COUNTERFACTUAL TAYLOR ATTRIBUTION FOR SEQUENTIAL POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing reinforcement-learning approaches have shown strong performance in sequential active perception by optimizing acquisition policies from long-term utility. However, these methods typically model only the next acquisition conditioned on the current observation history, leaving alternative multi-step acquisition strategies implicit in the policy. We introduce a diffusion policy with Counterfactual Taylor Attribution (CTA) to explicitly exploit such future plan variation. The diffusion policy models a joint distribution over multi-step acquisitions, enabling diverse candidate plans to be sampled and counterfactually evaluated from the same decision state. CTA uses the resulting utility variation to estimate action-level contributions and redistribute within-plan credit during policy optimization, without introducing an additional reward model or changing the original task objective. Our framework provides a general mechanism for exploiting counterfactual plan variation in active information acquisition, including active perception and sequential Bayesian experimental design.

1 INTRODUCTION

Active perception requires an agent to select measurements that improve inference about an unknown task quantity under a limited acquisition budget. Because each acquisition changes the information available to subsequent decisions, its value depends on its downstream effect on future observations and prediction rather than only on its immediate outcome (Foster et al., 2021). Reinforcement-learning (RL) approaches, such as Active Perception Policy Learning (APPLE; Schneider et al., 2026), optimize these long-term effects through sequential returns, while related Bayesian experimental design methods, such as Amortized Active Learning and Inference Engine (Huang et al., 2026), optimize sequential information acquisition under an analogous decision process. However, their acquisition policies typically represent a distribution over the next acquisition rather than a joint distribution over future acquisition sequences, leaving alternative multi-step strategies and their downstream utility variation implicit in the policy update.

A natural way to expose these alternatives is to model future acquisitions jointly. Diffusion policies provide expressive distributions over multi-step action sequences (Chi et al., 2025) and have recently been applied to information-gathering problems including autonomous exploration, informative path planning, active tracking, and 3D scanning (Cao et al., 2025; Lew et al., 2026; Liu et al., 2025; Hirako et al., 2026). From the same decision state, such a policy can generate multiple candidate acquisition plans whose different downstream outcomes reveal how variations in future decisions affect utility.

Reinforcement-learning optimization of diffusion policies has also been studied through methods such as DIPO, DDiffPG, and DPPO (**??**Ren et al., 2025). In DPPO, however, the sampled action sequence is optimized using a scalar plan-level advantage, which determines whether the plan as a whole should be reinforced or suppressed. This is limiting for sequential information acquisition, where measurements within the same plan can contribute very differently to downstream inference: a high-utility plan may contain both highly informative and redundant acquisitions. Moreover, in sequential Bayesian experimental design, consecutive actions in a plan need not correspond to nearby or smoothly varying design locations; the policy may move between distant regions of the design space as information is gathered. Thus, acquisition position alone provides little indication of informational contribution, and applying the same plan-level signal across all positions cannot distinguish

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which measurements are actually responsible for the observed utility. Thus, although the joint diffusion policy exposes structured counterfactual variation across plans, standard plan-level policy optimization does not use this variation to resolve within-plan credit.

A key advantage of active information acquisition is that candidate plans can be evaluated directly in terms of their downstream inference quality. In active perception, a prediction model provides the task utility associated with the observations induced by each plan, while in Bayesian experimental design the same role can be played by information gain or posterior improvement. This makes it possible to compare multiple plans from the same decision state and obtain counterfactual variation in utility without introducing an additional learned reward model. We introduce Counterfactual Taylor Attribution (CTA) to convert this variation into acquisition-level credit. CTA estimates the local utility variation from neighboring sampled plans through weighted regression (Conn et al., 2009; Roberts, 2025; Ma et al., 2025) and uses a first-order Taylor attribution to identify how individual acquisitions contribute relative to counterfactual alternatives (Deng et al., 2021; 2024). Integrated with DPPO, the resulting CTA-DPPO redistributes the existing plan-level update according to these contributions while preserving the original task objective.

We evaluate CTA-DPPO on active-perception benchmarks from APPLE (Schneider et al., 2026) and a sequential Bayesian experimental design benchmark from ALINE (Huang et al., 2026). We compare against the original acquisition methods as well as a protocol-matched DPPO baseline, allowing us to assess both overall sequential acquisition performance and the benefit of counterfactual within-plan credit assignment.

Our contributions are threefold:

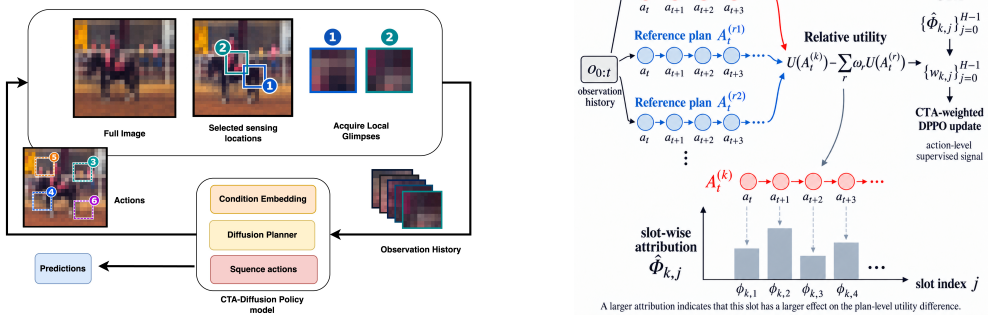
- We formulate multi-step information acquisition with a joint diffusion policy, allowing diverse future acquisition plans to be sampled from the same decision state and exposing counterfactual plan variation that remains implicit when policies optimize only the next acquisition.
- We introduce a counterfactual acquisition contribution that attributes utility differences between alternative plans to individual acquisitions. We show that these contributions exactly decompose the counterfactual utility margin—the improvement of a sampled plan over plausible alternatives—thereby connecting within-plan credit directly to the original task utility.
- We develop Counterfactual Taylor Attribution (CTA) to estimate these ideal contributions from sampled plan–utility pairs using a local first-order approximation, and use the estimated contributions to provide acquisition-level credit within the DPPO objective. Experiments on APPLE active-perception tasks and ALINE sequential Bayesian experimental design demonstrate the approach across distinct information-acquisition objectives.

2 Related Work

2.1 Active Perception and Sequential Information Acquisition

Active perception studies how sensing actions can be selected to improve task-relevant inference (Bajcsy, 1988). Learning-based methods have applied this idea to sequential visual attention (Mnih et al., 2014) and tactile exploration (Fleer et al., 2020; Smith et al., 2021; Xu et al., 2022; Hu et al., 2026). APPLE (Schneider et al., 2026) provides a general reinforcement-learning formulation for active perception by jointly learning an acquisition policy and a perception model under a common task objective. Related work in sequential Bayesian experimental design similarly optimizes measurements according to their downstream information value (Huan & Marzouk, 2016; Foster et al., 2021); ALINE (Huang et al., 2026) further couples amortized inference with active data acquisition.

These methods account for the long-term consequences of acquisition decisions, but their policies typically represent a distribution over the next acquisition rather than a joint distribution over future acquisition sequences. Alternative multi-step strategies therefore remain implicit in the policy update. We instead represent future acquisitions jointly as multi-step plans, allowing multiple plausible continuations to be sampled from the same decision state. This makes both the variation among alternative acquisition strategies and their downstream utility differences explicit, providing the counterfactual structure exploited by our method.



(a) Sequential acquisition with a diffusion policy.

(b) Counterfactual Taylor Attribution for diffusion-policy training.

Figure 1: Our method learns sequential acquisition policies by exploiting counterfactual variation among plans generated from the same decision state. Counterfactual Taylor Attribution (CTA) converts plan-level utility differences into action-level credit for diffusion-policy optimization. Shown in (a) is the sequential acquisition process, while (b) illustrates the CTA-based training procedure.

2.2 Diffusion Policies and Fine-Grained Policy Optimization

Diffusion policies provide expressive models for multimodal action distributions by generating multi-step action sequences through iterative denoising (Chi et al., 2025). Their use has subsequently been extended from behavior cloning to reinforcement learning, including offline Diffusion Q-learning (Wang et al., 2022), online model-free diffusion policies (**?**), Q-function-guided score matching (Psenka et al., 2023), and diffusion policy-gradient methods (**?**). DPPO (Ren et al., 2025) further enables policy-gradient fine-tuning of pretrained diffusion policies by treating the reverse denoising process as an auxiliary decision process and optimizing its transitions with proximal policy optimization (Schulman et al., 2017).

Beyond manipulation and continuous-control benchmarks, diffusion policies have also been applied to information-gathering and active-sensing problems, including autonomous exploration (Cao et al., 2025), informative path planning (Lew et al., 2026), active target tracking (Liu et al., 2025), and active 3D scanning (Hirako et al., 2026). These studies highlight the ability of diffusion policies to represent diverse multi-step sensing strategies. However, their learning objectives generally evaluate a sampled action sequence at the plan or trajectory level, without explicitly exploiting utility variation across alternative plans to resolve credit among the individual acquisitions within a plan.

This limitation is closely related to fine-grained credit assignment, where aggregate outcome feedback can obscure heterogeneous contributions from individual decisions (Deng et al., 2026; Li et al., 2026b;a). We address this problem in sequential acquisition by exploiting counterfactual plans sampled from the same decision state. Rather than introducing an additional reward model, Counterfactual Taylor Attribution estimates how individual acquisitions contribute to the utility differences among these plans and uses the resulting credits to reweight the existing DPPO update within each sampled plan.

3 Method

3.1 Problem Formulation

We consider active perception, where an agent actively acquires measurements to infer an unknown task property under a limited acquisition budget. Following APPLE (Schneider et al., 2026), we formulate the interaction as a partially observable sequential decision problem. At step t, the environment has hidden state $\widetilde{h}_t = (h_t, y_t^\star)$, where $y_t^\star$ is the target property to be inferred and h_t contains the remaining environment state. The agent receives observation o_t from the current environment state and conditions both its acquisition and prediction decisions on the observation history

$$\widetilde{a}_t = (a_t, y_t), \qquad a_t \sim \pi_a(\cdot \mid o_{0:t}), \qquad y_t = \pi_y(o_{0:t}), \tag{1}$$

where a_t determines the next acquisition and y_t is the current prediction of $y_t^\star$.

As in APPLE (Schneider et al., 2026), the active-perception reward is defined as

$$\widetilde{r}_t = r(h_t, a_t) - \ell(y_t^\star, y_t), \qquad J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t \widetilde{r}_t \right], \tag{2}$$

where $r(h_t, a_t)$ captures task-dependent acquisition reward, $\ell(y_t^\star, y_t)$ measures prediction error, T is the acquisition horizon, and γ is the discount factor. The same sequential formulation also covers Bayesian experimental design: $y_t^\star$ can represent an unknown model parameter, a_t the next experimental design, and the acquisition reward can be defined through information gain or posterior improvement, as in ALINE (Huang et al., 2026).

The value of an acquisition depends not only on its immediate outcome but also on the information obtained by subsequent measurements (Foster et al., 2021). Existing RL-based methods (Blau et al., 2022) account for these future consequences through long-term returns while parameterizing the policy one acquisition at a time, e.g. in APPLE (Schneider et al., 2026). This leaves alternative future acquisition sequences implicit in the policy update. We instead model a joint distribution over future acquisitions. From the same observation history, the policy can generate multiple candidate continuations whose different downstream outcomes provide structured supervision about how variations explored by the current policy affect utility. These multi-step plans serve as counterfactual training objects, with only a short prefix executed before replanning to preserve closed-loop adaptation.

3.2 Diffusion Policy for Sequential Acquisition

To explicitly represent alternative future acquisition strategies, we model the acquisition policy as a joint distribution over an H-step plan,

$$A_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}] \in \mathbb{R}^{H \times d_a}, \qquad A_t \sim \pi_\theta(\cdot \mid o_{0:t}), \tag{3}$$

where d_a is the dimension of each acquisition action. Following existing diffusion-policy approaches (Chi et al., 2025; Lew et al., 2026; Hirako et al., 2026), we parameterize this joint distribution with a conditional diffusion model that generates the complete action sequence through iterative denoising. Let A_t^n denote the noisy plan at diffusion step n, with $n = N, \ldots, 0$. Starting from Gaussian noise $A_t^N \sim \mathcal{N}(0, I)$, the reverse process progressively refines the full acquisition sequence,

$$A_t^{n-1} \sim p_\theta \left(A_t^{n-1} \mid A_t^n, o_{0:t}, n \right), \qquad n = N, \ldots, 1, \qquad A_t = A_t^0. \tag{4}$$

Each denoising transition operates jointly on all H future acquisitions, allowing the policy to model dependencies across the planned sequence.

We initialize the diffusion policy by behavior cloning expert acquisition trajectories with the standard diffusion objective (Lew et al., 2026). During deployment, the first $E \leq H$ acquisitions are executed before the new observations are incorporated and a new H-step plan is generated. This receding-horizon scheme combines joint multi-step planning with closed-loop adaptation.

We subsequently fine-tune the diffusion policy using DPPO (Ren et al., 2025), which treats the reverse denoising process as an auxiliary decision process and applies the PPO objective to the sampled denoising transitions. Given a plan-level advantage $\mathcal{A}_t$, the policy is updated by increasing or decreasing the likelihood of the denoising trajectory that generated A_t. The resulting update therefore uses a single scalar signal to evaluate the complete acquisition plan.

The joint policy exposes structured variation among alternative continuations, but Standard DPPO assigns a single plan-level advantage to all acquisitions within a sampled plan. While this determines whether the plan as a whole should be reinforced or suppressed, it does not distinguish acquisitions that contribute differently to the resulting utility. Exploiting this additional supervision therefore requires a within-plan credit signal. Prior work has similarly shown that aggregate trajectory feedback can obscure heterogeneous decision-level contributions (Deng et al., 2026; Li et al., 2026b;a). We therefore first characterize an ideal acquisition-level contribution, and defer its practical estimation to Section 3.3.

For a sampled plan $A_t^{(k)}$, let $A_t^{(r)}$ denote a plausible counterfactual plan generated from the same decision context. For each pair $(A_t^{(k)}, A_t^{(r)})$, consider any continuously differentiable path

$$\Xi_{kr} : [0, 1] \to \mathbb{R}^{H \times d_a}, \qquad \Xi_{kr}(0) = A_t^{(r)}, \qquad \Xi_{kr}(1) = A_t^{(k)}, \tag{5}$$

where $\xi_{kr,j}(\lambda) \in \mathbb{R}^{d_a}$ denotes the j-th acquisition component of $\Xi_{kr}(\lambda)$.

Definition 1 (Counterfactual acquisition contribution)**.**

$$z^{\star}_{k,j} = \mathbb{E}_{A_t^{(r)} \sim q_t} \left[\int_0^1 \nabla_{a_{t+j}} U_t(\Xi_{kr}(\lambda))^{\top} \dot{\xi}_{kr,j}(\lambda) \, \mathrm{d}\lambda \right], \tag{6}$$

where $\dot{\xi}_{kr,j}(\lambda) = \mathrm{d}\xi_{kr,j}(\lambda)/\mathrm{d}\lambda$, $U_t(A)$ denotes the expected downstream utility of plan A from the current decision context, and q_t denotes a distribution over plausible counterfactual plans.

The definition attributes the utility change between two plans to the acquisition variations along a counterfactual path. For any continuously differentiable path connecting the same endpoint plans, the total contribution recovers their utility difference.

To connect this decomposition directly to the task objective, we measure how much better a sampled plan performs than plausible alternatives from the same decision context. We call this gap the counterfactual utility margin,

$$M_t\Big(A_t^{(k)}\Big) = U_t\Big(A_t^{(k)}\Big) - \mathbb{E}_{A_t^{(r)} \sim q_t} \Big[U_t\Big(A_t^{(r)}\Big) \Big]. \tag{7}$$

A positive margin indicates that the sampled plan achieves higher expected downstream utility than its plausible counterfactual alternatives, while a negative margin indicates that it performs worse. For a fixed decision context and counterfactual distribution q_t, the second term is a common reference baseline, so maximizing the counterfactual utility margin is equivalent to maximizing the original downstream utility.

The acquisition-level contributions exactly decompose this margin. By the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} U_t(\Xi_{kr}(\lambda)) = \sum_{j=0}^{H-1} \nabla_{a_{t+j}} U_t(\Xi_{kr}(\lambda))^{\top} \dot{\xi}_{kr,j}(\lambda). \tag{8}$$

Integrating along the path and applying the fundamental theorem of calculus gives

$$\sum_{j=0}^{H-1} z^{\star}_{k,j} = \mathbb{E}_{A_t^{(r)} \sim q_t} \Big[U_t\Big(A_t^{(k)}\Big) - U_t\Big(A_t^{(r)}\Big) \Big] = M_t\Big(A_t^{(k)}\Big).$$

Thus, $z^{\star}_{k,j}$ is an acquisition-level decomposition of the same task-relevant utility margin that distinguishes a sampled plan from plausible alternatives. A larger contribution indicates that the variation of the corresponding acquisition accounts for a larger portion of this utility difference, while a negative contribution indicates that it opposes the overall improvement.

This property also clarifies what is hidden by uniform plan-level credit. DPPO observes only an aggregate plan-level advantage, although the associated utility margin may arise from highly heterogeneous acquisition contributions. We therefore retain the plan-level advantage to determine the overall reinforcement or suppression direction, while using acquisition-level contributions to determine how the update strength should be distributed within the plan.

Exact evaluation of $z^{\star}_{k,j}$ requires utility sensitivities along counterfactual paths together with an expectation over plausible alternatives. These quantities are unavailable in our black-box setting. CTA approximates this ideal contribution directly from sampled plan–utility pairs.

3.3 Counterfactual Taylor Attribution

Given an observation history $o_{0:t}$, we sample K candidate plans from the behavior policy,

$$A_t^{(k)} = \Big[a_t^{(k)}, a_{t+1}^{(k)}, \ldots, a_{t+H-1}^{(k)} \Big] \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid o_{0:t}), \qquad k = 1, \ldots, K. \tag{9}$$

All plans are evaluated from the same decision state. Let $J_t^{(k)}$ denote the realized downstream utility of plan $A_t^{(k)}$. We standardize utility within the same-state population, $\widetilde{\mathcal{U}}_t^{(k)} = \frac{J_t^{(k)} - \overline{J}_t}{\sigma_t}$, $\overline{J}_t = \frac{1}{K} \sum_{i=1}^{K} J_t^{(i)}$, where σ_t is the standard deviation of the K utilities.

For each target plan $A_t^{(k)}$, we construct a local counterfactual neighborhood from its M nearest same-state plans,

$$\mathcal{N}_k = \mathrm{TopM}_{i \neq k} \left(- \left\| \mathrm{vec}(A_t^{(i)}) - \mathrm{vec}(A_t^{(k)}) \right\|_2^2 \right). \tag{10}$$

We then fit a locally weighted linear model to the neighboring plan–utility pairs,

$$\widetilde{\mathcal{U}}_t^{(i)} - \widetilde{\mathcal{U}}_t^{(k)} \approx g_k^\top \left(\mathrm{vec}(A_t^{(i)}) - \mathrm{vec}(A_t^{(k)}) \right), \qquad i \in \mathcal{N}_k, \tag{11}$$

where

$$g_k = [g_{k,0}, \ldots, g_{k,H-1}] \in \mathbb{R}^{H d_a}, \qquad g_{k,j} \in \mathbb{R}^{d_a}. \tag{12}$$

In our implementation, g_k is estimated by weighted ridge regression. It provides a local first-order approximation to how downstream utility varies with the acquisition plan along directions represented by the sampled counterfactual population (Conn et al., 2009; Roberts, 2025; Ma et al., 2025).

The expectation over counterfactual alternatives is then approximated using a small set of sampled reference plans $\mathcal{R}_k$. Let ω_{ki} denote the normalized weight assigned to reference i, with $\sum_{i \in \mathcal{R}_k} \omega_{ki} = 1$. Replacing the path-wise utility sensitivity by the local slope g_k and the counterfactual expectation by these sampled references gives the practical CTA attribution

$$\Phi_{k,j} = \sum_{i \in \mathcal{R}_k} \omega_{ki}\, g_{k,j}^\top \left(a_{t+j}^{(k)} - a_{t+j}^{(i)} \right). \tag{13}$$

Under this local first-order approximation, the path integral in Definition 1 reduces to the difference between its endpoint acquisitions, so $\Phi_{k,j}$ serves as a practical estimate of the ideal contribution $z_{k,j}^\star$. We use $R = |\mathcal{R}_k| \in \{1, 2\}$ representative references selected from the same-state plan population, with preference given to plans whose standardized advantages are closest to zero. These plans provide a natural counterfactual baseline because they lie near the boundary between positive- and negative-advantage outcomes.

The plan-level advantage and CTA attribution play complementary roles in the policy update. The advantage $\mathcal{A}_t^{(k)}$ determines whether the sampled plan should be reinforced or suppressed, whereas $\Phi_{k,j}$ determines how this update is distributed across the acquisitions within the plan. We therefore center and normalize the raw attribution within each plan, using $\widehat{\Phi}_{k,j} = (\Phi_{k,j} - \overline{\Phi}_k)/\sqrt{H^{-1} \sum_l (\Phi_{k,l} - \overline{\Phi}_k)^2 + \epsilon}$, where $\overline{\Phi}_k = H^{-1} \sum_l \Phi_{k,l}$. This removes the common scale of the attribution and retains only the relative credit across acquisition positions.

We then align the relative attribution with the direction of the plan-level update by defining $c_{k,j} = \mathrm{sgn}(\mathcal{A}_t^{(k)}) \widehat{\Phi}_{k,j}$. The resulting acquisition weight is

$$w_{k,j} = (1 - \alpha) + \alpha H \frac{\exp(c_{k,j}/\tau_c)}{\sum_{l=0}^{H-1} \exp(c_{k,l}/\tau_c)}, \tag{14}$$

where α controls the strength of acquisition-level credit and τ_c controls its concentration. Thus, higher-contribution acquisitions receive stronger reinforcement in positive-advantage plans, while lower-contribution acquisitions receive stronger suppression in negative-advantage plans. The weights have unit mean and are detached during policy optimization.

CTA thus converts same-state cross-plan utility variation into within-plan credit using only plans already sampled and evaluated by the policy.

4 Experiments

We evaluate CTA-DPPO on benchmarks from two sequential acquisition frameworks: APPLE (Schneider et al., 2026) for active perception and ALINE (Huang et al., 2026) for Bayesian experimental design. These experiments assess whether joint multi-step planning with counterfactual acquisition credit improves sequential acquisition across different observation modalities and utility objectives. We additionally compare CTA-DPPO with standard DPPO using the same diffusion-policy backbone, training budget, and task objective to isolate the contribution of CTA.

4.1 Experimental Setup

Benchmarks and baselines. We consider three sequential acquisition tasks. CIFAR-10 active perception and TactileMNISTVolume follow the experimental settings of APPLE (Schneider et al., 2026), covering visual classification and continuous tactile estimation, respectively. We compare against the corresponding APPLE acquisition policies and evaluate our method under the same task objectives. We further consider the Location Finding benchmark from ALINE (Huang et al., 2026) to evaluate the method in a Bayesian optimal experimental design (BOED) setting, where acquisition quality is measured through sequential information gain. On this task, we compare against ALINE and the BOED baselines reported in the original work. Detailed task definitions, model architectures, rewards, and dataset-specific training protocols are provided in Appendix B.

Diffusion-policy implementation. Across all experiments, the acquisition policy is parameterized as the diffusion policy described in Section 3.2, initialized by behavior cloning and subsequently fine-tuned with CTA-DPPO. For CIFAR-10 and TactileMNISTVolume, we use a planning horizon of $H = 8$, execute $E = 2$ acquisitions before replanning, and sample $K = 32$ counterfactual plans from each decision state. For Location Finding, we use $H = 4$, $E = 1$, and $K = 64$. The neighborhood size M, reference count $R = |\mathcal{R}_k|$ are selected separately for each task. For tasks with a learned predictor, we alternate between policy optimization and predictor adaptation rather than updating both simultaneously. This protocol allows us to separately examine improvements arising from the acquisition policy itself and those introduced by subsequently adapting the predictor to the observations collected by the updated policy. The complete task-specific configurations are reported in Appendix B.

To separate the effect of CTA from that of the diffusion-policy parameterization itself, we additionally train a standard DPPO baseline using the same diffusion actor, task return, initialization, and optimization budget. CTA-DPPO differs only in using the counterfactual attribution described in Section 3.3 to redistribute the policy-learning signal across acquisition positions. This comparison is used as an ablation of our proposed credit attribution mechanism.

All results produced by our implementation are reported over three random seeds. All experiments are run on NVIDIA A100 GPUs.

Table 1: Final accuracy on CIFAR-10 active perception.

Method	Accuracy ↑
Reported APPLE results	
APPLE-RND	0.670
APPLE-CrossQ	0.730
APPLE-SAC	0.760
Our controlled evaluation	
APPLE-CrossQ	0.728 ± 0.052
DPPO + predictor adaptation	0.762 ± 0.038
CTA-DPPO + predictor adaptation (ours)	$\mathbf{0.802 \pm 0.018}$

4.2 Main Results

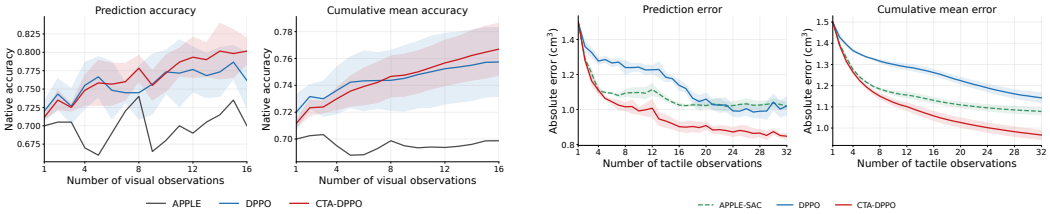
CIFAR-10 active perception. We follow the CIFAR-10 active-perception setting of APPLE (Schneider et al., 2026). At each step, the agent observes only a local 5×5 RGB glimpse of a 32×32 CIFAR-10 image and must classify the image into one of ten classes from the accumulated partial observations. The continuous action $a_t \in [-1, 1]^2$ controls the displacement of the glimpse across the image, so successful classification requires the policy to actively select informative regions over a sequence of observations.

Table 1 compares CTA-DPPO with the reported APPLE baselines and a DPPO baseline using the same diffusion-policy backbone. Both DPPO and CTA-DPPO are initialized by behavior cloning trajectories generated by APPLE-CrossQ, providing the same acquisition-policy initialization. DPPO already exceeds the strongest reported APPLE result, supporting the benefit of explicitly modeling multi-step acquisition plans with a diffusion policy. CTA-DPPO further improves final accuracy by 4.0 percentage points over DPPO, showing that counterfactual acquisition-level credit provides an additional benefit beyond the diffusion-policy formulation alone.

TactileMNISTVolume. We next consider TactileMNISTVolume, a tactile regression task from the APPLE benchmark suite (Schneider et al., 2026; 2025). The agent controls a GelSight Mini sensor to explore a randomly placed and oriented 3D MNIST digit and must estimate its volume

Table 2: TactileMNISTVolume results. Lower is better.

Method	Final Error (cm^3)	Mean Error (cm^3)
APPLE-CrossQ	1.088 ± 0.024	1.177 ± 0.017
APPLE-SAC	1.019 ± 0.021	1.078 ± 0.017
DPPO	0.962 ± 0.038	1.096 ± 0.020
CTA-DPPO	$\mathbf{0.847 \pm 0.016}$	$\mathbf{0.967 \pm 0.021}$



(a) CIFAR-10 active perception.

(b) TactileMNISTVolume.

Figure 2: Observation efficiency across sequential acquisition tasks. (a) Classification accuracy on CIFAR-10 as observations are acquired. (b) Volume-prediction error on TactileMNISTVolume over the acquisition trajectory. Curves show the mean over three policy-sampling seeds with 200 episodes per seed;

within a fixed acquisition budget. Unlike classification, accurate volume estimation requires substantially more complete information about the object's shape, making the quality of the exploration policy particularly important.

As shown in Table 2, CTA-DPPO achieves the lowest error under both evaluation metrics. Its final error is approximately 22% lower than APPLE-CrossQ and 17% lower than APPLE-SAC, while also improving over standard DPPO. The trajectory-averaged error shows the same trend, indicating that CTA-DPPO improves acquisition quality throughout the sensing process rather than only at the final prediction.

Bayesian experimental design. We further evaluate CTA on the Location Finding benchmark from ALINE (Huang et al., 2026) to test whether the same credit mechanism extends beyond the APPLE active-perception setting. Here, acquisitions correspond to experimental designs and performance is evaluated using sPCE.

Table 3: Results on the Location Finding task. Higher is better.

Method	sPCE ↑
Reported ALINE results	
Random	5.170 ± 0.050
VPCE	5.250 ± 0.220
DAD	7.330 ± 0.060
vsOED	7.300 ± 0.060
RL-BOED	7.700 ± 0.060
Our controlled evaluation	
Native ALINE	8.940 ± 0.227
DPPO (without CTA)	9.246 ± 0.146
CTA-DPPO	$\mathbf{9.336 \pm 0.186}$

Table 3 reports both published ALINE results and our controlled evaluation. CTA achieves the highest mean sPCE in our comparison, improving over Native ALINE while preserving the same underlying information objective. This result shows that counterfactual within-plan credit extends naturally to sequential Bayesian experimental design and can provide additional gains beyond the original acquisition strategy. The result therefore primarily shows that the counterfactual credit mechanism transfers to the BOED setting while preserving the underlying information objective. Representative acquisition trajectories and posterior localization are shown in Appendix E.

Observation efficiency. We further evaluate how efficiently the learned acquisition policies use the available observation budget. Figure 2 shows performance as a function of the number of observations on CIFAR-10 and TactileMNISTVolume. CTA-DPPO improves performance over much of the acquisition trajectory, indicating more effective use of the available observations.

4.3 Ablation Study and Analysis

We further analyze CTA on CIFAR-10 through two controlled studies. First, we compare CTA-DPPO with standard DPPO under the same alternating policy–predictor schedule to isolate the effect of acquisition-level credit. Second, we vary the local neighborhood size M to examine how the locality of the counterfactual population affects the first-order utility estimate. Additional ablations on the number of sampled counterfactual plans K and the number of reference plans R are reported in Appendix C.

Effect of acquisition-level credit. We compare CTA-DPPO with standard DPPO under identical initialization, training budget, and alternating policy–predictor schedules. This protocol separates gains from policy optimization from those introduced by subsequent predictor adaptation.

As shown in Table 4, CTA-DPPO achieves higher accuracy at most policy checkpoints compared with DPPO. The stage-wise results further help distinguish the direct effect of CTA on the acquisition policy from its downstream effect on predictor adaptation.

Local counterfactual neighborhood. CTA estimates the utility sensitivity around each sampled plan from a local population of neighboring counterfactuals. We therefore study the neighborhood size M, which directly controls the locality of this approximation. We fix $K = 32$, $R = 2$, and vary only the number of neighboring plans used to estimate the local utility slope.

Table 4: CIFAR-10 accuracy during alternating policy–predictor training.

Stage	DPPO	CTA-DPPO
5k	0.705 ± 0.018	0.715 ± 0.010
10k	0.703 ± 0.015	0.717 ± 0.023
15k	0.710 ± 0.005	0.713 ± 0.010
Predictor update	0.768 ± 0.015	0.775 ± 0.013
20k	0.787 ± 0.018	0.797 ± 0.013
25k	0.785 ± 0.010	0.780 ± 0.023
30k	0.765 ± 0.015	0.785 ± 0.010
Predictor update	0.762 ± 0.038	$\mathbf{0.802 \pm 0.018}$

Table 5 shows that CTA is sensitive to the locality of the counterfactual neighborhood. With $M = 16$, CTA-DPPO improves accuracy from 0.762 ± 0.038 for standard DPPO to 0.802 ± 0.018. In contrast, $M = 8$ provides almost no improvement over DPPO, while increasing the neighborhood to $M = 24$ degrades performance. This behavior is consistent with the local first-order approximation underlying CTA: too few neighbors provide insufficient variation for estimating the utility slope, whereas an overly broad neighborhood can include plans for which a single local linear approximation is less representative. The result therefore suggests that CTA benefits from a counterfactual population that is sufficiently diverse for slope estimation while remaining local to the target plan.

Intervention-based slot importance. We further test whether CTA correctly ranks the importance of individual acquisition positions. Following Appendix D, we freeze the trained policy and predictor, sample counterfactual plans from the same decision state, and intervene on each slot by replacing its action with the corresponding reference action. We then rank the $H = 8$ slots within each plan by increasing attribution magnitude $|\Phi_{k,j}|$ and measure the resulting absolute utility change.

Table 5: Effect of neighborhood size M on CIFAR-10.

Method	M	Accuracy ↑
DPPO	–	0.762 ± 0.038
CTA-DPPO	8	0.763 ± 0.015
CTA-DPPO	16	$\mathbf{0.802 \pm 0.018}$
CTA-DPPO	24	0.755 ± 0.025

Table 6 shows a monotonic increase in empirical intervention effect with CTA rank. Replacing the highest-attribution slot changes downstream utility by 0.104 on average, compared with 0.042 for the lowest-attribution slot, a $2.48\times$ difference. Thus, the acquisition positions assigned greater importance by CTA are also those whose replacement has larger realized effects on plan utility. Additional protocol, directional, and statistical analyses are provided in Appendix D.

Table 6: Utility change under slot replacement, ranked by CTA attribution.

Attribution rank	Mean $\lvert\Delta J\rvert$	95% CI
1: lowest	0.042	[0.032, 0.052]
2	0.050	[0.040, 0.062]
3	0.053	[0.041, 0.066]
4	0.059	[0.047, 0.072]
5	0.066	[0.052, 0.081]
6	0.071	[0.057, 0.087]
7	0.085	[0.069, 0.102]
8: highest	0.104	[0.085, 0.126]

5 Conclusion

We introduced a diffusion-policy framework with Counterfactual Taylor Attribution (CTA) for sequential information-acquisition tasks, including active perception and sequential Bayesian experimental design. CTA exploits utility variation among counterfactual plans from the same decision state to estimate acquisition-level contributions and refine within-plan credit while preserving the original task objective. Across both benchmark families, CTA-DPPO improves over uniform-credit diffusion-policy optimization, while intervention analysis shows that higher-attribution acquisition positions have larger realized effects on plan utility. These results support counterfactual plan variation as a useful source of structured supervision for sequential policy learning.

AI USE STATEMENT

We used generative AI tools to improve the readability and presentation of the manuscript and to assist with LaTeX formatting. All AI-assisted edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed algorithm, training procedure, and evaluation methodology in the main paper and appendices. The appendices additionally document the implementation and experimental settings necessary to reproduce the reported results. Our code and configuration files will be released publicly upon acceptance.

REFERENCES

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Tom Blau, Edwin V Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *International conference on machine learning*, pp. 2107–2128. PMLR, 2022.

Yuhong Cao, Jeric Lew, Jingsong Liang, Jin Cheng, and Guillaume Sartoretti. Dare: Diffusion policy for autonomous robot exploration. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11987–11993. IEEE, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. A unified Taylor framework for revisiting attribution methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11462–11469, 2021.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Unifying fourteen post-hoc attribution methods with Taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4625–4640, 2024.

Wenlong Deng, Yi Ren, Yushu Li, Boying Gong, Danica Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. Token hidden reward: Steering exploration-exploitation in group relative deep reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 122109–122133, 2026.

Sascha Fleer, Alexandra Moringen, Roberta L Klatzky, and Helge Ritter. Learning efficient haptic shape exploration with a rigid tactile sensor array. *PloS one*, 15(1):e0226880, 2020.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International conference on machine learning*, pp. 3384–3395. PMLR, 2021.

Itsuki Hirako, Ryo Hakoda, Yubin Liu, Matthew Hwang, Yoshihiro Sato, and Takeshi Oishi. ScanDP: Generalizable 3D scanning with diffusion policy. *arXiv preprint arXiv:2603.10390*, 2026.

Edward Hu, Jie Wang, Xingfang Yuan, Fiona Luo, Muyao Li, Gaspard Lambrechts, Oleh Rybkin, and Dinesh Jayaraman. Real-world reinforcement learning of active perception behaviors. *Advances in Neural Information Processing Systems*, 38:117272–117308, 2026.

Xun Huan and Youssef M Marzouk. Sequential Bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320*, 2016.

Daolang Huang, Xinyi Wen, Ayush Bharti, Samuel Kaski, and Luigi Acerbi. ALINE: joint amortization for Bayesian inference and active data acquisition. *Advances in Neural Information Processing Systems*, 38:54068–54102, 2026.

Jeric Lew, Yuhong Cao, Derek Ming Siang Tan, and Guillaume Sartoretti. AID: Agent intent from diffusion for multi-agent informative path planning. In *International Conference on Swarm Intelligence*, pp. 41–54. Springer, 2026.

Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual Shapley credit assignment. *arXiv preprint arXiv:2607.16999*, 2026a.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. SSVPO: Effective step-level credit assignment for RL training of language models. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Saida Liu, Nikolay Atanasov, and Shumon Koga. Matt-diff: Multimodal active target tracking by diffusion policy. *arXiv preprint arXiv:2511.11931*, 2025.

Kaiwen Ma, Luis Miguel Rios, Hua Zheng, Nikolaos V Sahinidis, and Sreekanth Rajagopalan. Model-and-search: a derivative-free local optimization algorithm. *Computational Optimization and Applications*, 92(3):889–921, 2025.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via Q-score matching. *arXiv preprint arXiv:2312.11752*, 2023.

Allen Ren, Justin Lidard, Lars Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *International Conference on Learning Representations*, volume 2025, pp. 77288–77329, 2025.

Lindon Roberts. Model construction for convex-constrained derivative-free optimization. *SIAM Journal on Optimization*, 35(2):622–650, 2025.

Tim Schneider, Guillaume Duret, Cristiana de Farias, Roberto Calandra, Liming Chen, and Jan Peters. Tactile mnist: Benchmarking active tactile perception. *arXiv preprint arXiv:2506.06361*, 2025.

Tim Schneider, Cristiana de Farias, Roberto Calandra, Liming Chen, and Jan Peters. APPLE: Toward general active perception via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 97598–97624, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3D shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 34:16064–16078, 2021.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Jingxi Xu, Shuran Song, and Matei Ciocarlie. Tandem: Learning joint exploration and decision making with tactile sensors. *IEEE Robotics and Automation Letters*, 7(4):10391–10398, 2022.

A Proof of Counterfactual Utility Decomposition

For a sampled plan

$$A_t^{(k)} = \left[a_t^{(k)}, \ldots, a_{t+H-1}^{(k)}\right]$$

and a counterfactual plan

$$A_t^{(r)} = \left[a_t^{(r)}, \ldots, a_{t+H-1}^{(r)}\right],$$

consider any continuously differentiable path

$$\Xi_{kr} : [0,1] \to \mathbb{R}^{H \times d_a}, \qquad \Xi_{kr}(0) = A_t^{(r)}, \qquad \Xi_{kr}(1) = A_t^{(k)}. \tag{15}$$

We write the path component-wise as

$$\Xi_{kr}(\lambda) = \left[\xi_{kr,0}(\lambda), \ldots, \xi_{kr,H-1}(\lambda)\right], \tag{16}$$

where $\xi_{kr,j}(\lambda) \in \mathbb{R}^{d_a}$ describes the acquisition at position j along the path. Its endpoints satisfy

$$\xi_{kr,j}(0) = a_{t+j}^{(r)}, \qquad \xi_{kr,j}(1) = a_{t+j}^{(k)}. \tag{17}$$

Assume that U_t is continuously differentiable along Ξ_{kr}. Since

$$U_t(\Xi_{kr}(\lambda)) = U_t(\xi_{kr,0}(\lambda), \ldots, \xi_{kr,H-1}(\lambda)),$$

the multivariate chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} U_t(\Xi_{kr}(\lambda)) = \sum_{j=0}^{H-1} \nabla_{a_{t+j}} U_t(\Xi_{kr}(\lambda))^\top \dot{\xi}_{kr,j}(\lambda), \tag{18}$$

where $\dot{\xi}_{kr,j}(\lambda) = \mathrm{d}\xi_{kr,j}(\lambda)/\mathrm{d}\lambda$. Starting from Definition 1 and summing over acquisition positions,

$$\sum_{j=0}^{H-1} z_{k,j}^\star = \sum_{j=0}^{H-1} \mathbb{E}_{A_t^{(r)} \sim q_t} \left[\int_0^1 \nabla_{a_{t+j}} U_t(\Xi_{kr}(\lambda))^\top \dot{\xi}_{kr,j}(\lambda)\, \mathrm{d}\lambda\right] \tag{19}$$

$$= \mathbb{E}_{A_t^{(r)} \sim q_t} \left[\int_0^1 \sum_{j=0}^{H-1} \nabla_{a_{t+j}} U_t(\Xi_{kr}(\lambda))^\top \dot{\xi}_{kr,j}(\lambda)\, \mathrm{d}\lambda\right] \tag{20}$$

$$= \mathbb{E}_{A_t^{(r)} \sim q_t} \left[\int_0^1 \frac{\mathrm{d}}{\mathrm{d}\lambda} U_t(\Xi_{kr}(\lambda))\ \mathrm{d}\lambda\right], \tag{21}$$

Applying the fundamental theorem of calculus gives

$$\sum_{j=0}^{H-1} z_{k,j}^\star = \mathbb{E}_{A_t^{(r)} \sim q_t} \left[U_t(\Xi_{kr}(1)) - U_t(\Xi_{kr}(0))\right] \tag{22}$$

$$= \mathbb{E}_{A_t^{(r)} \sim q_t} \left[U_t\left(A_t^{(k)}\right) - U_t\left(A_t^{(r)}\right)\right] \tag{23}$$

$$= U_t\left(A_t^{(k)}\right) - \mathbb{E}_{A_t^{(r)} \sim q_t} \left[U_t\left(A_t^{(r)}\right)\right] \tag{24}$$

$$= M_t\left(A_t^{(k)}\right). \tag{25}$$

Therefore, the counterfactual acquisition contributions exactly decompose the counterfactual utility margin:

$$\sum_{j=0}^{H-1} z_{k,j}^\star = M_t\left(A_t^{(k)}\right). \tag{26}$$

The decomposition assigned to individual acquisition positions may depend on the chosen counterfactual path, but their sum depends only on the two endpoint plans.

B EXPERIMENTAL DETAILS

This appendix provides the task definitions, diffusion-policy configurations, training procedures, and evaluation protocols used in our experiments. CIFAR-10 and TactileMNISTVolume use the active-perception environments and prediction objectives of APPLE (Schneider et al., 2026), while Location Finding uses the simulator and pretrained inference model of ALINE (Huang et al., 2026).

B.1 DIFFUSION-POLICY AND CTA SETTINGS

At each decision state, the diffusion policy generates an H-step acquisition plan and executes the first E actions before replanning. For CTA-DPPO, K plans are sampled from the same decision state to construct counterfactual training signals. For each target plan, M nearest plans are used for local regression, while $R = |\mathcal{R}_k|$ denotes the number of counterfactual references used for Taylor attribution, and d_a denotes the acquisition-action dimension,

Table 7 summarizes the configurations associated with the main reported results. The neighborhood size M and reference count R play different roles and are not shared across all experiments.

Table 7: Task-specific diffusion-policy and CTA settings used for the reported results.

Experiment	d_a	H	E	K	M	R	τ_c
CIFAR-10	2	8	2	32	16	2	1.0
Volume, Cross-Q	3	8	2	32	16	2	1.0
Volume, SAC	3	8	2	32	16	1	0.5
Location Finding	2	4	1	64	32	2	1.0

For the Volume experiments, CTA is emphasized during the early stage of policy optimization, when counterfactual plans exhibit larger variation and provide a stronger acquisition-level supervision signal. As the sampled plans become more concentrated later in training, we gradually reduce the CTA-based redistribution and transition toward the standard DPPO update.

Diffusion actor. For the APPLE tasks, the diffusion actor is a conditional one-dimensional U-Net operating on the H-step action sequence. It uses channel widths $(64, 128)$, kernel size 5, GroupNorm with 8 groups, Mish activations, residual connections, and a 32-dimensional sinusoidal diffusion-time embedding. We use 20 denoising steps with a cosine noise schedule and the standard noise-prediction objective. The policy is conditioned on the frozen 128-dimensional history representation from the corresponding APPLE perception model.

Alternating policy and predictor optimization. We alternate between policy optimization and predictor adaptation. During each policy stage, the perception encoder and prediction head are frozen, and only the diffusion actor is optimized with CTA-DPPO. Policy optimization updates the final ten, lowest-noise denoising transitions while retaining the first ten transitions from the frozen behavior-cloned denoiser. Each batch contains eight root states with $K = 32$ counterfactual plans per root and is reused for ten PPO epochs. We use AdamW with learning rate 10^{-5}, $(\beta_1, \beta_2) = (0.5, 0.999)$, zero weight decay, and PPO clipping coefficient 10^{-3}.

After each policy stage, the diffusion policy and perception encoder are frozen, and only the prediction head is updated using observations collected by the current policy. The adapted predictor is then used in the subsequent policy stage. This alternating procedure allows the acquisition policy to improve under a fixed inference model and the predictor to subsequently adapt to the observation distribution induced by the updated policy.

B.2 CIFAR-10 ACTIVE PERCEPTION

Task and perception model. We use the CIFAR-10 active-perception environment from APPLE (Schneider et al., 2026). Each episode contains a 32×32 RGB image, while the agent observes a local 5×5 glimpse. The two-dimensional acquisition action controls the displacement of the glimpse, and the predictor outputs logits over the ten CIFAR-10 classes.

The perception model uses the pretrained APPLE history encoder with embedding dimension 128, three transformer layers, eight attention heads, and feed-forward dimension 64. The prediction head is a linear ten-class classifier.

Evaluation. Each checkpoint is evaluated on the same fixed set of 200 test episodes. Policy fine-tuning is repeated using random seeds $\{0, 1, 2\}$, and we report the mean and sample standard deviation across these three training runs.

B.3 TACTILEMNISTVOLUME

Task and perception model. We use the TactileMNISTVolume environment from APPLE (Schneider et al., 2026; 2025). The agent explores a simulated three-dimensional MNIST object using a tactile sensor and predicts its physical volume. Each observation contains a 32×32 tactile image together with the three-dimensional sensor position. The action is a three-dimensional relative sensor motion in $[-1, 1]^3$.

The tactile image is encoded with the pretrained APPLE vision transformer and combined with the remaining observations through the APPLE history transformer. The prediction head outputs a single continuous volume estimate.

Evaluation. Each trained model is evaluated on 200 test episodes using each of three policy-sampling seeds. The same 200 environment-reset cases are reused across the three evaluations and across compared policies.

B.4 LOCATION FINDING

Task. We use the Location Finding environment from ALINE (Huang et al., 2026). A hidden source location $\theta \in [0, 1]^2$ is inferred from noisy measurements collected at sequentially selected locations $x \in [0, 1]^2$. Each episode begins with one random observation and contains 30 observations in total, corresponding to 29 policy-selected acquisitions.

For the reported candidate-pool experiment, 2,000 candidate measurement locations and their noisy outcomes are sampled at episode initialization. Native ALINE selects among the remaining candidates using its acquisition score, while a diffusion action is mapped to $[0, 1]^2$ and projected to the nearest unused candidate. Both methods are evaluated on matched source locations, initial observations, candidate pools, and simulator seeds.

Evaluation. Acquisition quality is evaluated using sequential Prior Contrastive Estimation (sPCE). For generating parameter θ_0 and $L = 10^6$ prior contrast samples $\theta_1, \ldots, \theta_L$, we compute

$$u_t(\theta) = \sum_{i=1}^{t} \log p(y_i \mid x_i, \theta), \tag{27}$$

and

$$\mathrm{sPCE}_t = u_t(\theta_0) - \log \left[\frac{1}{L+1} \sum_{l=0}^{L} \exp u_t(\theta_l) \right]. \tag{28}$$

We report final sPCE_{30} over 100 matched episodes together with a normal-approximation 95% confidence interval.

C ADDITIONAL CTA ABLATIONS

We provide additional ablations on the size of the counterfactual plan population and the number of reference plans used by CTA. Unless otherwise specified, experiments use the same two-round policy–predictor schedule as the main CIFAR-10 evaluation, with $M = 16$, ridge coefficient 0.1, credit temperature 1, and $E = 2$.

Number of counterfactual plans. We vary the number of plans K sampled from each decision state. Unlike M, which determines the local neighborhood used to estimate the utility slope, K controls the size of the counterfactual population from which neighbors and reference plans are selected. Because increasing K also increases rollout and policy-update cost, these experiments are not compute-matched.

Table 8: Effect of candidate-plan count K on CIFAR-10.

Method	K	Accuracy ↑
APPLE-CrossQ (Schneider et al., 2026)	–	0.728 ± 0.052
CTA-DPPO	24	0.782 ± 0.023
CTA-DPPO	32	$\mathbf{0.802 \pm 0.018}$
CTA-DPPO	48	0.783 ± 0.010
CTA-DPPO	64	0.772 ± 0.006
CTA-DPPO	128[†]	0.768 ± 0.004

[†]Preliminary result over two training seeds.

Performance does not improve monotonically with K, indicating that simply increasing the number of sampled plans is not sufficient to improve credit estimation. Nevertheless, CTA-DPPO achieves higher mean accuracy than the reported APPLE-crossQ result across all tested values of K. This suggests that the benefit of CTA is not restricted to a particular candidate-plan count and that obtaining informative local counterfactual variation is more important than maximizing the total population size.

Number of counterfactual references. We next vary the number of reference plans $R = |\mathcal{R}_k|$ while fixing $K = 32$, $M = 16$. For $R = 1$, each acquisition is attributed relative to a single near-zero-advantage counterfactual plan. For $R = 2$, CTA averages the contributions obtained from two such references, reducing dependence on any single counterfactual baseline.

Table 9: Effect of reference count R on CIFAR-10.

Method	R	Accuracy ↑
DPPO	–	0.762 ± 0.038
CTA-DPPO	1	0.780 ± 0.005
CTA-DPPO	2	$\mathbf{0.802 \pm 0.018}$

Increasing the reference count from $R = 1$ to $R = 2$ improves the mean accuracy from 0.780 to 0.802. Using multiple plausible counterfactual references therefore appears to provide a more representative estimate of acquisition contribution and reduces sensitivity to the choice of an individual reference plan.

D Intervention-Based Validation of CTA Slot Importance

We provide additional details for the intervention analysis in Section 4.3. The purpose of this experiment is to test whether acquisition positions assigned larger CTA attributions have larger realized effects on plan utility when directly perturbed. The analysis is purely diagnostic: the policy and predictor are frozen throughout, and intervention rollouts are not used for training or for recomputing CTA.

Evaluation protocol. We evaluate the 30k-step policy checkpoints from three training seeds $(0, 1, 2)$, together with the predictors used during their corresponding policy training. Each checkpoint is evaluated on the same 64 CIFAR-10 images, with one prescribed decision root per image at acquisition step $t \in \{0, 4, 8\}$. At each root, we sample $K = 32$ candidate plans, compute CTA using $M = 16$, and uniformly select eight plans as intervention targets. This analysis contains 1,528 target plans, and 12,224 slot-level interventions.

Slot intervention. For a target plan

$$A_t^{(k)} = \left[a_t^{(k)}, \ldots, a_{t+H-1}^{(k)}\right], \tag{29}$$

we construct an intervention by replacing a single acquisition position j with the corresponding action from a counterfactual reference $A_t^{(r)}$,

$$A_t^{(k,j\leftarrow r)} = \left[a_t^{(k)}, \ldots, a_{t+j}^{(r)}, \ldots, a_{t+H-1}^{(k)}\right]. \tag{30}$$

The original and modified plans are replayed from the same decision state, with all other planned actions unchanged. We measure the resulting signed utility change as

$$\Delta J_{k,j}^{(r)} = J_t\left(A_t^{(k)}\right) - J_t\left(A_t^{(k,j\leftarrow r)}\right). \tag{31}$$

For the $R = 2$ reference construction used in the main configuration, both reference actions are evaluated independently. Their signed effects are averaged before taking the magnitude,

$$\Delta J_{k,j}^{\text{pair}} = \frac{1}{2}\left(\Delta J_{k,j}^{(r_1)} + \Delta J_{k,j}^{(r_2)}\right). \tag{32}$$

This matches the signed averaging used when constructing the corresponding CTA attribution.

Attribution-rank analysis. For each target plan, the eight acquisition positions are ranked by the raw attribution magnitude $|\Phi_{k,j}|$, before centering, sign alignment, or softmax weighting. Rank 1 denotes the lowest-attribution slot and rank 8 the highest; these ranks are unrelated to the execution order of the acquisitions.

For each rank, we report the mean empirical intervention magnitude $\left|\Delta J_{k,j}^{\text{pair}}\right|$.

As shown in Table 6, the intervention effect increases monotonically across all eight attribution ranks. The highest-attribution slot produces approximately $2.48\times$ the utility change of the lowest-attribution slot. The rank-8 minus rank-1 difference is 0.0622, with a 95% confidence interval of $[0.0494, 0.0768]$. This monotonic relationship provides direct empirical support that CTA attribution magnitude ranks acquisition positions according to their realized importance under counterfactual replacement.

Directional agreement. We additionally examine whether the sign of the raw attribution predicts the direction of the intervention effect. A positive attribution indicates that retaining the original acquisition is predicted to be preferable to replacing it with its counterfactual reference. We therefore count a directional match when

$$\text{sgn}(\Phi_{k,j}) = \text{sgn}\left(\Delta J_{k,j}^{\text{pair}}\right). \tag{33}$$

Near-zero attributions and intervention effects are excluded from this analysis. The resulting sign agreement is 61.4%, compared with an independence baseline of 51.1% computed from the marginal positive and negative frequencies.

Statistical analysis. All confidence intervals are obtained by bootstrap resampling at the image level. This keeps interventions originating from the same image within the same bootstrap unit rather than treating the individual slot-level comparisons as independent observations. The reported intervals are conditional on the three evaluated policy checkpoints.

E Additional Location Finding Analysis

We provide additional qualitative and observation-efficiency analyses for Location Finding. These analyses complement the final-budget sPCE comparison in the main text by showing where the policies acquire measurements and how their performance evolves as observations accumulate.

Acquisition locations and posterior localization. Figure 3 compares ALINE, DPPO, and CTA-DPPO on a representative illustrative episode, using the same hidden source and initial observation. Each method acquires 29 additional measurements, giving 30 observations in total. For each resulting history, we compute the approximate posterior using the same frozen ALINE posterior head. The full-domain views show the acquisition locations and log-posterior contours, while the lower panels show the posterior's concentration near the source.

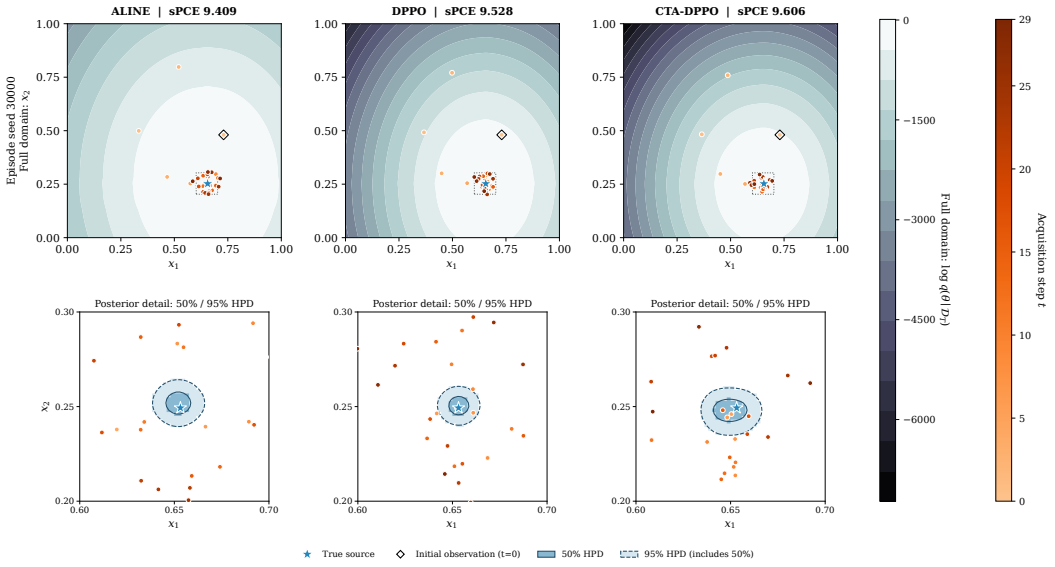


Figure 3: Acquisition locations and approximate posteriors for a shared Location Finding episode. Orange point colors indicate acquisition order; blue stars mark the true source, and hollow diamonds identify the initial observation. Upper panels show full-domain log-density contours of the frozen ALINE posterior approximation $q(\theta \mid \mathcal{D}_{30})$. Lower panels show enlarged, nested 50% and 95% highest-posterior-density regions, with posterior mass normalized over the displayed domain $[0, 1]^2$.

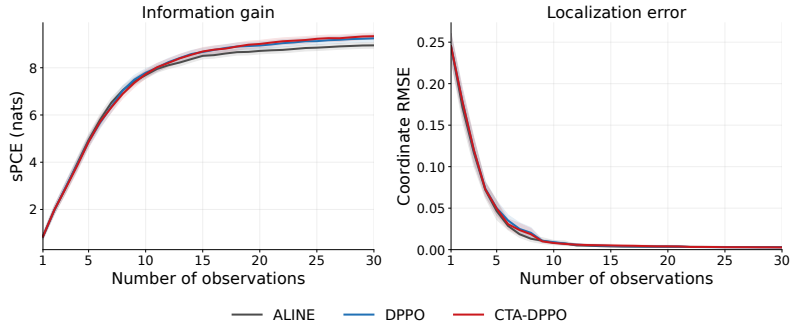


Figure 4: Observation efficiency on Location Finding. Left: mean sPCE in nats (higher is better). Right: coordinate RMSE of the source-location estimate (lower is better).

Observation efficiency. Figure 4 reports mean sPCE and coordinate root-mean-squared error (RMSE) across our full evaluation set of 200 Location Finding episodes. Performance is evaluated after each observation. The horizontal axis includes the initial observation: 30 observations correspond to 29 policy-selected acquisitions. The localization estimates use the same frozen ALINE posterior head, conditioned on each method's observation history. These curves characterize acquisition efficiency across the evaluation set throughout the observation budget.

F Qualitative Acquisition Trajectories

Figure 5 compares representative CIFAR-10 acquisition trajectories produced by APPLE, uniform-credit DPPO, and CTA-DPPO on the same input images. Arrows indicate the order in which acquisitions are executed. The examples show that the methods often share similar early acquisition patterns, while CTA-DPPO can explore additional object regions rather than repeatedly sampling within a small local neighborhood.

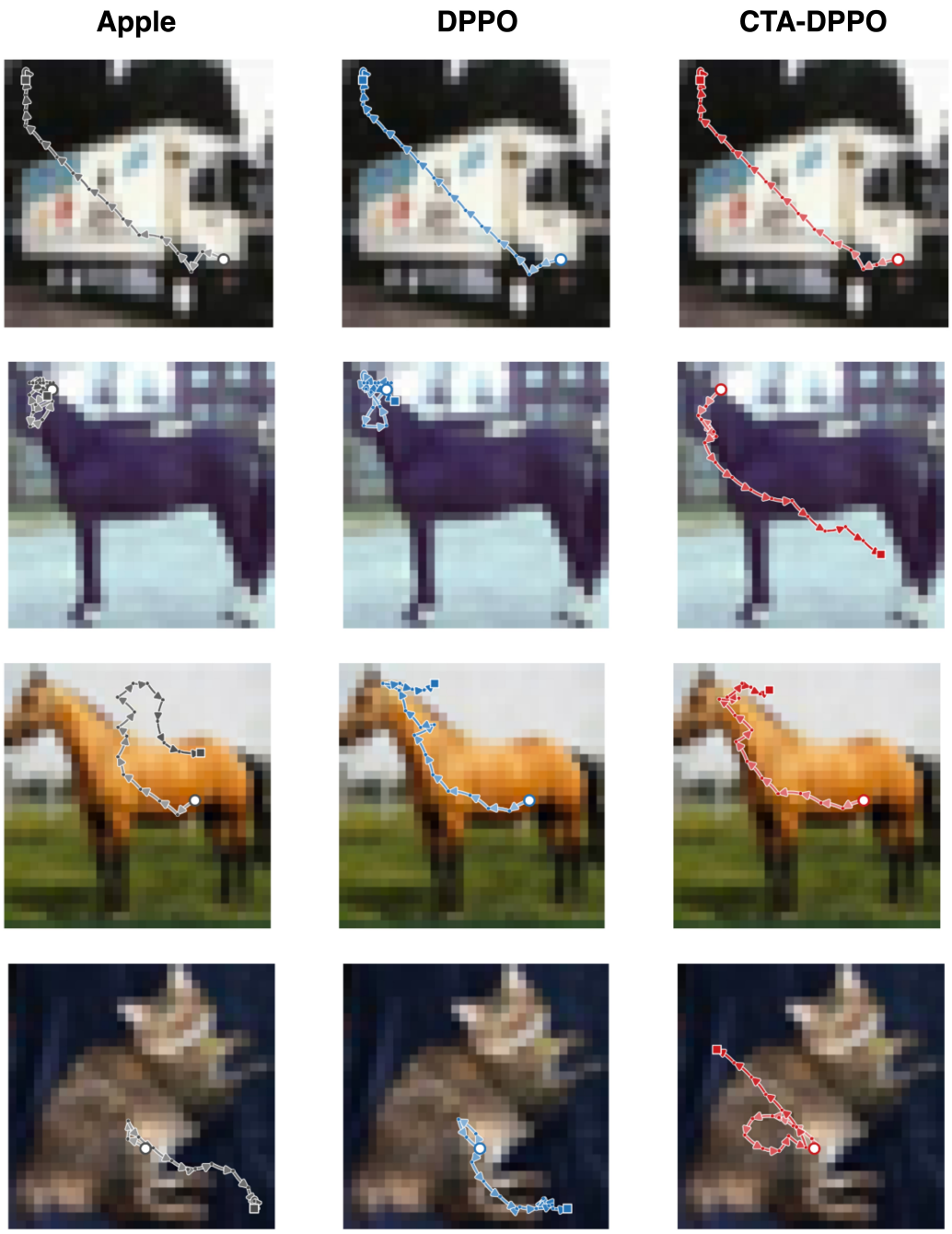


Figure 5: Representative CIFAR-10 acquisition trajectories for APPLE, DPPO, and CTA-DPPO on the same input images. Arrows indicate acquisition order.