}^B s_b \cdot \sigma(q_b \cdot A_b)$$

This formulation allows Diffusion-RKO to focus learning on difficult samples where winners resemble losers and viceversa.

a) Similarity reweighting and convergence intuition:

The similarity matrix $\omega_{i,j}$ defines a kernel on the representation space induced by the encoder embeddings (Eq. 5). Under this view, s_i^{pos} increases when a preferred sample i has a nearby dispreferred neighbor (large $\max_j \omega_{i,j}$), i.e., when the sample lies close to the local decision boundary between preferred and dispreferred behaviors. Conversely, s_j^{neg} increases with the local density of preferred neighbors around a dispreferred sample (large $\sum_i \omega_{i,j}$), highlighting *hard negatives* that are difficult to separate in representation space. This yields a principled mechanism: Diffusion-RKO allocates larger gradient mass to regions where preferred and dispreferred behaviors overlap, while leaving isolated, high-margin samples close to the unweighted KTO regime. From an optimization perspective, recent work proves contraction-style stability/convergence guarantees under a softmax bandit abstraction for DPO [39]; since KTO can be cast in the same KL/log-ratio form and RKO applies bounded similarity reweighting to the same per-sample terms, we expect analogous theoretical convergence guarantees to hold for KTO and RKO up to constant rescaling. We refer the reader to our website for the full convergence analysis¹.

D. Dataset Generation

For each preference, a policy is trained on 60 winning and 60 losing demonstrations, named D_{pref} , with 40 demos shared across all policies and 20 collected via human takeover [40]. These 20 demonstrations are collected by running the corresponding policy individually, after it has

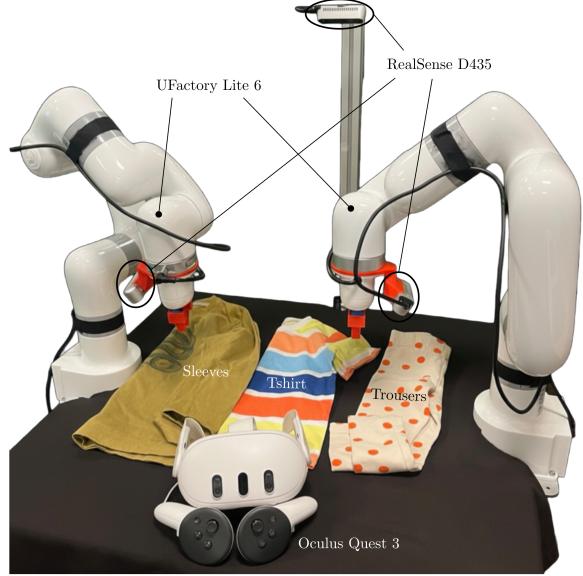


Fig. 4: Experimental setup.

been initially trained on 40 demonstrations. This yields 60 winning demonstrations per preference dataset in total, with takeover demonstrations comprising 50% of the data. Takeover is particularly advantageous as it allows to efficiently collect targeted demonstrations in out-of-distribution states that could be missing from the initial dataset. Reference policies are trained on a dataset of 100 demonstrations (70 standard, 30 takeover), which we name D_{ref} . Each demonstration consists of image observations, actions, and robot state, with an average length of 24s for trousers, 29s for sleeves and 25s for tshirt. We construct D_{pref} and D_{ref} by rotating the "roles" of three preference datasets in a fixed cycle. As shown in Fig. 3, there are three preference sets: pref_1, pref_2, and pref_3. For each preference setting, one set provides *winning* demonstrations, a different set provides *losing* demonstrations, and the remaining set is used to train the reference model (using 100 demonstrations). The assignment is as follows:

- For pref_1: winners come from pref_1, losers come from pref_2, and the reference model is trained on pref_3.
- For pref_2: winners come from pref_2, losers come from pref_3, and the reference model is trained on pref_1.
- For pref_3: winners come from pref_3, losers come from pref_1, and the reference model is trained on pref_2.

This cyclic setup ensures that, for each preference, the *winning* demonstrations match the target preference, while the *losing* demonstrations and the reference model are drawn from distinct, non-overlapping preference sources.

IV. EXPERIMENTS

We now describe how we evaluated different preference alignment frameworks. The main goal is to compare the proposed Diffusion-RKO with Diffusion-DPO, -RPO, -KTO, and a vanilla DDPM baseline in terms of sample efficiency. To ensure fairness, all policies are trained on the same set

¹github.com/Preference-DOM

of *winning* demonstrations and evaluated on their ability to reproduce the preferred behavior. We focus on real-world cloth folding, a domain where user preferences strongly influence task execution but are often overlooked. All experiments are conducted on real robots, considering that physically accurate simulation of deformable objects remains challenging [41] and large-scale demonstration datasets for training diffusion policies on cloth manipulation tasks are scarce both in simulation and the real world. Since our goal is to evaluate the models in terms of sample efficiency, we consider two experimental settings:

- **Performance comparison:** we compare all models on a fixed set of 60 *winning* demonstrations for challenging cloth-folding tasks. We chose 60 as we observed a good trade-off between policy performance and demonstration collection effort.
- **Sample efficiency vs. # demonstrations:** we compare all models while varying the number of training demonstrations (20 to 95 *winning* demonstrations, in increments of 15), tracking performance to analyze how each framework scales with the availability of demonstrations. In addition, we perform an ablation experiment for RKO where we set the similarity-reweighting factor to 1, in order to empirically evaluate its influence.

A. Experimental settings and implementation details

Cloth folding is notoriously difficult to evaluate [42]. To address this, we adopt a step-wise evaluation scheme based on the correctness of individual pick-and-place actions during execution, providing a more informative signal than binary success/failure. Scoring details for each garment are shown in Fig. 3, with task scores normalized to sum to 1. A pick action is only counted if it includes a successful lift, making it slightly more valuable than a place action, since failed lifts invalidate the attempt. The aim is not perfect folding, but alignment to the preferred behavior under sample-efficiency constraints. Due to the proximity of many joint singularities, the model must be precise to avoid failure. Executions stop when: the task is completed; the arms enter a singularity or collide with the sponge-covered table; or the system reaches an out-of-distribution state in robot poses or garment configuration from which recovery is impossible. Objects are placed with slight variations across trials, and additional variability arises from garment deformability. Recovery is allowed without penalty, as some demonstrations could include it, and we observe some generalization, with policies recovering autonomously from minor failures.

We use two UFactory Lite6 robots equipped with three RealSense D435 cameras (Fig. 4): two wrist-mounted and one bird’s-eye view. For trousers, only the bird’s-eye and right wrist cameras are used; for sleeves and t-shirts, all three are employed. Demonstrations are collected via teleoperation using the Quest2ROS Oculus app [43]. Our experiments cover three garment types (trousers, tshirt, and sleeves), and three task variations with different levels of task complexity. Trousers involve single-arm manipulation with two-camera input, while sleeves and tshirt require bimanual coordination

and synchronized actions. Each policy is evaluated over 10 runs per task, for a total of 1170 executions. All policies are trained using RGB image observations. The model architecture consists of a CNN encoder and a U-Net-based denoising backbone. DDPM baselines are initialized from a model pretrained on a separate folding task for fairness, though we found no significant performance boost from this. All models are trained for 100k training steps. For preference models, we select checkpoints based on the best reward difference between *winning* and *losing* demonstrations. All models use a batch size of 64 and a learning rate of $3 \cdot 10^{-5}$. Following [32], we incorporate the time-dependent factor T into the scaling constant β . We use the following hyperparameters, selected based on their stable and consistent performance during preliminary evaluations:

- $\beta = 10$ for DPO,
- $\beta = 20$ for RPO,
- $\beta = 12$ for KTO and RKO,
- $\tau = 0.15$ for RPO and RKO.

B. Results

We report results from two evaluation settings: (1) performance comparison across models using a fixed number of demonstrations, and (2) sample efficiency analysis by varying the number of demonstrations used for training. Videos of the executions and further details are available on our website².

Performance comparison: Table I, II and III summarize the results for the Trousers, Sleeves and Tshirt performance comparison experiments.

Model	Pref 1	Pref 2	Pref 3
π_{DDPM}	0.701 ± 0.377	0.667 ± 0.381	0.453 ± 0.416
π_{DPO}	0.680 ± 0.336	0.780 ± 0.225	0.590 ± 0.348
π_{RPO}	0.790 ± 0.238	0.790 ± 0.233	0.570 ± 0.271
π_{KTO}	0.860 ± 0.206	0.650 ± 0.381	0.520 ± 0.322
$\pi_{\text{RKO}} (\text{ours})$	0.910 ± 0.166	0.760 ± 0.334	0.620 ± 0.367

TABLE I: Trousers - Performance Comparison

Model	Pref 1	Pref 2	Pref 3
π_{DDPM}	0.510 ± 0.301	0.563 ± 0.306	0.626 ± 0.279
π_{DPO}	0.570 ± 0.275	0.350 ± 0.163	0.460 ± 0.279
π_{RPO}	0.455 ± 0.314	0.685 ± 0.238	0.690 ± 0.156
π_{KTO}	0.730 ± 0.249	0.660 ± 0.297	0.730 ± 0.314
$\pi_{\text{RKO}} (\text{ours})$	0.715 ± 0.274	0.695 ± 0.207	0.695 ± 0.199

TABLE II: Sleeves - Performance Comparison

The results show that preference optimization frameworks consistently outperform the vanilla DDPM baseline across all tasks. Among individual models, π_{RKO} achieves the best performance in 4 out of 9 tasks, followed by π_{KTO}

²github.com/Preference-DOM

Model	Pref 1	Pref 2	Pref 3
π_{DDPM}	0.496 ± 0.333	0.453 ± 0.436	0.593 ± 0.405
π_{DPO}	0.515 ± 0.368	0.520 ± 0.307	0.520 ± 0.391
π_{RPO}	0.525 ± 0.192	0.560 ± 0.350	0.440 ± 0.338
π_{KTO}	0.595 ± 0.274	0.540 ± 0.389	0.695 ± 0.219
$\pi_{\text{RKO}} (\text{ours})$	0.605 ± 0.306	0.550 ± 0.251	0.660 ± 0.302

TABLE III: Tshirt - Performance Comparison

in 3 and π_{RPO} in 2, suggesting that these methods are generally more effective than DPO for preference alignment and sample efficiency. This highlights the value of using more expressive preference frameworks even in scenarios not directly related to preference alignment but that still use preference optimization frameworks, as in [38]. Three of four preference-based methods consistently outperform the DDPM baseline, and RKO does so across all tasks. They are also more training-efficient, typically reaching strong performance within 40–50k steps versus 100k for DDPM. Overall, these results indicate that preference learning (especially RKO, KTO, and RPO) provides substantially better sample and training efficiency than standard DDPM or DPO for aligning diffusion policies to user-preferred behaviors.

Sample efficiency vs. # demos: results are shown in Fig. 5: RKO consistently outperforms DDPM in terms of sample efficiency across all settings and typically matches or outperforms the other baselines. The main exceptions are Sleeves Task 1, where KTO performs better at 35 and 65 demonstrations, and Trousers Task 1 with 20 demonstrations, where RKO performs worst. In the latter low-data regime, we frequently observe the robot going out-of-distribution during executions, likely due to the limited number of *winning* and *losing* demonstrations which may confuse RKO during training. This effect vanishes as the dataset grows, after which RKO again surpasses DDPM, indicating that its advantage is robust when sufficient preference data are available. For the ablated variant without reweighting (RKO-noRW), performance is consistently lower than full RKO across the experiments, indicating that the reweighting term contributes in boosting RKO. Given the relatively large standard deviations observed in the experiments, we further assessed the robustness of the results using a Bayesian A/B test, following the approach discussed in [44], defining a task as successful when the achieved score is at least 0.75. The posterior analysis shows that RKO outperforms DDPM with probabilities of 87% on Trousers and 70.1% on Sleeves, consistent across varying numbers of demonstrations, providing statistical evidence of RKO’s reliability over DDPM.

Overall, results from both experiments indicate that using a preference alignment framework is generally more effective than training a vanilla diffusion model from scratch, especially when demonstrations of alternative behaviors and pre-trained reference policies are available. Preference-aligned policies not only learn the desired behavior more accurately, but also benefit from being explicitly constrained to avoid

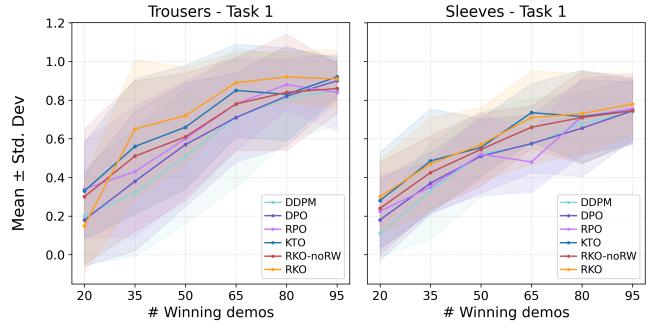


Fig. 5: Sample efficiency experiment: influence of the number of demonstrations (20 to 95 *winning* demos, increments of 15) on the performance, in Trousers - Pref 1 and Sleeves - Pref 1.

dispreferred behaviors. This leads to improved performance and generally faster training. These advantages are particularly evident in scenarios where additional demonstrations (even if not representing the preferred behavior) can be repurposed as *losing* examples instead of being discarded. In terms of failure modes, we observe no method-specific failure patterns across the different preference optimization frameworks. Most failures arise when the robot enters states that are out of distribution relative to the training data, often due to imprecise end-effector orientation or incorrect initial actions that lead to unseen states. Failures are more frequent in bimanual tasks, where coordination between arms increases the likelihood of distributional shift, and are more pronounced for DDPM- and DPO-based policies.

V. CONCLUSION

In this work, we studied alignment of pretrained visuomotor diffusion policies to user-preferred behaviors for deformable object manipulation, and provided the first systematic comparison of diffusion preference objectives (DPO, RPO, KTO) while introducing RKO, which combines the benefits of KTO and RPO. Across multiple real-world cloth-folding tasks, preference alignment consistently improves performance and sample efficiency over standard DDPM fine-tuning, suggesting a practical path to incorporating user preferences using limited demonstrations instead of retraining from scratch. Future work will study robustness to noisy or suboptimal demonstrations, automatic selection of negative samples from offline data, and extension beyond cloth folding to other types of deformable objects.

REFERENCES

- [1] G. Canal, G. Alenyà, and C. Torras, “Adapting robot task planning to user preferences: an assistive shoe dressing example,” *Autonomous Robots*, vol. 43, no. 6, pp. 1343–1356, 2019.
- [2] B. Woodworth, F. Ferrari, T. E. Zosa, and L. D. Riek, “Preference learning in assistive robotics: Observational repeated inverse reinforcement learning,” in *Machine learning for healthcare conference*, pp. 420–439, PMLR, 2018.
- [3] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, *et al.*, “Challenges and outlook in robotic manipulation of deformable objects,” *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

- [4] A. Longhini, Y. Wang, I. Garcia-Camacho, D. Blanco-Mulero, M. Moletta, M. Welle, G. Alenyà, H. Yin, Z. Erickson, D. Held, *et al.*, “Unfolding the literature: A review of robotic cloth manipulation,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 8, 2024.
- [5] G. Canal, “Adapting robot behavior to user preferences in assistive scenarios,” 2020.
- [6] M. Moletta, M. K. Wozniak, M. C. Welle, and D. Kragic, “A virtual reality framework for human-robot collaboration in cloth folding,” in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–7, IEEE, 2023.
- [7] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, “ π_0 : A vision-language-action flow model for general robot control,” 2024.
- [8] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [9] S. Liu, W. Fang, Z. Hu, J. Zhang, Y. Zhou, K. Zhang, R. Tu, T.-E. Lin, F. Huang, M. Song, *et al.*, “A survey of direct preference optimization,” *arXiv preprint arXiv:2503.11701*, 2025.
- [10] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in neural information processing systems*, vol. 36, pp. 53728–53741, 2023.
- [11] Y. Gu, Z. Wang, Y. Yin, Y. Xie, and M. Zhou, “Diffusion-rpo: Aligning diffusion models through relative preference optimization,” *arXiv preprint arXiv:2406.06382*, 2024.
- [12] S. Li, K. Kallidromitis, A. Gokul, Y. Kato, and K. Kozuka, “Aligning diffusion models by optimizing human utility,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 24897–24925, 2024.
- [13] R. Laezza, “Robot learning for deformable object manipulation tasks,” 2024.
- [14] A. Longhini, *Adapting to Variations in Textile Properties for Robotic Manipulation*, PhD thesis, KTH Royal Institute of Technology, 2025.
- [15] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, “Learning reward functions by integrating human demonstrations and preferences,” *arXiv preprint arXiv:1906.08928*, 2019.
- [16] N. Ingelhag, J. Munkeby, J. van Haastregt, A. Varava, M. C. Welle, and D. Kragic, “A robotic skill learning system built upon diffusion policies and foundation models,” in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 748–754, IEEE, 2024.
- [17] M. Dalal, A. Mandlekar, C. Garrett, A. Handa, R. Salakhutdinov, and D. Fox, “Imitating task and motion planning with visuomotor transformers,” *arXiv preprint arXiv:2305.16309*, 2023.
- [18] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [19] A. Sridhar, D. Shah, C. Glossop, and S. Levine, “Nomad: Goal masked diffusion policies for navigation and exploration,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 63–70, IEEE, 2024.
- [20] I. Kapelyukh, V. Vosylius, and E. Johns, “Dall-e-bot: Introducing web-scale diffusion models to robotics,” *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 3956–3963, 2023.
- [21] U. A. Mishra, S. Xue, Y. Chen, and D. Xu, “Generative skill chaining: Long-horizon skill planning with diffusion models,” in *Proceedings of The 7th Conference on Robot Learning* (J. Tan, M. Toussaint, and K. Darvish, eds.), vol. 229 of *Proceedings of Machine Learning Research*, pp. 2905–2925, PMLR, 06–09 Nov 2023.
- [22] M. Reuss, M. Li, X. Jia, and R. Lioutikov, “Goal-conditioned imitation learning using score-based diffusion policies,” *arXiv preprint arXiv:2304.02532*, 2023.
- [23] T. Pearce, T. Rashid, A. Kanervisto, D. Bignell, M. Sun, R. Georgescu, S. V. Macua, S. Z. Tan, I. Momennejad, K. Hofmann, *et al.*, “Imitating human behaviour with diffusion models,” *arXiv preprint arXiv:2301.10677*, 2023.
- [24] J. Fürnkranz and E. Hüllermeier, “Preference learning and ranking by pairwise comparison,” in *Preference learning*, pp. 65–82, Springer, 2010.
- [25] D. Sadigh, A. Dragan, S. Sastry, and S. Seshia, *Active preference-based learning of reward functions*. 2017.
- [26] C. Wirth, R. Akrou, G. Neumann, and J. Fürnkranz, “A survey of preference-based reinforcement learning methods,” *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.
- [27] W. Cheng, J. Fürnkranz, E. Hüllermeier, and S.-H. Park, “Preference-based policy iteration: Leveraging preference learning for reinforcement learning,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 312–327, Springer, 2011.
- [28] A. Wilson, A. Fern, and P. Tadepalli, “A bayesian approach for policy learning from trajectory preference queries,” *Advances in neural information processing systems*, vol. 25, 2012.
- [29] Y. Yin, Z. Wang, Y. Gu, H. Huang, W. Chen, and M. Zhou, “Relative preference optimization: Enhancing ILM alignment through contrasting responses across identical and diverse prompts,” *arXiv preprint arXiv:2402.10958*, 2024.
- [30] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
- [31] G. I. Winata, H. Zhao, A. Das, W. Tang, D. D. Yao, S.-X. Zhang, and S. Sahu, “Preference tuning with human feedback on language, speech, and vision tasks: A survey,” *Journal of Artificial Intelligence Research*, vol. 82, pp. 2595–2661, 2025.
- [32] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, “Diffusion model alignment using direct preference optimization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.
- [33] J. Lee, S. Park, Y. Kwon, J. Lee, M. Ahn, and S. Choi, “Visual preference inference: An image sequence-based preference reasoning in tabletop object manipulation,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9745–9752, IEEE, 2024.
- [34] F. Zhang, A. Cully, and Y. Demiris, “Personalized robot-assisted dressing using user modeling in latent spaces,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3603–3610, IEEE, 2017.
- [35] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler, “Human preferences for robot-human hand-over configurations,” in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1986–1993, IEEE, 2011.
- [36] A. Jain, S. Sharma, T. Joachims, and A. Saxena, “Learning preferences for manipulation tasks from online coercive feedback,” *The International Journal of Robotics Research*, vol. 34, no. 10, pp. 1296–1313, 2015.
- [37] X. Yuan, Z. Shang, Z. Wang, C. Wang, Z. Shan, M. Zhu, C. Bai, X. Li, W. Wan, and K. Harada, “Preference aligned diffusion planner for quadrupedal locomotion control,” *arXiv preprint arXiv:2410.13586*, 2024.
- [38] W. Chen, H. Xue, F. Zhou, Y. Fang, and C. Lu, “Deformpam: Data-efficient learning for long-horizon deformable object manipulation via preference-based action alignment,” *arXiv preprint arXiv:2410.11584*, 2024.
- [39] R. Shi, R. Zhou, and S. S. Du, “The crucial role of samplers in online direct preference optimization,” *arXiv preprint arXiv:2409.19605*, 2024.
- [40] N. Ingelhag, J. Munkeby, M. C. Welle, M. Moletta, and D. Kragic, “Real-time operator takeover for visuomotor diffusion policy training,” *arXiv preprint arXiv:2502.02308*, 2025.
- [41] H. Yin, A. Varava, and D. Kragic, “Modeling, learning, perception, and control methods for deformable object manipulation,” *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.
- [42] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borras, C. Torras, A. Marino, G. Alenyà, *et al.*, “Benchmarking bimanual cloth manipulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1111–1118, 2020.
- [43] M. C. Welle, N. Ingelhag, M. Lippi, M. Wozniak, A. Gasparri, and D. Kragic, “Quest2ros: An app to facilitate teleoperating robots,” in *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions*, 2024.
- [44] H. Kress-Gazit, K. Hashimoto, N. Kuppuswamy, P. Shah, P. Horgan, G. Richardson, S. Feng, and B. Burchfiel, “Robot learning as an empirical science: Best practices for policy evaluation,” *arXiv preprint arXiv:2409.09491*, 2024.