Active Target Discovery under Uninformative Prior: The Power of Permanent and Transient Memory

Anindya Sarkar*, **Binglin Ji***, **Yevgeniy Vorobeychik**
{anindya, binglin.j, yvorobeychik}@wustl.edu,
Department of Computer Science and Engineering
Washington University in St. Louis, USA

Abstract

In many scientific and engineering fields, where acquiring high-quality data is expensive—such as medical imaging, environmental monitoring, and remote sensing—strategic sampling of unobserved regions based on prior observations is crucial for maximizing discovery rates within a constrained budget. The rise of powerful generative models, such as diffusion models, has enabled active target discovery in partially observable environments by leveraging learned priors—probabilistic representations that capture underlying structure from data. With guidance from sequentially gathered task-specific observations, these models can progressively refine exploration and efficiently direct queries toward promising regions. However, in domains where learning a strong prior is infeasible due to extremely limited data or high sampling cost (such as rare species discovery, diagnostics for emerging diseases, etc.), these methods struggle to generalize. To overcome this limitation, we propose a novel approach that enables effective active target discovery even in settings with uninformative priors, ensuring robust exploration and adaptability in complex real-world scenarios. Our framework is theoretically principled and draws inspiration from neuroscience to guide its design. Unlike black-box policies, our approach is inherently interpretable, providing clear insights into decision-making. Furthermore, it guarantees a strong, monotonic improvement in prior estimates with each new observation, leading to increasingly accurate sampling and reinforcing both reliability and adaptability in dynamic settings. Through comprehensive experiments and ablation studies across various domains, including species distribution modeling and remote sensing, we demonstrate that our method substantially outperforms baseline approaches.

1 Introduction

Active Target Discovery (ATD) is a fundamental problem in scientific and engineering domains where data acquisition is expensive and environments are only partially observable. Beginning with an unobservable task, the goal is to strategically and sequentially sample glimpses to infer and uncover the underlying target, such as disease regions in medical imaging, pollution hotspots in environmental monitoring, or objects of interest in satellite imagery—all while adhering to a strict sampling budget. This task is critical in scenarios like detecting rare tumors in MRI scans, localizing contaminants in remote areas, or identifying missing persons in search-and-rescue missions. In all these cases, the cost of each observation is high, and acquiring feedback from the ground-truth reward function often requires expert judgment, specialized equipment, and substantial time. Diffusion-guided Active Target Discovery (DiffATD) [1] has been recently proposed as a promising solution to the ATD task. It operates by learning a strong prior over the search space using samples from the domain of interest.

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-2163

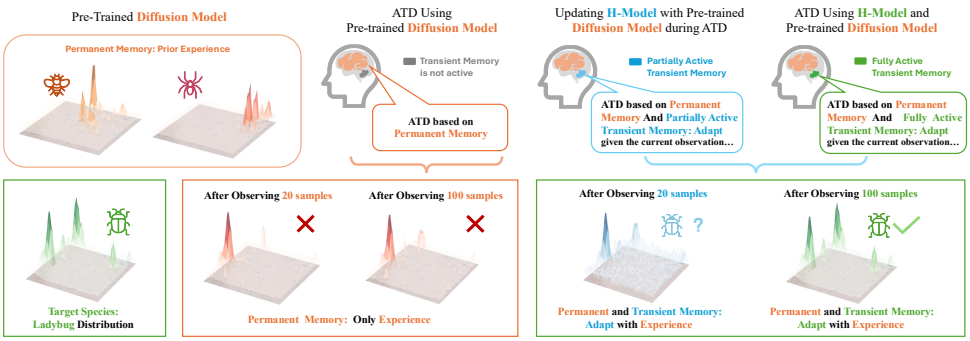


Figure 1: Interplay of Permanent and Transient Memory to Guide Active Target Discovery.

Specifically, DiffATD [1] leverages a diffusion model trained on domain-specific samples to serve as a strong prior, guiding its sampling strategy via learned diffusion dynamics conditioned on the gathered observations to efficiently uncover target regions within a constrained sampling budget. This prior-driven approach allows it to balance exploration and exploitation in a principled way, often outperforming baseline methods in well-studied domains. However, the effectiveness of DiffATD [1] fundamentally hinges on the availability of representative prior samples from the target domain. In many real-world scenarios—such as emerging diseases, rare environmental hazards, or underexplored geographic regions—such samples are either unavailable or prohibitively difficult to obtain. In these cases, the strong prior assumption breaks down, and methods like DiffATD [1] struggle to generalize, limiting their applicability in precisely the settings where efficient target discovery is most critical.

To confront the challenge of ATD without access to prior samples, we draw inspiration from Neuroscience. The real world is inherently dynamic, continually presenting novel and unforeseen situations. Yet, humans navigate such environments seamlessly by drawing on two complementary systems: a long-term memory that encodes structured, generalizable knowledge, and a short-term memory that rapidly adapts to immediate context. For instance, when driving in an unfamiliar city, we rely on our understanding of traffic rules (long-term knowledge) while dynamically responding to unexpected road closures or detours (short-term adaptation). Kumaran [2] suggests that this dual mechanism is rooted in the neocortex and hippocampus, that is, the brain leverages a dual memory system comprised of *permanent memory*—mediated by the neocortex—which encodes structured knowledge through slow learning and supports generalization, and *transient memory*—associated with the hippocampus—which enables rapid acquisition of precise, task-specific information. Inspired by this cognitive mechanism, we argue that emulating such a dual memory structure in artificial agents is essential for solving new tasks in dynamic environments, where both stable generalization and fast adaptability are crucial.

Inspired by this mechanism, we draw on Bayesian inference to tackle ATD in the absence of prior domain data. At the heart of our approach lies an Expectation-Maximization-style algorithm that guarantees monotonic improvement of the sampling prior—modeled as a diffusion process—as new observations unfold across the environment. This evolving prior actively guides the sampling strategy to maximize discovery under a strict budget. Our prior model architecture reflects mechanisms from the brain's dual-memory system by pairing a powerful *pretrained diffusion model* as a long-term or *permanent memory*—capturing rich, generalizable structure across domains—with a lightweight, adaptive module based on *Doob's h-transform* that serves as *transient memory*, enabling rapid contextual adaptation from new observations. The lightweight Doob's h-transform module is updated after every few observations, allowing it to swiftly adapt to the evolving dynamics of the current task. This synergy enables our system to combine the power of global knowledge with the flexibility of fast local adaptation, allowing for efficient, intelligent exploration in novel and data-scarce environments. Crucially, the integration of a pretrained diffusion model with a lightweight network that implements *Doob's h-transform* enables principled sampling from the posterior distribution conditioned on the accumulated observations. This design ensures that each sampling decision is grounded in both prior knowledge and task-specific context, lending theoretical rigor and practical effectiveness to our architecture. Finally, our sampling strategy ranks unobserved points by balancing exploration and exploitation scores, driven by an updated prior and an online-trained reward model that learns the target's characteristics from gathered observations. As more observations are revealed, the prior progressively improves, leading to more accurate rankings that fuel a highly informed active discovery

process, reinforced by theoretical guarantees and compelling empirical evidence. Figure 1 illustrates the core motivation behind our approach. We summarize our contributions as follows:

- We introduce a novel and principled framework, Expectation Maximized Permanent Temporary Diffusion Memory (EM-PTDM), designed to uncover targets of interest within a strict sampling budget in partially observable environments. Unlike prior methods, EM-PTDM operates without relying on task-specific prior domain data, significantly broadening its applicability. Furthermore, it integrates a white-box, interpretable sampling policy grounded in Bayesian experiment design, enabling transparent and strategically guided exploration.
- EM-PTDM framework is backed by rigorous empirical evidence and ensures a monotonic refinement of the prior as new observations are gathered. This continual refinement directly enhances the accuracy of the scoring mechanism used for sampling, resulting in more precise and efficient active target discovery over time.
- We demonstrate the significance of each component in our proposed EM-PTDM method through extensive quantitative and qualitative ablation studies across a range of datasets, including species distribution modeling and remote sensing.

2 Problem Formulation

In this section, we present the details of our proposed Active Target Discovery (ATD) task setup. ATD involves actively uncovering one or more targets within a search area, represented as an (initially unobserved) region x divided into N grid cells, such that $x = (x^{(1)}, x^{(2)}, ..., x^{(N)})$. ATD operates under a query budget $\mathcal{B}$, representing the maximum number of measurements allowed. Each grid cell represents a sub-region and serves as a potential measurement location. A measurement of location i *provides feedback*, revealing both *the content* of a specific sub-region $x^{(i)}$ for the i-th grid cell, as well as yielding *an outcome* $y^{(i)} \in [0, 1]$, where $y^{(i)}$ is the fraction of pixels in the grid cell $x^{(i)}$ that belong to the target of interest. In each task configuration, the target's content is initially unknown and is revealed incrementally through observations from measurements. The goal is to identify as many grid cells belonging to the target as possible by strategically exploring the grid within the given budget $\mathcal{B}$. Denoting a query performed in step t as q_t, the overall task optimization objective is:

$$U(x^{\{(q_t)\}}; \{q_t\}) \equiv \max_{\{q_t\}} \sum_t y^{(q_t)} \text{ subject to } t \leq \mathcal{B} \tag{1}$$

With objective 1 in focus, we aim to develop a search policy that efficiently explores the search area ($x_{\text{test}} \sim X_{\text{test}}$) to identify as many target regions as possible within a measurement budget $\mathcal{B}$— **all without requiring access to any prior samples from the domain of interest.**

3 Methodology

A central challenge in ATD under an initially uninformative prior lies in achieving three distinct goals: (1) progressively improving the prior as sequential observations accumulate—since the prior fundamentally shapes the ATD process; (2) designing the prior to be swiftly adaptable, capable of rapidly incorporating knowledge from limited observations, as demanded by the task's high data-efficiency requirements, and (3) utilizing the sequentially improved prior model to strategically sample unobserved regions to maximize target discovery within limited sampling budget.

Progressively Improving the Prior model We begin by addressing the crucial challenge of enabling the prior to evolve progressively—ensuring that each newly acquired observation from the search space contributes meaningfully to refining the model's understanding and guiding sampling more effectively. To this end, we formulate our approach within the principled structure of the Expectation-Maximization (EM) framework, utilizing its iterative inference and optimization steps to systematically refine the prior with each new observation. Specifically, our goal is to learn the parameters of a prior model (e.g., a DDPM), denoted as $q_\phi(x, y)$, by maximizing the log-evidence $\log q_\phi(y)$ for a given observation y. Considering a distribution over observations $p(y)$, this corresponds to maximizing the expected log-evidence, which is equivalent to minimizing the KL divergence between the true distribution $p(y)$ and $q_\phi(y)$:

$$\phi^* = \arg\max_\phi \mathbb{E}_{p(y)}[\log q_\phi(y)] = \arg\min_\phi \mathrm{KL}(p(y) \| q_\phi(y)). \tag{2}$$

Next, we derive an iterative optimization procedure for ϕ that ensures monotonic improvement of expected log-evidence $\mathbb{E}_{p(y)}[\log q_\phi(y)]$. We present the result in the following Proposition.

https://doi.org/10.52202/085713-2163

Proposition 1. *Let ϕ_k denote the parameters of the current prior model, then improving this prior by maximizing the expected log-evidence $\mathbb{E}_{p(y)}[\log q_{\phi_k}(y)]$ with respect to ϕ_k is equivalent to maximizing the following surrogate maximization:*

$$\phi_{k+1} = \arg\max_{\phi} \mathbb{E}_{p(y)} \mathbb{E}_{q_{\phi_k}(x|y)} [\log q_{\phi}(x)] \tag{3}$$

such that $\mathbb{E}_{p(y)}[\log q_{\phi_{k+1}}(y)] \geq \mathbb{E}_{p(y)}[\log q_{\phi_k}(y)]$. Where $q_{\phi_k}(x|y)$ represents the approximate posterior under the current prior ϕ_k. Thus, this update ensures a strict improvement in the expected log-evidence. We present detailed proof in the Appendix.

As established in 1, optimizing the prior using the objective in Eqn. 3 ensures that each update step systematically refines the prior toward a more faithful approximation of the underlying data distribution. Specifically, $q_{\phi_{k+1}}(x)$ is more consistent with the distribution of observations $p(y)$ than $q_{\phi_k}(x)$. The update of the prior parameters ϕ_{k+1} follows a two-stage procedure: first, samples are drawn from the posterior distribution $q_{\phi_k}(x|y)$; next, the prior model $q_{\phi_{k+1}}(x)$ is trained to approximate this posterior by fitting to the generated samples. A central challenge in this framework is the accurate sampling from the posterior distribution $q_{\phi_k}(x|y)$, particularly during the early stages of active discovery when observations are inherently sparse. This data scarcity significantly limits the efficacy of traditional data-driven optimization approaches. Moreover, the challenge is intensified by the fact that repeated updates ($k > 1$) are both computationally and time-demanding. Consequently, it is imperative to design a prior model that is both swiftly adaptable to limited observation and lightweight in complexity. Large-scale diffusion models, while powerful, are typically data-intensive and ill-suited for such low-resource settings. Our objective is *to develop a model that can efficiently adapt under constrained data regimes without compromising the quality of posterior approximation.*

Swiftly Adaptable Prior model with Limited Observation To address this challenge, we draw inspiration from Doob's h-transform, which enables sampling conditioned on the observations gathered. This perspective offers a powerful lens through which we construct an adaptive mechanism capable of efficiently guiding the prior with minimal data. Specifically, conditional sampling from the posterior distribution $q_{\phi_k}(x|y)$ requires access to the posterior score $\nabla_x \log p_t(x \mid Y = y)$, conditioned on a given observation y. Since diffusion models are trained to approximate the score function of the marginal data distribution, i.e., $s_t^{\theta^*} \approx \nabla_x \log p_t(x)$, a pre-trained diffusion model can be adapted for conditional inference by applying Bayes' theorem:

$$\nabla_x \log p_t(x \mid Y = y) \approx s_t^{\theta^*} + \nabla_x \log p_t(Y = y \mid x). \tag{4}$$

The term $\nabla_x \log p_t(Y = y \mid x)$ is commonly referred to as the *guidance* term, as it effectively steers the reverse diffusion process toward samples consistent with the observation y. However, computing this guidance term is generally analytically intractable, which poses a significant challenge in performing conditional generation. Before delving into how Doob's h-transform can be harnessed to efficiently approximate the conditional guidance term, especially in data-scarce regimes, we first introduce the formal definition of Doob's h-transform. This foundational understanding will later enable us to leverage its powerful structure for efficient and adaptive conditional sampling.

Definition 1 (Doob's *h*-transform)**.** *[3, 4] Consider the following unconditional reverse-time SDE with f_t as the drift and σ_t as the diffusion coefficient.*

$$dX_t = \left(f_t(X_t) - \sigma_t^2 \nabla_{X_t} \ln p_t(X_t)\right) dt + \sigma_t dW_t, \quad X_T \sim P_T \sim \mathcal{N}(0, 1)$$

The corresponding conditional process $(X_t \mid X_0 \in B)$ evolves according to the SDE:

$$dH_t = \left[\overleftarrow{b_t(H_t)} - \sigma_t^2 \nabla_{H_t} \ln p_{0|t}(X_0 \in B \mid H_t)\right] dt + \sigma_t \overleftarrow{dW_t}, \quad H_T \sim \mathcal{P}_T \sim \mathcal{N}(0, 1) \tag{5}$$

where the backward drift $\overleftarrow{b_t(H_t)}$ is defined as: $\overleftarrow{b_t(H_t)} = f_t(H_t) - \sigma_t^2 \nabla_{H_t} \ln p_t(H_t)$, and the conditional law satisfies $\mathrm{Law}(H_s \mid H_t) = \overrightarrow{p}_{s|t,0}(x_s \mid x_t, x_0 \in B)$ with $\mathbb{P}(X_0 \in B) = 1$. We denote the conditional process as H_t and the corresponding unconditional process as X_t. Doob's h-transform illustrates that conditioning a diffusion process to reach a target set $X_0 \in B$ at terminal time T yields a new diffusion process governed by an adjusted drift term (as shown in blue). This conditional process is guaranteed to reach the desired event within a finite time T. The function $h(t, H_t) \triangleq \overleftarrow{p_{0|t}}(X_0 \in B \mid H_t)$ is known as the h-transform.

Definition 1 highlights a key insight: conditional sampling, informed by the observations collected up to the current step, hinges on the interplay between the unconditional score and a conditional score term, which is solely characterized by Doob's h-transform. Consequently, comparing Equation 4 and 5, the posterior score can be formulated as:

$$\nabla_x \log p_t(x \mid Y = y) \approx \underbrace{s_t^{\theta^*}(x)}_{\text{Permanent Memory}} + \underbrace{\nabla_x \log p_t(Y = y \mid x)}_{\approx h_t^{\zeta}(x,\, y)\text{: Transient Memory}} . \tag{6}$$

As shown in Equation 6, the posterior score naturally decomposes into two complementary components. The first term, derived from a pretrained diffusion model, serves as a form of *permanent memory*—encapsulating broad, generalizable patterns acquired from large-scale training data. While this offers a strong prior, it alone is insufficient in our setting, where no access to domain-specific prior samples is available. To address this limitation, we introduce the second term as a *transient memory* that modulates the pretrained dynamics in response to the limited observations collected during the active discovery process. This adaptive correction is governed by Doob's h-transform, parameterized by ζ, allowing the model to align with newly encountered observations in the data. In the Appendix, we provide an interpretation that frames the h-transform as a correction mechanism for the unconditional score, offering insights into how it adapts the pretrained dynamics to the conditional setting. Next, we detail how ζ can be efficiently learned to enable rapid adaptation in data-scarce scenarios. Interestingly, the parameter ζ can be optimized using an approach analogous to denoising score matching, as formalized in the following Lemma.

Lemma 1. *Consider the following stochastic differential equation for the conditional process:*

$$dH_t = \left[f_t(H_t) - \sigma_t^2 \left(\nabla_{H_t} \ln p_t(H_t) + h_t(H_t)\right)\right] dt + \sigma_t dW_t,$$

where $H_T \sim \mathcal{Q}_T^{f_t}[p(x_0 \mid y)] = \int p_{T|0}(x \mid x_0)\, p(x_0 \mid y)\, dx_0$. *The h-transform function* h_t^* *admits a denoising score-matching representation:* $h_t^* = \arg\min_{h_t \in \mathcal{H}} \mathcal{L}_{\mathrm{DSM}}(h_t)$, *where*

$$\mathcal{L}_{\mathrm{DSM}}(h_t) := \mathbb{E}_{\substack{X_0 \sim p(x_0|y) \\ t \sim \mathcal{U}(0,T);\, H_t \sim p_{t|0}(x_t|x_0)}} \left[\left\| \left(h_t(H_t) + \nabla_{H_t} \ln p_t(H_t)\right) - \nabla_{H_t} \ln p_{t|0}(H_t \mid X_0) \right\|^2 \right].$$

The proof is in the Appendix. Based on this Lemma, we optimize ζ as follows:

$$\min_{\zeta} \mathbb{E}_{(X_0, Y),\, \varepsilon,\, t} \left\| \left(h_t^{\zeta}(H_t, Y) + s_t^{\theta^*}(H_t) \right) - \varepsilon \right\|^2. \tag{7}$$

Let $H_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1 - \bar{\alpha}_t}\, \varepsilon$, where $(X_0, Y) \sim q_\phi(x_0, y)$ and $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. The function h_t^{ζ} denotes a neural network designed to approximate the h-transform. Crucially, the loss function in Equation 7 operates exclusively through forward evaluations of the pre-trained model and does not necessitate backpropagation through the fixed parameters θ^*. Consequently, the h-transform is tasked solely with learning the residual noise component, effectively serving as a corrective mechanism. This design choice permits the h-model to remain lightweight and shallow, thereby enabling rapid adaptation even in regimes with extremely limited observations. Following each observation, we draw posterior samples $X_0 \sim q_\phi(x_0 \mid y)$ using the posterior score composed of the frozen primary memory $s_t^{\theta^*}(x)$ and the transient memory $h_t^{\zeta}(x, y)$. These samples are then used to update the transient memory parameters ζ via the objective in 7, thereby ensuring that the resulting prior ϕ—consisting of both θ^* and the updated ζ—yields a strict improvement in the expected log-evidence, as formalized in 3. Formally, the following theorem establishes that updating the transient memory parameter ζ according to 7 leads to a monotonic improvement of the composite prior $\phi = (\zeta, \theta)$.

Theorem 1. *Let* $\phi = (\theta, \zeta)$ *denote the parameters of the prior, where* θ *corresponds to a fixed pre-trained model and* ζ *parameterizes the learnable h-transform module. Suppose* ζ *is updated according to Equation 7, yielding an updated parameter* ζ_{new} *and an updated prior* $\phi_{new} = (\theta, \zeta^{new})$. *Then, the marginal expected log-evidence satisfies the following:*

$$\mathbb{E}_{p(y)}[\log q_{\phi_{new}}(y)] \geq \mathbb{E}_{p(y)}[\log q_\phi(y)].$$

We present detailed proof in the Appendix. Next, we present our methodology for updating the h-model, designed to enable a seamless adaptation of the reverse diffusion dynamics.

https://doi.org/10.52202/085713-2163

Learning Dynamics of h-model A natural strategy is to iteratively update the h-model after each new observation, thereby refining the prior and leveraging this improved prior to guide subsequent rounds of ATD more effectively. However, in the early stages of the discovery process, the available observations are often too sparse to capture the structure of the underlying search space. Updating the h-model solely based on such limited data can lead to premature convergence to suboptimal regions in the parameter space, from which recovery becomes increasingly difficult—even as more data becomes available gradually as search progresses. This challenge is further intensified by the h-model's intended adaptability, as rapid updates may amplify the risk of overfitting to early, uninformative signals. To validate this hypothesis, we conducted a toy experiment (illustrated in Fig. 2), comparing two strategies for updating the h-model. In the first, the h-model is updated after every observation; in the second, updates occur only after accumulating several observations. We visualize the resulting posterior samples at different stages. When the h-model is updated after each individual observation, the posterior becomes noticeably noisy during the early phase of discovery, likely due to limited data. In contrast, postponing the updates results in more stable and coherent posterior samples that better reflect the gathered observations, enabling the diffusion process to incorporate new information more effectively at each stage. Motivated by this observation, we introduce a simple yet effective scheduling strategy for updating the h-model. The proposed scheduler (detailed in Appendix) adapts dynamically with the query budget—starting with infrequent updates and increasing in frequency as the search progresses. This design leverages the fact that updates become more reliable over time due to the increasing accumulation of informative observations.

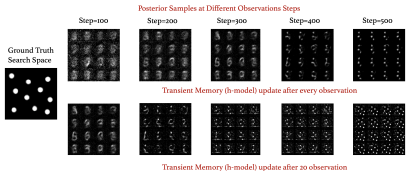


Figure 2: h-model adaptability dynamics. We use a Diffusion model trained on MNIST as permanent memory (16 samples per Fig).

Theorem 2. *Let $\phi_t = (\theta, \zeta_t)$ be the prior parameters at observation step t. Suppose ζ_t is updated following Equation 7, yielding $\phi_{t+k} = (\theta, \zeta_{t+k})$ after k additional observations. Then the expected log-evidence improves monotonically:*

$$\mathbb{E}_{p(y_{t+k})}[\log q_{\phi_{t+k}}(y)] \geq \mathbb{E}_{p(y_{t+k})}[\log q_{\phi_t}(y)].$$

where y_{t+k} represents the set of observations gathered uptill time $t+k$, and $k \geq 1$.

We present detailed proof in the Appendix. Next, we introduce a sampling strategy that capitalizes on the updated prior developed thus far, guiding effective and budget-conscious ATD.

Sampling Strategy for ATD Utilizing the Updated Prior An effective sampling strategy that strikes the right balance between exploration and exploitation is crucial for solving ATD. We first describe how we approach the exploration using the observations collected thus far. To achieve this, we adopt a maximum-entropy strategy, selecting the measurement q_t^{exp} at the t-th query step as

$$q_t^{\text{exp}} = \arg\max_{q_t} [H(\hat{x}_t|Q_t, y_{t-1})] = \arg\max_{q_t} -\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \tag{8}$$

Here, H represents entropy, $\hat{x}_t$ denotes samples from the approximate posterior $q_{\phi_{t-1}}(x \mid y_{t-1})$. Note that we utilize the final updated prior $\phi_{t-1} = (\theta, \zeta_{t-1})$ to sample from the posterior. Q_t represents the set of locations queried up to time t, and y_{t-1} represents the set of observations up to time $t-1$. As defined in Equation 8, we select the query location that corresponds to the maximum entropy. To compute $\log p(\hat{x}_t|Q_t, y_{t-1})$, we begin by drawing P samples from the posterior $q_{\phi_{t-1}}(x \mid y_{t-1})$, denoting the i-th sample as $\hat{x}_t^i$. We then approximate $\log p(\hat{x}_t|Q_t, y_{t-1})$ using a mixture of Gaussians: $p(\hat{x}_t|Q_t, y_{t-1}) = \sum_{i=0}^{P} \alpha_i \mathcal{N}(\hat{x}_t^i, \sigma_x^2 I)$. In the following Theorem, we derive a result that allows us to compute q_t^{exp} utilizing the expression of $p(\hat{x}_t|Q_t, y_{t-1})$.

Theorem 3. *Assuming k represents the set of possible measurement locations at step t, and that all samples from the posterior $q_{\phi_{t-1}}(x \mid y_{t-1})$ have equal weights ($\alpha_i = \alpha_j$, $\forall i, j$), then,*

$$q_t^{exp} = \arg\max_{q_t} \left[\sum_{i=0}^{P} \log \sum_{j=0}^{P} \exp\left(\frac{\sum_{q_t \in k} ([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2} \right) \right],$$

where $[.]_{q_t}$ selects element of $[.]$ indexed by q_t. Detailed proof is in the Appendix.

The implication is that the optimal next observation location (q_t^{exp}) lies in the region of the search space where the predicted semantics ($\hat{x}_t^{(i)}$) show the highest disagreement among the *posterior samples* ($i \in 0, \ldots, P$). This disagreement is quantified by a metric, denoted as $\text{expl}_\phi^{\text{score}}(q_t)$, which enables us to rank candidate observation locations and thereby guide exploration more effectively. In scenarios where the observation space is composed of pixels, each $\hat{x}_t^{(i)}$ represents a full image conditioned on the observed pixels y_{t-1}.

$$\text{expl}_\phi^{\text{score}}(q_t) = \left[\sum_{i=0}^{P} \sum_{j=0}^{P} \frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2} \right] \tag{9}$$

We now describe how the collected observations are used to guide effective exploitation. Specifically, we: (i) employ a *reward model* (r_η), parameterized by η and incrementally trained on supervised data gathered from observations, to predict whether an observed location corresponds to the target of interest; and (ii) estimate the *expected log-likelihood score* at a location q_t, denoted as $\text{likeli}_\phi^{\text{score}}(q_t)$, defined as $\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t \mid Q_t, y_{t-1})]_{q_t}$. This score is key to prioritizing locations during exploitation. The following proposition provides a closed-form expression for computing $\text{likeli}_\phi^{\text{score}}(q_t)$.

Proposition 2. *The expected log-likelihood score at q_t can be expressed as follows:*

$$\text{likeli}_\phi^{\text{score}}(q_t) = \sum_{i=0}^{P} \sum_{j=0}^{P} \exp\left\{ -\frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2} \right\}$$

The derivation of Proposition 2 is provided in Appendix. The term $\text{likeli}_\phi^{\text{score}}(q_t)$ plays a central role in computing the exploitation score at measurement location q_t. Specifically, the exploitation score, $\text{exploit}_{(\phi,\eta)}^{\text{score}}(q_t)$, is defined as the *reward-weighted expected log-likelihood*, as shown below:

$$\text{exploit}_{(\phi,\eta)}^{\text{score}}(q_t) = \underbrace{\text{likeli}_\phi^{\text{score}}(q_t)}_{\textit{Expected log-likelihood}} \times \underbrace{\sum_{i=0}^{P} r_\eta([\hat{x}_t^{(i)}]_{q_t})}_{\textit{reward}} \tag{10}$$

As per Eqn. 10, a measurement location q_t is favored for exploitation if it meets two criteria: (1) the predicted content at q_t is highly likely to correspond to the target of interest, as indicated by the reward model (r_η); and (2) the estimated content at q_t, denoted as $[\hat{x}_t^{(i)}]_{q_t}$, exhibit strong agreement across the *posterior samples*, suggesting high predictability about the content at that location. Note that the reward model parameters (η) are randomly initialized and progressively updated using *binary cross-entropy loss* as new supervised data becomes available after each observation. While the reward model is initially unrefined—resulting in less reliable exploitation scores—this is not a limitation, as our sampling strategy is designed to emphasize exploration during the early stages of the discovery process. We now bring together the core components to formulate the proposed sampling strategy for efficient active target discovery. At each time step t, potential measurement locations (q_t) are ranked by jointly considering both exploration and exploitation scores, as below:

$$\text{Score}_{(\phi,\eta)}(q_t) = \alpha(\mathcal{B}) \cdot \text{expl}_\phi^{\text{score}}(q_t) + (1 - \alpha(\mathcal{B})) \cdot \text{exploit}_{(\phi,\eta)}^{\text{score}}(q_t) \tag{11}$$

Here, $\alpha(\mathcal{B})$ is a budget-dependent function that enables a dynamic and adaptive balance between exploration and exploitation. After computing the combined score for each candidate location q_t (as defined in Eqn. 11), the location with the highest score is selected for sampling. The choice of $\alpha(\mathcal{B})$ is problem-specific and can be tailored accordingly. A straightforward and effective option is to define it as a linear function of the remaining budget, such as $\alpha(\mathcal{B}) = \frac{\mathcal{B}-t}{\mathcal{B}+t}$. We empirically show that this simple formulation performs well across diverse application domains. The complete training and sampling algorithm is detailed in the Appendix, with a visual overview provided in the Appendix. Finally, we demonstrate that the improvement of the prior ϕ in turn leads to increasingly accurate estimates of the score defined in Equation 11 as formally stated in Theorem 4.

Theorem 4. *Let ϕ and ϕ_{new} denote the prior parameters before and after updating the h-model using the objective in Equation (7). Then the following relation holds,*

$$||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi,\eta)}(q_t)|| \geq ||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi_{new},\tilde{\eta})}(q_t)||$$

Here, $\text{Score}^*_{(\phi^*,\eta^*)}(q_t)$ represents the true score estimate computed using the optimal prior ϕ^* and optimal reward model η^*, and $\tilde{\eta}$ denotes the updated reward model parameter.

https://doi.org/10.52202/085713-2163

We present the proof in the Appendix. When tackling consecutive tasks from the same domain, we update the long-term memory using posterior samples gathered at the end of each discovery process, promoting faster and more efficient adaptation to subsequent tasks. We present a pictorial illustration of our proposed EM-PTDM approach in Figure 3. Next, we present our detailed experimental results

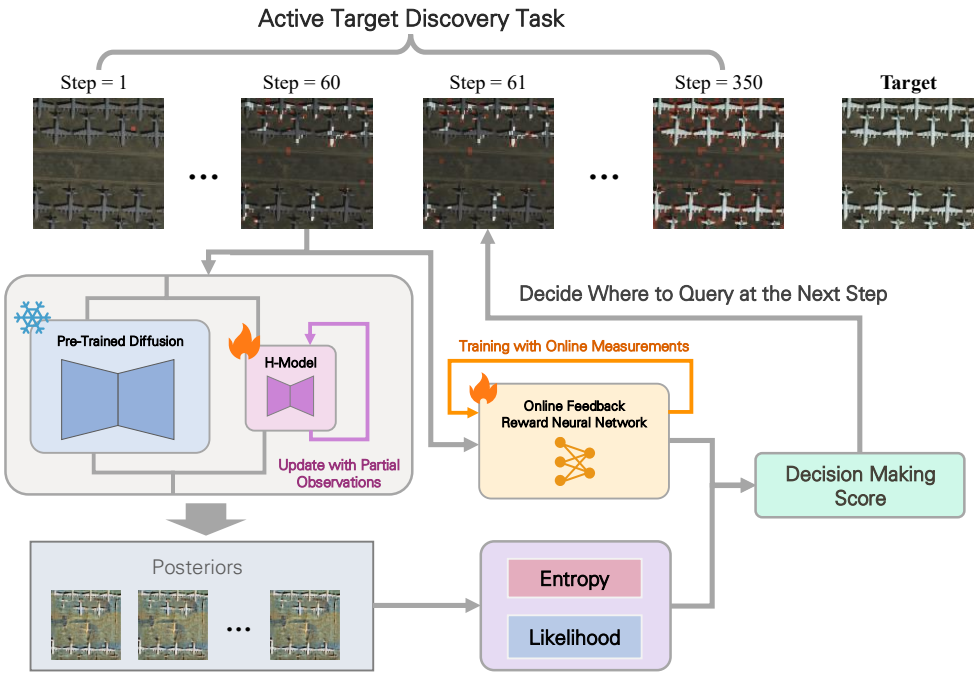


Figure 3: An Overview of EM-PTDM Framework.

to validate the efficacy of our proposed method.

4 Empirical Analysis

Evaluation metrics. Since ATD seeks to maximize the identification of measurement locations containing the target of interest, we assess performance using the *success rate (SR) of selecting measurement locations that belong to the target* during exploration in partially observable environments. Therefore, SR is defined as: $\text{SR} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{\min\{\mathcal{B},U_i\}}\sum_{t=1}^{\mathcal{B}} y_i^{(q_t)}$; where, L = number of tasks. Here, U_i denotes the maximum number of measurement locations containing the target in the i-th search task. We evaluate EM-PTDM and the baselines across different measurement budgets $\mathcal{B} \in \{150, 200, 250, 300, 350\}$ for various target categories and application domains.

Baselines. We compare our proposed EM-PTDM policy to the following baselines:

- *Random Search* (RS), in which unexplored measurement locations are selected uniformly at random.
- DiffATD [1] leverages a pre-trained diffusion model to guide sampling based on partially observed data, efficiently discovering target regions within a fixed budget.
- *Greedy-Adaptive* (GA) selects q_t with the highest $\text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t)$ among unexplored locations and updates reward model r_η via binary cross-entropy after each observation.

Active Discovery of Unknown Species from Known Species Distribution We begin our evaluation of EM-PTDM in a setting where the permanent memory is trained to approximate the distribution of species from iNaturalist [5]. The distribution is formed by dividing a large geospatial region into equal-sized grids and counting target species occurrences on each grid, as detailed in Appendix. The goal is to actively discover Coccinella Septempunctata (CS) under a strict sampling budget using a prior trained on the distribution of Gladicosa and Gonioctena (GG). The results presented in Table 1, suggest that EM-PTDM significantly outperforms the baseline in terms of SR, highlighting the capability of the proposed framework in ATD under an uninformative prior. We further visualize (in 4) how EM-PTDM, starting from an uninformative prior, rapidly approximates the true distribution with only a few observations, demonstrating its swift adaptability and efficiency in data-scarce scenarios.

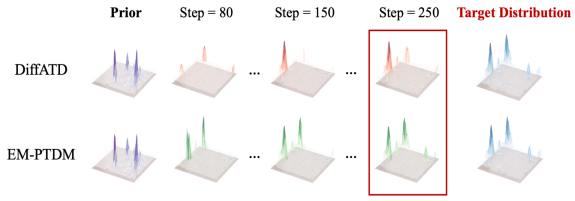


Figure 4: Active Discovery of CS species.

Table 1: *SR* Comparison with Baselines.

Active Discovery of CS Species with Species GG as Prior.			
Method	$\mathcal{B} = 150$	$\mathcal{B} = 200$	$\mathcal{B} = 250$
RS	0.1624	0.2327	0.2775
DiffATD	0.3420	0.4365	0.4808
GA	0.4061	0.5067	0.5567
EM-PTDM	**0.4983**	**0.6495**	**0.6989**

Active Discovery of Unknown Overhead Objects from Ground-view Prior We further assess EM-PTDM in a challenging cross-view setting, where the task is to actively discover novel objects from DOTA overhead imagery [6], leveraging an uninformative prior learned solely from ground-level images of semantically disjoint classes from Imagenet [7]. The results are presented in Table 2. We observe that EM-PTDM consistently outperforms the baseline, with the performance gap widening as the search budget increases, highlighting the strength of its cumulatively refined prior in driving efficient exploration and informed target discovery. Additionally, we compare the exploration dynamics of EM-PTDM and DiffATD at two stages of the active discovery process. As shown in Figure 5, while both models initially struggle due to uninformative priors, **EM-PTDM rapidly adapts through Doob's h-transform, enabling more efficient target discovery as observations accumulate.**

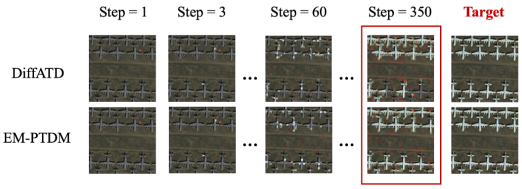


Figure 5: Active Discovery of Overhead objects.

Table 2: *SR* Comparison with Baselines.

Active Discovery of Overhead Objects with ImageNet as Prior.			
Method	$\mathcal{B} = 250$	$\mathcal{B} = 300$	$\mathcal{B} = 350$
RS	0.2325	0.2852	0.3207
DiffATD	0.5143	0.6391	0.7348
GA	0.4784	0.5659	0.6562
EM-PTDM	**0.5620**	**0.7013**	**0.8256**

Enhancing Permanent Memory with Domain Cues for Improved In-Domain Target Discovery In real-world settings, it's common to encounter scenarios where ATD tasks must be addressed consecutively within similar domains, requiring models that can retain and transfer knowledge effectively. To support efficient adaptation in sequential ATD tasks within the same domain, we update the permanent memory—specifically, the pretrained diffusion model (s^{θ^*})—using posterior samples generated at the final step of the previous task, conditioned on all collected observations. This update allows the permanent memory to gain domain-specific structural knowledge, thereby reducing the divergence between prior and posterior distributions in subsequent tasks. As a result, the learning task of doob's h-model becomes easier as the correction needed for rapid adaptation diminishes, enabling the model to explore and discover new targets more effectively with fewer observations. We assess the impact of updating permanent memory after each task by comparing it against a static-memory baseline. For this comparison, we consider the task of ATD of overhead objects with ImageNet as the prior. As reported in 3, the results reveal a marked improvement in the SR when updates are applied, underscoring the value of **accumulating domain-specific knowledge in permanent memory to accelerate adaptation and enhance target discovery in related tasks.**

Table 3: Effect of Permanent Memory

ATD of Overhead Objects with ImageNet as Prior.			
s^{θ^*} update	$\mathcal{B} = 250$	$\mathcal{B} = 300$	$\mathcal{B} = 350$
No	0.5620	0.7013	0.8256
Yes	**0.5859**	**0.7194**	**0.8461**

The Role of Transient Memory in Rapid Adaptation to Novel ATD Tasks To assess the impact of transient memory, we compare posterior estimates generated with and without the proposed transient memory mechanism, starting from an uninformative prior. As illustrated in Fig. 7, incorporating transient memory via the h-model enables rapid adaptation from sparse early observations, resulting in posterior estimates that closely approximate the ground truth and facilitate more efficient target discovery. In contrast, the baseline, which relies solely on permanent memory, yields an incorrect posterior, resulting in limited exploration capability in early stages, underscoring the critical role of the h-model in accelerating effective exploration. We quantitatively assess the role of the h-model by measuring $L2$ distance-based semantic dissimilarity between the posterior and ground-truth targets. Results (6) show that **incorporating the h-model significantly accelerates convergence to the true posterior**, as dissimilarity drops much faster compared to using the permanent memory alone, demonstrating the h-model's effectiveness in rapidly correcting the prior with minimal observations.

https://doi.org/10.52202/085713-2163

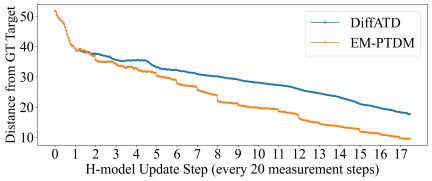
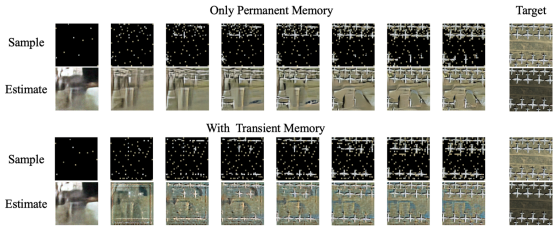


Figure 6: Effectiveness of Transient Memory.

Figure 7: Posterior Dynamics with and without h-model.



An Important Observation: ATD is not just About Reconstruction One might ask: if the primary goal is for the posterior samples to accurately reconstruct the search space, isn't that sufficient for efficient target discovery? Interestingly, in our setting, a precise reconstruction of the entire search space is not strictly necessary, as long as the model effectively identifies and reconstructs the target regions, efficient discovery can still be achieved. To validate this hypothesis, we conduct an experiment and visualize the posterior at intermediate stages of active target discovery. We compare posterior samples from EM-PTDM and DiffATD using a representative task where EM-PTDM significantly outperforms DiffATD, allowing us to understand how the posterior contributes to improved target discovery. We present the visualization in the Figure 8. Our observations reveal that, **while EM-PTDM's posterior samples exhibit lower overall reconstruction quality compared to those from DiffATD, they more effectively focus on target-rich regions** (For example, see the highlighted Red box in the Figure 8), leading to significantly improved target discovery performance. This highlights a key insight: successful target discovery relies more on accurately modeling the regions of interest than on reconstructing the entire search space.

DiffATD Posterior at Intermediate Active Target Discovery Phase

EM-PTDM Posterior at Intermediate Active Target Discovery Phase

Ground Truth Posterior

Figure 8: ATD is about Discovering Targets, NOT just Reconstructing the Search Space.

5 Related Work

Several RL-based methods [8, 9, 10, 11] have been developed for ATD; however, they typically assume full observability of the search space and require access to large-scale, pre-labeled datasets to learn efficient exploration strategies. Training-free approaches inspired by Bayesian decision theory [12, 13, 14] offer an appealing alternative, yet their dependence on complete observability limits their effectiveness in partially observable settings. More recently, a new class of methods [15, 16, 17] has emerged, explicitly addressing active discovery under partial observability. Nevertheless, these approaches still face a major challenge—their heavy reliance on extensive annotated datasets to reach optimal performance. DiffATD [1] advances active target discovery by enabling efficient exploration in partially observable environments without requiring task-specific labeled data. However, it still depends on domain-specific prior samples to learn an effective prior model, limiting its applicability in domains where such data is unavailable. To address this critical gap, we propose EM-PTDM, a novel framework that enables active target discovery without relying on any prior domain samples.

6 Conclusion

We propose EM-PTDM, a principled and generalizable framework for ATD in partially observable environments, remarkably operating without any task-specific prior data, significantly broadening its applicability in diverse domains. EM-PTDM bridges solid theoretical foundations with compelling empirical performance across diverse domains, ranging from rare species discovery to remote sensing.

https://doi.org/10.52202/085713-2163

Acknowledgement This research was partially supported by the National Science Foundation (CCF-2403758, IIS-2214141), Army Research Office (W911NF2510059), Office of Naval Research (N000142412663), and Amazon.

References

[1] Anindya Sarkar, Binglin Ji, and Yevgeniy Vorobeychik. Online feedback efficient active target discovery in partially observable environments, 2025.

[2] Dharshan Kumaran. What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience*, 6:157, 2012.

[3] L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales: Volume 2, Itô Calculus*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2000.

[4] Jeremy Heng, Valentin De Bortoli, Arnaud Doucet, and James Thornton. Simulating diffusion bridges with score matching, 2021. arXiv preprint arXiv:2111.07243.

[5] *iNaturalist*. https://www.inaturalist.org/.

[6] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12345–12354, 2020.

[9] Anindya Sarkar, Michael Lanier, Scott Alfeld, Jiarui Feng, Roman Garnett, Nathan Jacobs, and Yevgeniy Vorobeychik. A visual active search framework for geospatial exploration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8316–8325, 2024.

[10] Anindya Sarkar, Nathan Jacobs, and Yevgeniy Vorobeychik. A partially-supervised reinforcement learning framework for visual active search. *Advances in Neural Information Processing Systems*, 36:12245–12270, 2023.

[11] Quan Nguyen, Anindya Sarkar, and Roman Garnett. Amortized nonmyopic active search via deep imitation learning. *arXiv preprint arXiv:2405.15031*, 2024.

[12] Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. Bayesian optimal active search and surveying. *arXiv preprint arXiv:1206.6406*, 2012.

[13] Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search. In *International Conference on Machine Learning*, pages 1714–1723. PMLR, 2017.

[14] Shali Jiang, Roman Garnett, and Benjamin Moseley. Cost effective active search. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Samrudhdhi B Rangrej, Chetan L Srinidhi, and James J Clark. Consistency driven sequential transformers attention model for partially observable scenes. *arXiv preprint arXiv:2204.00656*, 2022.

[16] Aleksis Pirinen, Anton Samuelsson, John Backsund, and Kalle Aström. Aerial view goal localization with reinforcement learning. *arXiv preprint arXiv:2209.03694*, 2022.

 https://doi.org/10.52202/085713-2163

[17] Anindya Sarkar, Srikumar Sastry, Aleksis Pirinen, Chongjie Zhang, Nathan Jacobs, and Yevgeniy Vorobeychik. Gomaa-geo: Goal modality agnostic active geo-localization. *arXiv preprint arXiv:2406.01917*, 2024.

[18] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[19] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[20] Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1998–2007. PMLR, 2018.

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: section 3 discusses our algorithm and section 4 includes the results from our experiments, both of which claims in the abstract and introduction reflect.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: Our method outperformed existing methods on a newly proposed task.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/085713-2163

Justification: We provide complete and detailed theoretical proof in section 3 and Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discuss the experiment details in main paper and in the Appendix. We also provide the code for reference.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use public available datasets for all our experiments and we provide the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We discuss these details in the main paper and in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide these details in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/085713-2163

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide these details in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics. Our research does not involve human participants, and our data were obtained in a way that respects the corresponding licenses, and processed to preserve anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impacts of our work in the Appendix, which also includes considerations that should made even when our method is used as intended.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We judge that our work poses no immediate risk of misuse of our released data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All works whose data are used by us are appropriately credited and the corresponding licenses were respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

 https://doi.org/10.52202/085713-2163

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The README file in our code includes documentation on how to use the data and the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve human participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve human participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our method doesn't involve LLM.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-2163

Appendix: Active Target Discovery under Uninformative Prior: The Power of Permanent and Transient Memory

Overview of the Contents

A Proofs of Theoretical Results

A.1 Proof of Proposition 1

Proof.

$$\theta^* = \arg\max_\theta \mathbb{E}_{p(y)} \left[\log q_\theta(y)\right] \tag{12}$$

$$= \arg\min_\theta \mathrm{KL}\left(p(y) \,\|\, q_\theta(y)\right) \tag{13}$$

The core principle of the EM algorithm is that, for any two parameter sets θ_a and θ_b, the following identity holds:

$$\log \frac{q_{\theta_a}(y)}{q_{\theta_b}(y)} = \log \frac{q_{\theta_a}(x,y)}{q_{\theta_b}(x,y)} \cdot \frac{q_{\theta_b}(x \mid y)}{q_{\theta_a}(x \mid y)} \tag{14}$$

$$= \mathbb{E}_{q_{\theta_b}(x|y)} \left[\log \frac{q_{\theta_a}(x,y)}{q_{\theta_b}(x,y)}\right] + \mathrm{KL}\left(q_{\theta_b}(x \mid y) \,\|\, q_{\theta_a}(x \mid y)\right) \tag{15}$$

$$\geq \mathbb{E}_{q_{\theta_b}(x|y)} \left[\log q_{\theta_a}(x,y) - \log q_{\theta_b}(x,y)\right] \tag{16}$$

This inequality remains valid when taking the expectation over $p(y)$. Consequently, starting from an initial parameter setting θ_0, the EM update rule can be expressed as:

$$\theta_{k+1} = \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x|y)} \left[\log q_\theta(x,y) - \log q_{\theta_k}(x,y)\right] \tag{17}$$

$$= \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x|y)} \left[\log q_\theta(x,y)\right] \tag{18}$$

In the empirical Bayes setting, the forward model $p(y \mid x)$ is known and only the parameters of the prior q_θ(x) should be optimized. In this case, Eq. 18 becomes:

$$\theta_{k+1} = \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x|y)} \left[\log q_\theta(x) + \log p(y \mid x)\right] \tag{19}$$

$$\theta_{k+1} = \arg\max_\theta \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x|y)} \left[\log q_\theta(x)\right] \tag{20}$$

□

A.2 Proof of Lemma 1

Proof. We can express the conditional score as follows:

$$\nabla_{x_t} \ln p_t(x \mid y) = \int \nabla_{x_t} \ln \overrightarrow{p}_{t|0}(x_t \mid x_0)\, \overleftarrow{p}_{0|t}(x_0 \mid x_t, y)\, dx_0 \tag{21}$$

The minimizer is obtained by exploiting the property that the conditional expectation yields the optimal solution under the mean squared error criterion:

$$\mathbb{E}\left[\nabla_{H_t} \ln p_{t|0}(H_t \mid X_0) - \nabla_{H_t} \ln p_t(H_t) \,\middle|\, Y = y, H_t = x\right] \tag{22}$$

$$h_t^*(x,y) = \left(\int \left[\nabla_x \ln p_{t|0}(x \mid x_0) - \nabla_x \ln p_t(x)\right] p_{0|t}(x_0 \mid H_t = x, Y = y)\, dx_0\right) \tag{23}$$

$$= \int \nabla_x \ln p_{t|0}(x \mid x_0)\, p_{0|t}(x_0 \mid H_t = x, Y = y)\, dx_0 - \nabla_x \ln p_t(x) \tag{24}$$

$$= \nabla_x \ln p_t(x \mid y) - \nabla_x \ln p_t(x) \tag{25}$$

$$= \nabla_x \ln p_t(y \mid x) \tag{26}$$

□

https://doi.org/10.52202/085713-2163

A.3 Proof of Theorem 1

Proof.

$$\arg\max_{\phi} \mathbb{E}_{p(y)}[\log q_\phi(y)] \tag{27}$$

Utilizing the result of Theorem 1, we can express the above expression as follows:

$$\phi^{\text{new}} = \arg\max_{\phi} \mathbb{E}_{p(y)} \mathbb{E}_{q_\phi(x|y)} [\log q_\phi(x)] \tag{28}$$

Maximizing the above objective involves computing $\nabla_x \log q_\phi(x)$. Following the approach in denoising score matching, the above optimization can be equivalently reformulated as:

$$\zeta^{\text{new}} = \min_{\zeta} \mathbb{E}_{(X_0,Y),\,\varepsilon,\,t} \left\| \left(h_t^\zeta(x_t, Y) + s_t^{\theta^*}(x_t) \right) - \varepsilon \right\|^2 \tag{29}$$

Let $x_t = \sqrt{\bar{\alpha}_t}\, X_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$, where $X_0 \sim q_\phi(x_0 \mid y), Y \sim y$ and $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. Also, $\phi^{\text{new}} = (\theta^*, \zeta^{\text{new}})$.

Thus, optimizing the objective in Equation 29, is equivalent to one step of EM. Hence, guarantees the improvement of expected log-evidence. As a result, we can write:

$$\mathbb{E}_{p(y)}[\log q_{\phi_{\text{new}}}(y)] \geq \mathbb{E}_{p(y)}[\log q_\phi(y)].$$

□

A.4 Proof of Theorem 2

Proof.

$$\mathbb{E}_{p(y_{t+k})}[\log q_{\phi_{t+k}}(y)] \tag{30}$$

Assuming the observations collected during the active discovery process are independent, we can decompose the above expression into two parts as follows:

$$= \underbrace{\mathbb{E}_{p(y_t)}[\log q_{\phi_{t+k}}(y)]}_{\textit{part1}} \times \underbrace{\mathbb{E}_{p(y_{t+k:(t+1)})}[\log q_{\phi_{t+k}}(y)]}_{\textit{part2}} \tag{31}$$

The set of observations gathered from time step $t+1$ to $t+k$ is denoted as $y_{t+k:(t+1)}$.

For *Part1*, we can write,

$$\mathbb{E}_{p(y_t)}[\log q_{\phi_{t+k}}(y)] \geq \mathbb{E}_{p(y_t)}[\log q_{\phi_t}(y)] \quad \text{(Following Theorem 2)} \tag{32}$$

For *Part2*, we can write,

$$\mathbb{E}_{p(y_{t+k:(t+1)})}[\log q_{\phi_{t+k}}(y)] \geq \mathbb{E}_{p(y_{t+k:(t+1)})}[\log q_{\phi_t}(y)] \quad \text{(By Definition)} \tag{33}$$

The above relation holds because, unlike ϕ_k, the model ϕ_{t+k} is trained on posterior samples explicitly conditioned on the observations $y_{t+k:(t+1)}$. As a result, when ϕ_{t+k} is optimally trained, the left-hand side of Equation 33 reaches its optimal value. In contrast, since ϕ_k is never exposed to $y_{t+k:(t+1)}$ during training, it cannot attain the optimal value of the left-hand side in Equation 33.

Finally, combining the results of Equation 32 and 33, we can write:

$$\mathbb{E}_{p(y_{t+k})}[\log q_{\phi_{t+k}}(y)] \geq \mathbb{E}_{p(y_{t+k})}[\log q_{\phi_t}(y)]$$

□

A.5 Proof of Theorem 3

Proof. We start with the definition of q_t^{exp} as follows:

$$q_t^{\text{exp}} = \arg\max_{q_t} -\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \quad \text{where} \quad p(\hat{x}_t|Q_t, y_{t-1}) = \sum_{i=0}^{P} \alpha_i \mathcal{N}(\hat{x}_t^i, \sigma_x^2 I)$$

According to [18],

$$q_t^{\text{exp}} \propto \sum_{i=0}^{P} \alpha_i \log \sum_{j=0}^{P} \alpha_j \exp\left\{\frac{||\hat{x}_t^{(i)} - \hat{x}_t^{(j)}||_2^2}{2\sigma_x^2}\right\}$$

We can write,

$$q_t^{\text{exp}} = \arg\max_{q_t} \sum_{i=0}^{P} \alpha_i \log \sum_{j=0}^{P} \alpha_j \exp\left\{\frac{||\hat{x}_t^{(i)} - \hat{x}_t^{(j)}||_2^2}{2\sigma_x^2}\right\}$$

$$q_t^{\text{exp}} = \arg\max_{q_t} \sum_{i,j} \log\left(\exp\left(\frac{\|\hat{x}_t^{(i)} - \hat{x}_t^{(j)}\|^2}{2\sigma_x^2}\right)\right) \text{(By assuming, } \alpha_i = \alpha_j, \forall i, j)$$

$$q_t^{\text{exp}} = \arg\max_{q_t} \sum_{i,j} \log\left(\exp\left(\frac{\sum_{a \in Q_t}([\hat{x}_t^{(i)}]_a - [\hat{x}_t^{(j)}]_a)^2}{2\sigma_x^2}\right)\right)$$

We decompose it into two parts: one representing the set of potential measurement locations at the query step t, and the other corresponding to the set of locations already selected in Q_{t-1}. Hence, we can express it as follows:

$$q_t^{\text{exp}} = \arg\max_{q_t} \sum_{i,j} \log\left(\exp\left(\frac{\sum_{q_t \in k}([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2 + \sum_{r \in Q_{t-1}}([\hat{x}_t^{(i)}]_r - [\hat{x}_t^{(j)}]_r)^2}{2\sigma_x^2}\right)\right)$$

$$q_t^{\text{exp}} = \arg\max_{q_t} \sum_{i,j} \log\left(\prod_{q_t \in k} \exp\left(\frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}\right) \prod_{r \in Q_{t-1}} \exp\left(\frac{([\hat{x}_t^{(i)}]_r - [\hat{x}_t^{(j)}]_r)^2}{2\sigma_x^2}\right)\right)$$

$$q_t^{\text{exp}} \propto \arg\max_{q_t} \sum_{i,j} \left(\sum_{q_t \in k} \frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2} + \underbrace{\sum_{r \in Q_{t-1}} \frac{([\hat{x}_t^{(i)}]_r - [\hat{x}_t^{(j)}]_r)^2}{2\sigma_x^2}}_{\text{We can ignore as it doesn't depend on the choice of measurement location at time t.}}\right)$$

$$q_t^{\text{exp}} \propto \arg\max_{q_t} \sum_{i,j} \sum_{q_t \in k} \frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}.$$

$$q_t^{\text{exp}} = \arg\max_{q_t} \left[\sum_{i=0}^{P} \log \sum_{j=0}^{P} \exp\left(\frac{\sum_{q_t \in k}([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}\right)\right]$$

□

https://doi.org/10.52202/085713-2163

A.6 Proof of Theorem 4

Proof. Using Equation 12, we can decompose the score as follows:

$$||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi,\eta)}(q_t)|| = \alpha(\mathcal{B}) \underbrace{\left(\text{expl}^{\text{score}}_{\phi^*}(q_t) - \text{expl}^{\text{score}}_{\phi}(q_t)\right)}_{A}$$
$$+ (1-\alpha(\mathcal{B})) \underbrace{\left(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t)\right)}_{B} \tag{34}$$

By following a similar decomposition, we can write,

$$||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi_{\text{new}},\tilde{\eta})}(q_t)|| = \alpha(\mathcal{B}) \underbrace{\left(\text{expl}^{\text{score}}_{\phi^*}(q_t) - \text{expl}^{\text{score}}_{\phi_{\text{new}}}(q_t)\right)}_{C}$$
$$+ (1-\alpha(\mathcal{B})) \underbrace{\left(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi_{\text{new}},\tilde{\eta})}(q_t)\right)}_{D} \tag{35}$$

We can write,

$$\underbrace{\left(\text{expl}^{\text{score}}_{\phi^*}(q_t) - \text{expl}^{\text{score}}_{\phi}(q_t)\right)}_{A} \geq \underbrace{\left(\text{expl}^{\text{score}}_{\phi^*}(q_t) - \text{expl}^{\text{score}}_{\phi_{\text{new}}}(q_t)\right)}_{C}$$
(Follows from Proposition 1 and Theorem 1) (36)

The above relation holds as ϕ_{new} is obtained by applying one iteration of EM, thus guaranteeing improvement from the previous iteration prior (ϕ). Hence, ensure a more accurate exploration score estimation compared to the prior of previous iteration.

Next, we will compare the terms B and D and show that

$$\underbrace{\left(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t)\right)}_{B} \geq \underbrace{\left(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi_{\text{new}},\tilde{\eta})}(q_t)\right)}_{D}$$

Utilizing the expression in 11, we compare the exploit scores computed via different parameterizations of the prior as follows:

$$\text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t) = \underbrace{\text{likeli}^{\text{score}}_{\phi}(q_t)}_{\textit{Expected log-likelihood}} \times \underbrace{\sum_{i=0}^{P} r_\eta([\hat{x}_t^{(i)}]_{q_t})}_{\textit{reward}}$$

We can rewrite the above expression as follows:

$$\text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t) = \text{likeli}^{\text{score}}_{\phi}(q_t) + \mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi(x|y)}}[r_\eta([\hat{x}_t^{(i)}]_{q_t})]$$

Following exactly similar steps, we can also write

$$\text{exploit}^{\text{score}}_{(\phi_{\text{new}},\tilde{\eta})}(q_t) = \text{likeli}^{\text{score}}_{\phi_{\text{new}}}(q_t) + \mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi_{\text{new}}(x|y)}}[r_{\tilde{\eta}}([\hat{x}_t^{(i)}]_{q_t})]$$

Following the same reasoning as in 37, we can write

$$\left(\text{likeli}^{\text{score}}_{\phi^*}(q_t) - \text{likeli}^{\text{score}}_{\phi}(q_t)\right) \geq \left(\text{likeli}^{\text{score}}_{\phi^*}(q_t) - \text{likeli}^{\text{score}}_{\phi_{\text{new}}}(q_t)\right)$$
(Follows from Proposition 1 and Theorem 1) (37)

Now, note that at the beginning of the active target discovery process, $r_{\tilde{\eta}}([\hat{x}_t^{(i)}]_{q_0}) = r_\eta([\hat{x}_t^{(i)}]_{q_0})$.

As ϕ_{new} is a improved parameterization of the previous iteration prior ϕ, thus posterior samples from $q_{\phi_{\text{new}}}(x \mid y)$ are more reliable and accurate compared to the posterior samples from $q_\phi(x \mid y)$ and thus reward evaluated on the posterior samples from $q_{\phi_{\text{new}}}(x \mid y)$ are more reliable. Furthermore, the reward model $r_{\tilde{\eta}}$ benefits from training on more diverse samples, as entropy computed using the updated prior ϕ_{new} is more accurate than that from ϕ. This leads to improved data diversity during collection, enabling $r_{\tilde{\eta}}$ to converge faster than r_η. The following lemma supports this hypothesis:

Lemma 2 (Diverse Data Improves Convergence)**.** *[19, 20] Let θ_t be the parameters of a neural network trained using SGD with batch size 1 and learning rate η on dataset S. Let $\mathcal{L}_S(\theta)$ be the empirical loss. Assume the loss is L-smooth and gradients are bounded by G. Let S_1 and S_2 be two datasets of size n, with $D(S_1) > D(S_2)$. Then, for the same number of iterations T, the expected generalization gap*

$$\mathbb{E}[\mathcal{L}(\theta_T^{(1)}) - \mathcal{L}(\theta_T^{(2)})] < 0 \quad \textit{where,}\ D(S) = \frac{1}{n^2}\sum_{i,j}\|x_i - x_j\|^2 \quad \textit{for}\ (x_i, y_i), (x_j, y_j) \in S$$

where $\theta_T^{(1)}$ and $\theta_T^{(2)}$ are trained on S_1 and S_2 respectively, assuming the data distribution $\mathcal{D}$ has high support over $\mathcal{X}$.

Hence, $\theta_T^{(1)}$ is closer to optimal solution of $\mathcal{L}(\theta)$ than $\theta_T^{(2)}$ in fewer steps, assuming same training budget.

Thus, utilizing the result of the lemma 2, we can write:

$$\begin{aligned}&\big(\mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi^*(x|y)}}[r_{\eta^*}([\hat{x}_t^{(i)}]_{q_t})] - \mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi(x|y)}}[r_{\eta}([\hat{x}_t^{(i)}]_{q_t})]\big) \geq \\ &\big(\mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi^*(x|y)}}[r_{\eta^*}([\hat{x}_t^{(i)}]_{q_t})] - \mathbb{E}_{[\hat{x}_t^{(i)}]\sim q_{\phi_{\text{new}}(x|y)}}[r_{\tilde{\eta}}([\hat{x}_t^{(i)}]_{q_t})]\big)\end{aligned} \tag{38}$$

Now, combining the results of (38) and (39), we can write

$$\underbrace{\big(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi,\eta)}(q_t)\big)}_{B} \geq \underbrace{\big(\text{exploit}^{\text{score}}_{(\phi^*,\eta^*)}(q_t) - \text{exploit}^{\text{score}}_{(\phi_{\text{new}},\tilde{\eta})}(q_t)\big)}_{D} \tag{39}$$

Finally, leveraging the results of (37) and (40), and utilizing the definition of (35), we can write

$$||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi,\eta)}(q_t)|| \geq ||\text{Score}^*_{(\phi^*,\eta^*)}(q_t) - \text{Score}_{(\phi_{\text{new}},\tilde{\eta})}(q_t)||$$

□

A.7 Proof of Proposition 2

Proof. We start with the definition of entropy H:

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] = -H(\hat{x}_t|Q_t, y_{t-1})$$

Following the results of [18], and by setting $\alpha_i = \alpha_j = 1$, we obtain:

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \propto -\sum_{i=0}^{P}\log\sum_{j=0}^{P}\exp\left\{\frac{||\hat{x}_t^{(i)} - \hat{x}_t^{(j)}||_2^2}{2\sigma_x^2}\right\}$$

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \propto -\sum_{i,j}\log\left(\exp\left(\frac{\sum_{a\in Q_t}([\hat{x}_t^{(i)}]_a - [\hat{x}_t^{(j)}]_a)^2}{2\sigma_x^2}\right)\right)$$

Assuming k is the set of potential measurement locations at time step t, and $Q_t = Q_{t-1} \cup q_t$, where $q_t \in k$.

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \propto -\sum_{i,j}\log\left(\exp\left(\frac{\sum_{q_t\in k}([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2 + \sum_{r\in Q_{t-1}}([\hat{x}_t^{(i)}]_r - [\hat{x}_t^{(j)}]_r)^2}{2\sigma_x^2}\right)\right)$$

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \propto -\sum_{i,j}\log\left(\prod_{q_t\in k}\exp\left(\frac{([\hat{x}_t^{(i)}]_{q_t} - [\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}\right)\prod_{r\in Q_{t-1}}\exp\left(\frac{([\hat{x}_t^{(i)}]_r - [\hat{x}_t^{(j)}]_r)^2}{2\sigma_x^2}\right)\right)$$

https://doi.org/10.52202/085713-2163

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})] \propto -\sum_{i,j}\left(\sum_{q_t\in k}\frac{([\hat{x}_t^{(i)}]_{q_t}-[\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}+\sum_{r\in Q_{t-1}}\frac{([\hat{x}_t^{(i)}]_{r}-[\hat{x}_t^{(j)}]_{r})^2}{2\sigma_x^2}\right)$$

We then compute the expected log-likelihood at a specified measurement location q_t, discarding all terms independent of q_t. This key observation allows us to simplify the expression as follows:

$$\underbrace{\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})]\Big|_{q_t}}_{\text{The expected log-likelihood at a measurement location } q_t} \propto \sum_{i,j}\left(-\frac{([\hat{x}_t^{(i)}]_{q_t}-[\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}\right)$$

Equivalently, we can write the above expression as:

$$\mathbb{E}_{\hat{x}_t}[\log p(\hat{x}_t|Q_t, y_{t-1})]\Big|_{q_t} \propto \left(\underbrace{\sum_{i=0}^{P}\sum_{j=0}^{P}\exp\left\{-\frac{([\hat{x}_t^{(i)}]_{q_t}-[\hat{x}_t^{(j)}]_{q_t})^2}{2\sigma_x^2}\right\}}_{\text{likeli}^{\text{score}}(q_t)}\right)$$

By definition, the left-hand side of the above expression corresponds to $\text{likeli}^{\text{score}}(q_t)$. □

B Empirical Analysis of h model

B.1 Doob's h-transform as the Correction Factor

We demonstrate that the h-transform serves as a correction term for Tweedie's estimate. Specifically, the conditional Tweedie estimate can be expressed as:

$$\mathbb{E}[x_0 \mid x_t, y] \approx \hat{x}_0(x_t, y) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\left(h_t^{\zeta}(x_t, y) + s_t^{\theta^*}(x_t)\right)}{\sqrt{\bar{\alpha}_t}}$$

$$= \underbrace{\left(\frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}s_t^{\theta^*}\right)}_{\text{Unconditional Tweedie estimate}} - \underbrace{\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}h_t^{\zeta}(x_t, y)}_{\text{Correction Factor (i.e., } h\text{-transform)}}$$

In Equation (40), the first term represents the unconditional Tweedie estimate, illustrating that the h-transform can be viewed as a correction to the unconditional denoised prediction.

B.2 Details of h-model Update Scheduler

As highlighted in the main paper, the pessimistic updating of the h-model parameters plays a critical role in stabilizing the prior model's adaptability dynamics. While a straightforward heuristic might suggest updating the h-model after a fixed number of observations, this approach only provides marginal improvements. A more effective scheduling strategy requires fewer updates during the early stages of discovery, allowing the model to gather ample data and avoid the risk of erroneous updates when the observations are still sparse. However, as the discovery process progresses and the model's understanding of the search space strengthens, more frequent updates of the h-model become essential. This shift enables more effective exploitation of the environment, as the model has already gathered enough information, making the need for pessimistic updates unnecessary. Motivated by this observation, we propose the following h-model update scheduler:

$$\Delta t_i = \frac{\mathcal{B}}{U}\cdot(1-\frac{i}{U+1})^{\gamma} \tag{40}$$

Here, $\mathcal{B}$ denotes the overall sampling budget, U is the total number of h-model updates throughout the active discovery process, i indicates the current sampling step, γ governs the decay rate, and Δt_i defines the interval between two successive updates of the h-model parameters at the i-th sampling step. Note that since $\gamma \geq 1$, the update frequency of the h-model naturally accelerates with increasing i, leading to more frequent updates during the later stages of the active discovery process.

B.3 Effect of h-model Update Scheduler

In this section, we evaluate the effectiveness of the h-model update scheduler by comparing two variants of EM-PTDM: one using a uniform update schedule and the other employing the adaptive scheduler defined in Equation 40. For the uniform scheduler, the h-model is updated at fixed intervals—specifically, every 20 update steps. For the adaptive scheduler, we set $\gamma = 1$, $U = 30$, and an observation budget of $\mathcal{B} = \{200, 250\}$. The comparative results under this configuration are presented in Table 4. For this analysis, we consider active discovery of balls with a diffusion model trained on MNIST data as the prior model. Our empirical results show that **EM-PTDM with an adaptive h-model update scheduler consistently outperforms its uniform counterpart across various measurement budgets. This improvement can be attributed to fewer updates in the early stages, allowing the model to collect more informative observations before updating and thus reducing the risk of premature or noisy updates.** As the discovery progresses and the model gains a stronger understanding of the search space, the increased update frequency of h-model in later stages proves beneficial for accelerating active target discovery.

Table 4: Effect of Adaptive h-model Update Scheduler.

Active Discovery of Balls with MNIST Digit Images as the Prior.		
h-model update Schedule	$\mathcal{B} = 200$	$\mathcal{B} = 250$
Uniform	0.6856	0.7875
Adaptive	**0.7364**	**0.8268**

C Exploratory Nature of EM-PTDM

In active target discovery under an uninformative prior, exploration isn't just helpful—it's essential. Especially in the early stages, when the permanent memory offers little insight into the target domain and the correction factor is large, the h-model must rapidly adapt. This demands smart, strategic exploration of the search space to collect informative observations, enabling the h-model to efficiently learn and calibrate the correction factor. To assess EM-PTDM's exploratory behavior, we tackle active target discovery of overhead objects using ground-level imagery as prior knowledge, evaluating across a wide range of observation budgets—from very sparse (200) to less sparse (350)—and benchmark against baseline methods. We present the results in the following Table 5.

Table 5: Importance of Exploration

ATD of Overhead Objects with ImageNet as Prior.				
Method	$\mathcal{B} = 200$	$\mathcal{B} = 250$	$\mathcal{B} = 300$	$\mathcal{B} = 350$
DiffATD	0.3873	0.5143	0.6391	0.7348
GA	0.3479	0.4784	0.5659	0.6562
EM-PTDM	**0.4127**	**0.5620**	**0.7013**	**0.8256**

We observe that EM-PTDM's performance improvement over baselines grows with the observation budget. When the budget is low, the performance gap is narrow, reflecting limited opportunity for exploration. As the budget increases, EM-PTDM leverages richer exploration to adapt its h-model, leading to significantly more effective target discovery.

D Training and Inference Pseudocode

Below we present the training and inference pseudocode.

https://doi.org/10.52202/085713-2163

Algorithm 1 EM-PTDM SAMPLING STRATEGY (AT THE t-TH OBSERVATION STEP)

Require: Current State of Transient Memory: Trained h-transform $h_t^{\zeta}(x, \hat{x}_0, y)$ with parameters ζ.
Require: Permanent Memory as Unconditionally trained noise predictor $s_t^{\theta^*}(x_t)$
Require: Noise schedule $\beta_t = \beta(t)$, $\bar{\alpha}_t = \bar{\alpha}(t)$
Require: Sampling schedule $\sigma_t = \sigma(t)$
Require: Observation y, Posterior samples list $ps = []$, Success = $R = 0$.

$x_T \sim P_T = \mathcal{N}(0, 1)$ ▷ Sample a starting point
for $i = P$ to 1 **do**
  $\hat{x}^i = 0$
  for t in $(T, T-1, \ldots, 1)$ **do**
    $\hat{\epsilon}_\theta \leftarrow s_t^{\theta^*}(x_t)$ ▷ Predict unconditional noise
    $\hat{x}_0 \leftarrow \dfrac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\theta}{\sqrt{\bar{\alpha}_t}}$
    $\hat{\epsilon}_\zeta \leftarrow h_t^{\zeta}(x_t, \hat{x}_0, y)$ ▷ Predict correction noise via h-transform
    $\hat{\epsilon} \leftarrow \hat{\epsilon}_\theta + \hat{\epsilon}_\zeta$ ▷ Estimate posterior noise
    if $t > 1$ **then**
      Sample $\epsilon_t \sim \mathcal{P}_{\text{noise}}$
    else
      $\epsilon_t \leftarrow 0$
    end if
    $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\dfrac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\hat{\epsilon} + \sigma_t\epsilon_t$
    if $t = 1$ **then**
      $\hat{x}^i = x_{t-1}$.
    end if
  end for
  ps.append($\hat{x}^i$)
end for
Utilize Posterior samples in ps to compute $\text{expl}^{\text{score}}(q_t)$ and $\text{exploit}^{\text{score}}(q_t)$ using Eqn. 9 and 10 respectively for each $q_t \in k$.
Compute $\text{score}(q_t)$ using Eqn. 11 for each $q_t \in k$ and sample a location q_t with the highest score.
$\mathcal{B} \leftarrow \mathcal{B} - 1$, $\{k\} \leftarrow \{k\} \setminus q_t$
Update: $Q_t \leftarrow Q_{t-1} \cup q_t$, $y_t \leftarrow y_{t-1} \cup [x]_{q_t}$.
Update: $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{[x]_{q_t}, y^{(q_t)}\}$, R += $y^{(q_t)}$
Train r_η with updated $\mathcal{D}_t$ and optimize η with Cross-Entropy loss.
return R

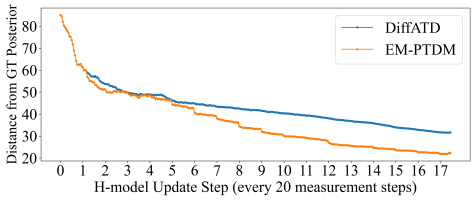
Algorithm 2 H-TRANSFORM FINE-TUNING (AT THE t-TH OBSERVATION STEP)

Require: Posterior Samples drawn from $q_{\phi_{t-1}}(x \mid y_{t-1})$
Require: Noise schedule $\beta_t = \beta(t)$, $\bar{\alpha}_t = \bar{\alpha}(t)$
Require: Permanent Memory (i.e. Pre-Trained Noise predictor function) $s_t^{\theta^*}(x)$ with parameters θ^*.
Require: Current state of Transient Memory (i.e. h-transform) $h_t^{\zeta}(x, \hat{x}_0, y)$ with parameters ζ.

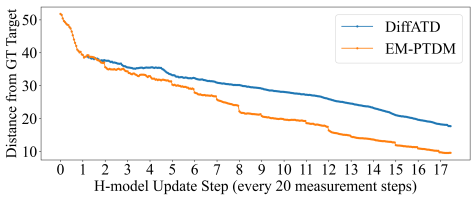
repeat
  $x_0 \sim P_0 = q_{\phi_{t-1}}(x \mid y_{t-1})$
  $t \sim$ Uniform($\{1, \ldots, T\}$)
  $\varepsilon_t \sim \mathcal{N}(0, I)$ ▷ Sample noise
  $x_t \leftarrow \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_t$
  $\hat{\varepsilon}_\theta \leftarrow s_t^{\theta^*}(x_t)$ ▷ Estimate noise with pretrained model
  $\hat{x}_0 \leftarrow \frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\varepsilon}_\theta}{\sqrt{\bar{\alpha}_t}}$
  $\hat{\varepsilon}_\zeta \leftarrow h_t^{\zeta}(x_t, \hat{x}_0, y)$ ▷ Estimate correction via h-transform
  Take gradient descent step w.r.t. ζ on
  $\nabla_\zeta \mathcal{L}(\varepsilon_t, \hat{\varepsilon}_\theta + \hat{\varepsilon}_\zeta)$ ▷ $\mathcal{L}$ is defined in Equation 7
until convergence or maximum epochs reached
return Updated $h-$model Parameter ζ.

E Efficacy of h-model in Estimating The Search Space with only Few Observations

We analyze the adaptability of the h-model using quantitative visualizations in the context of an active target discovery of overhead objects, where ground-level ImageNet images serve as the prior. Specifically, we assess the h-model's role by computing L2-based semantic similarity between predicted and ground-truth targets (right 9b), and between predicted and ground-truth posteriors (left 9a). As depicted in Figure 9, the inclusion of the h-model leads to significantly faster convergence to the true posterior and the true targets, thus enabling more informed target discovery. **The rapid drop in dissimilarity compared to using permanent memory alone highlights the h-model's ability to quickly correct the prior with minimal observations.**



(a) L2 Distance between predicted and ground-truth posterior.



(b) L2 Distance between predicted and ground-truth targets.

Figure 9: Quantitative analysis of h-model's adaptability in Active Target Discovery.

F Analyzing the Role of Permanent and Transient Memory for Enhancing In-Domain Target Discovery

Since we've seen that inside the EM-PTDM framework, **updating permanent memory with posteriors from prior in-domain ATD tasks boosts performance, it raises a key question: do we still need the h-model?** To investigate, we turn to DiffATD—a state-of-the-art baseline that uses only permanent memory. To update the permanent memory of DiffATD, we apply the same continual memory update strategy as in EM-PTDM, incorporating accumulated posteriors after each task, to see how far performance can go without the h-model. For this comparison, we examine active target discovery of overhead objects using ground-level ImageNet images as prior knowledge. While updating only the permanent memory leads to improved discovery rates compared to DiffATD with fixed permanent memory across different observation budgets, a clear and consistent performance gap remains when compared to EM-PTDM, particularly when EM-PTDM updates its permanent memory after each in-domain ATD task. **These results, summarized in Table 6, highlight the added value of the h-model in driving more effective exploration and adaptation irrespective of whether permanent memory is being updated or not.**

Table 6: Importance of h-model with or without Permanent Memory (**PM**) Update

ATD of Overhead Objects with ImageNet as Prior.			
Method	$\mathcal{B} = 250$	$\mathcal{B} = 300$	$\mathcal{B} = 350$
DiffATD	0.5143	0.6391	0.7348
DiffATD w/ **PM** Update	0.5294	0.6623	0.7589
EM-PTDM	0.5620	0.7013	0.8256
EM-PTDM w/ PM Update	**0.5859**	**0.7194**	**0.8461**

G Effect of $\kappa(\mathcal{B})$

We conduct experiments to assess the impact of $\kappa(\mathcal{B})$ on EM-PTDM's active discovery performance. Specifically, we investigate how amplifying the exploration weight, by setting $\kappa(\mathcal{B}) = \max\{0, \kappa(\alpha \cdot \mathcal{B})\}$ with $\alpha > 1$, and enhancing the exploitation weight by setting $\alpha < 1$, influence the overall effectiveness of the approach. We report results for $\alpha \in \{0.2, 1, 5\}$ in two settings: using

https://doi.org/10.52202/085713-2163

Table 7: Effect of $\kappa(\mathcal{B})$

Performance across varying α with $\mathcal{B} = 250$				
Target	Prior	$\alpha = 0.2$	$\alpha = 1.0$	$\alpha = 5.0$
Balls	MNIST	0.7416	0.7875	**0.9272**

overhead objects as targets with ground-level images as the prior (first row), and discovering target balls using MNIST digit images as the prior (second row). The results are summarized in Table 7. The best performance is achieved with $\alpha = 5$, and **the results suggest that higher values of α boosts performance, reinforcing the fact that exploration is key to success in active target discovery under an uninformative prior.**

H EM-PTDM's Capability of Discovering Isolated Targets within Observation Budget

To evaluate EM-PTDM's ability to uncover disjoint target regions within a limited sampling budget, we design a series of controlled toy experiments. In each task, the goal is to discover a varying number of balls—positioned differently and with different radii—using a diffusion model pretrained on MNIST digits as the permanent memory. We systematically increase task difficulty by varying the number of target balls from 5 to 10. Notably, tasks with more targets demand effective exploration of the search space to successfully locate all disjoint regions within the budget constraints. We present comparative visualizations of the exploration behavior of EM-PTDM and DiffATD with an active discovery task involving uncovering 10 target balls with MNIST Images as the prior, shown in Fig. 11. As the number of disjoint targets increases, making the task more challenging and exploration-intensive, EM-PTDM consistently succeeds in discovering most, if not all, targets within the given budget. In contrast, the baseline (i.e., DiffATD) relying solely on permanent memory struggles as task complexity rises, as seen in Plot 10. **These visualizations clearly demonstrate EM-PTDM's superior exploration capabilities, which are crucial for efficient target discovery in settings with multiple disjoint targets.**

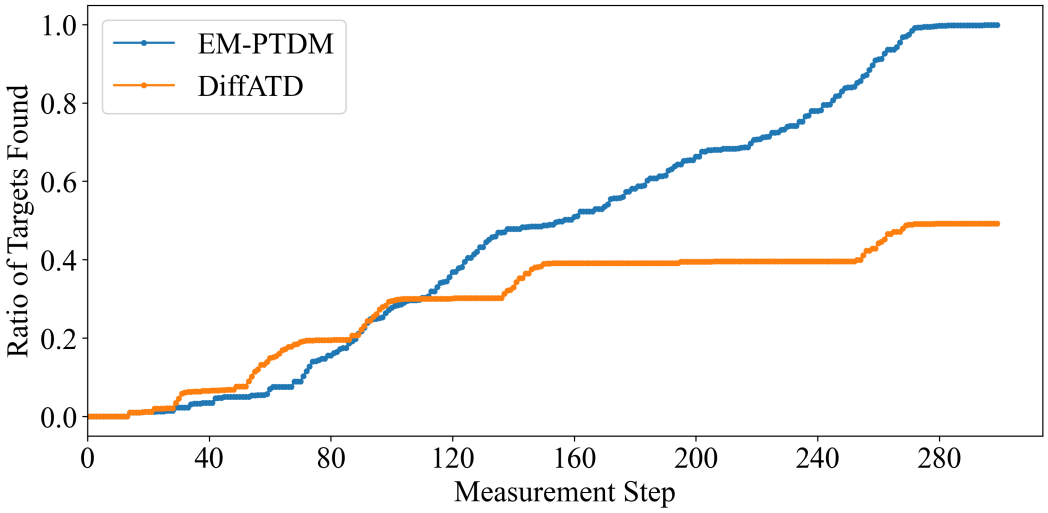


Figure 10: Comparison of the Discovery Process: EM-PTDM vs. DiffATD. In this experiment, we evaluate the active discovery of 10 target balls using a diffusion model trained on MNIST images as the prior. As the search budget increases, EM-PTDM consistently uncovers more disjoint target balls, thanks to its inherently exploratory behavior. In contrast, DiffATD struggles to discover as many targets as it lacks the same efficiency as EM-PTDM in exploring the search space.

https://doi.org/10.52202/085713-2163

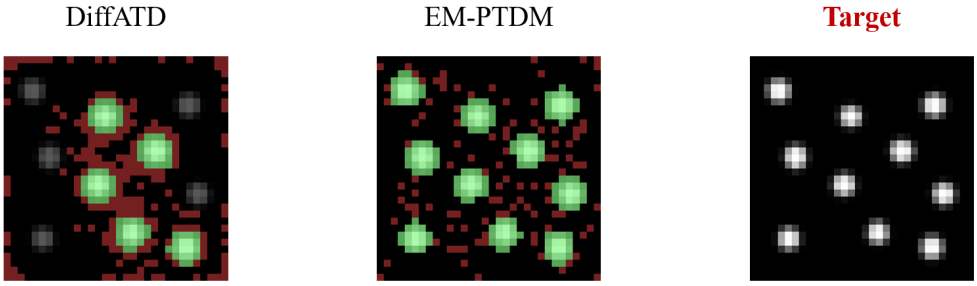


Figure 11: Comparison between EM-PTDM and DiffATD's Discovery. **Green** patches correspond to successful target discovery, and **Red** Patches correspond to unsuccessful observations. In this example, the task is to discover the balls with MNIST images as the prior. The exploratory behavior of EM-PTDM, in contrast to DiffATD, is evident in the visualization.

I Species Distribution Modelling as Active Target Discovery Problem

We constructed our species distribution experiment using observation data of the chosen species from iNaturalist. Center points were randomly sampled within North America (latitude 25.6°N to 55.0°N, longitude 123.1°W to 75.0°W). Around each center, we defined a square region approximately 480 km × 480 km in size (roughly 5 degrees in both latitude and longitude). Each retained region was discretized into a 64×64 grid, where the value of each cell represents the number of observed species. To simulate the querying process, each 2×2 block of grid cells was treated as a query.

J Additional Results on Active Discovery of Unknown Species from Known Species Distribution

In the main paper, we explored active discovery of Coccinella septempunctata (CS) using the known species distribution of Gladicosa and Gonioctena (GG) as the prior. This section extends our analysis to a different species from the iNaturalist dataset. Specifically, we evaluate EM-PTDM on a task that involves discovering Species Cedar Waxwing from the known distribution of Species Black-capped Chickadee. Figure 12 compares the exploration behavior of EM-PTDM and DiffATD at various stages of target discovery. As shown, EM-PTDM—starting from the same prior as DiffATD—progressively and efficiently adapts the prior toward the true target distribution with only a few task-specific observations. In contrast, DiffATD, which relies solely on static permanent memory, struggles to approximate the ground-truth distribution within the given observation budget.

Table 8: *SR* Comparison with Baselines.

Active Discovery of Cedar Waxwing Species with Species Black-capped Chickadee as Prior.			
Method	$\mathcal{B} = 150$	$\mathcal{B} = 200$	$\mathcal{B} = 250$
DiffATD	0.2319	0.3309	0.4453
GA	0.2232	0.3098	0.3116
EM-PTDM	**0.3079**	**0.3854**	**0.6347**

We further compare the performance of EM-PTDM against baseline methods using Success Rate (SR) as the evaluation metric, with results summarized in Table 8. The task involves actively discovering Species Cedar Waxwing from the known distribution of Species Black-capped Chickadee. Consistent with other experimental settings, EM-PTDM significantly outperforms all baselines across varying observation budgets. These results further highlight the effectiveness of EM-PTDM in tackling active target discovery under an uninformative prior.

https://doi.org/10.52202/085713-2163

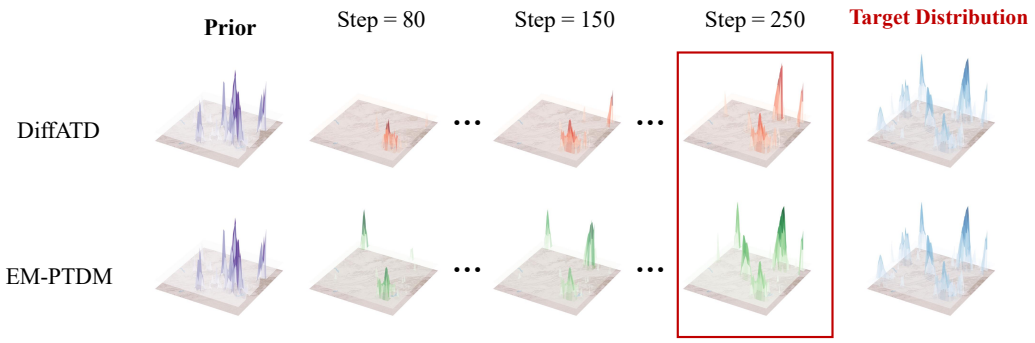


Figure 12: Exploration Behavior of Different Approaches. We visualize the explored Regions at Different Active Target Discovery Phases. We consider the task of Active Discovery of the species Cedar Waxwing with a known distribution of the species Black-capped Chickadee. EM-PTDM discovers most target regions (i.e., more accurately discovers the existence of the species Cedar Waxwing).

K More Visualizations on Efficiency of h-model's Adaptability From Very Sparse Observations

In this section, we provide additional visualizations of posterior samples generated by EM-PTDM and DiffATD across different stages of active target discovery. As shown in Figures (13, 14, 15), EM-PTDM produces samples that are more semantically aligned with the ground-truth posterior compared to DiffATD. Notably, even with sparse observations, EM-PTDM effectively simulates the search space, enabling more informed exploration and leading to improved target discovery under an uninformative prior.

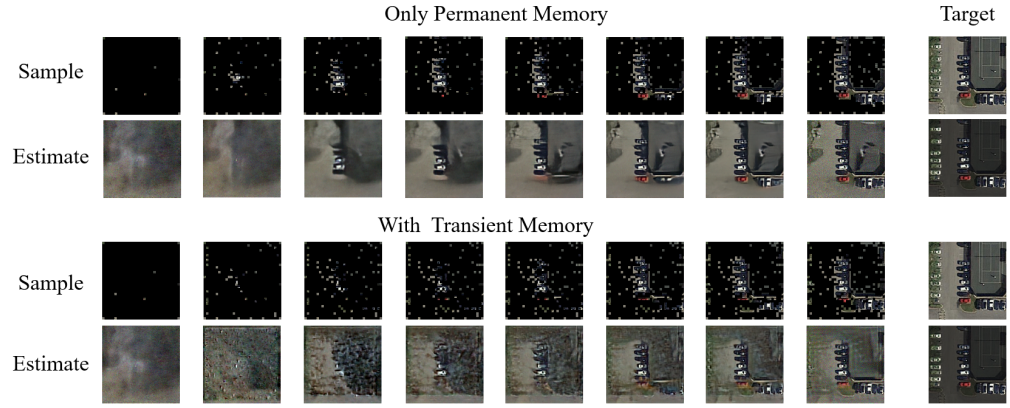


Figure 13: h-model's Adaptability From Very Sparse Observations. Posterior Samples at Different Active Target Discovery Phases. Overhead Object Discovery (i.e., **Car**) with Ground Level images from ImageNet as the Prior. Observation Budget of 300.

https://doi.org/10.52202/085713-2163

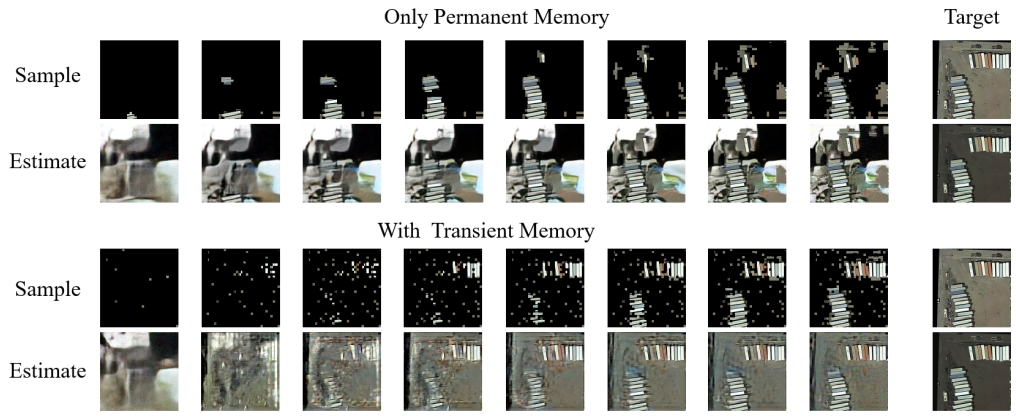


Figure 14: h-model's Adaptability From Very Sparse Observations. Posterior Samples at Different Active Target Discovery Phases. Overhead Object Discovery (i.e., **Truck**) with Ground Level images from ImageNet as the Prior. Observation Budget of 300.

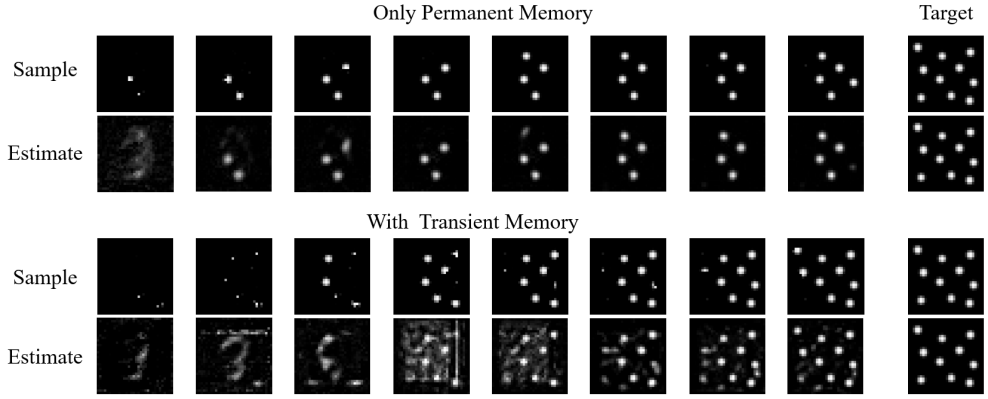


Figure 15: h-model's Adaptability From Very Sparse Observations. Posterior Samples at Different Active Target Discovery Phases. Uncovering Disjoint Balls with MNIST digit images as the Prior.

L More Visualizations of the Exploration Behavior of EM-PTDM at Different Active Target Discovery Phases

In this section, we present additional exploration behavior of EM-PTDM at different active target discovery phases. We also provide a similar exploration behavior of the baseline approaches, including DifffATD and Greedy Adaptive, for the comparison. These visualizations are provided in Figures 16, 17, 18. These additional visualizations further reinforce the effectiveness of EM-PTDM in addressing active target discovery under an uninformative prior.

https://doi.org/10.52202/085713-2163

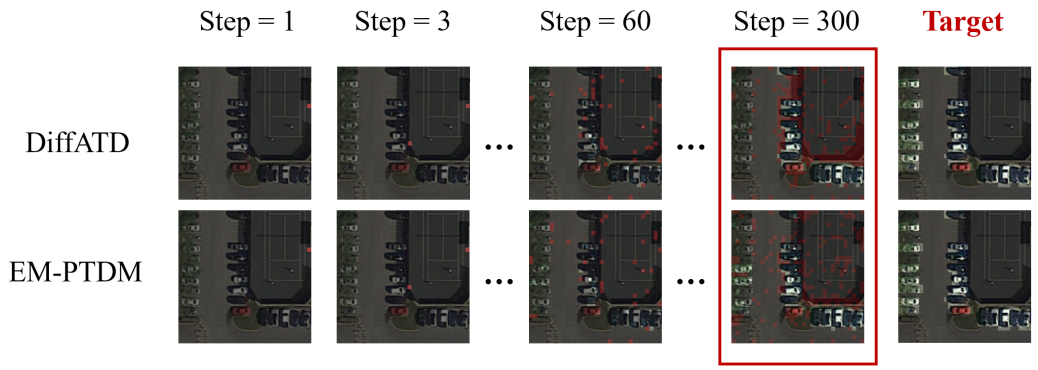


Figure 16: Visualizing the regions explored by each method at various stages of the ATD process. In these visualizations, patches corresponding to **successful queries are unmasked**, while those resulting in **unsuccessful queries are highlighted in Red**. The task focuses on discovering overhead objects (**cars**), using ground-level images from ImageNet as the prior. The results demonstrate that EM-PTDM effectively identifies and explores most of the target regions containing **cars**.

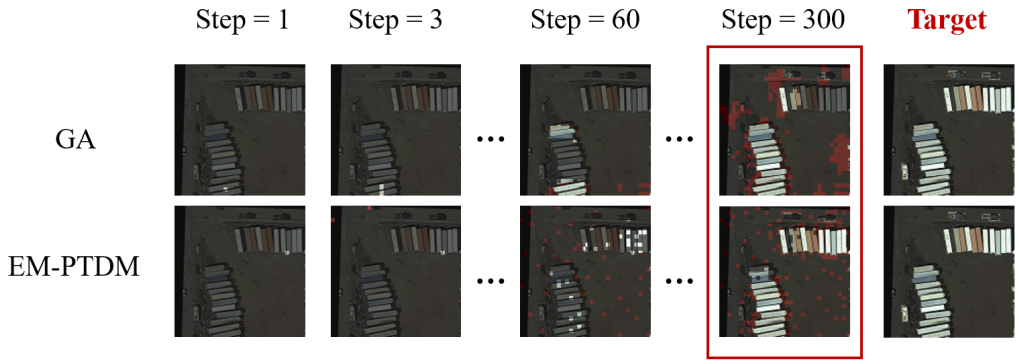


Figure 17: Visualizing the regions explored by each method at various stages of the ATD process. In these visualizations, patches corresponding to **successful queries are unmasked**, while those resulting in **unsuccessful queries are highlighted in Red**. The task focuses on discovering overhead objects (**trucks**), using ground-level images from ImageNet as the prior. The results demonstrate that EM-PTDM effectively identifies and explores most of the target regions containing **trucks**.

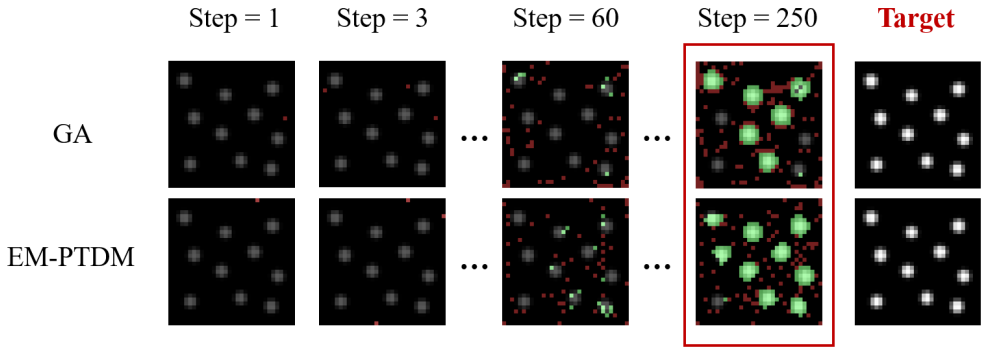


Figure 18: Visualizing the regions explored by each method at various stages of the ATD process. In these visualizations, patches corresponding to **successful queries are Green**, while those resulting in **unsuccessful queries are highlighted in Red**. The task focuses on uncovering Disjoint Balls from MNIST digit images as the Prior. EM-PTDM discovers most target regions (i.e., Disjoint Balls).

https://doi.org/10.52202/085713-2163

M Active Target Discovery of Balls Using MNIST Images as the Prior

To enable a comprehensive analysis of EM-PTDM, we introduce a custom-designed dataset tailored for this task. It simulates active discovery scenarios involving an unknown number of balls with unknown locations and radii, with MNIST images as the prior. Success in this setting requires effective exploration of the search space to accurately localize the targets, capturing the core challenge of the problem. The details of the proposed dataset are provided below.

M.1 Dataset Creation Procedure

In this dataset, we generate each sample by randomly placing 5 to 10 identical balls within a 32×32 2D grid. The radius of all balls in a given sample is either 3 or 4 pixels, randomly selected per sample. All placements are performed uniformly at random, subject to the non-overlapping constraint and the boundary condition that each ball lies entirely within the 32×32 space.

M.2 SR Comparisons with Baseline Approaches

As in previous settings, we quantitatively evaluate EM-PTDM and baseline methods using the Success Rate (SR) metric. In this experiment, the task involves actively discovering target balls using a diffusion model trained on the MNIST dataset as the prior. The results, summarized in Table 9, show a consistent trend: EM-PTDM significantly outperforms all baselines across different measurement budgets. This further reinforces the effectiveness of EM-PTDM in handling active target discovery under an uninformative prior.

Table 9: *SR* Comparison with Baselines.

Active Discovery of balls with MNIST Digit Images as Prior.			
Method	$\mathcal{B} = 150$	$\mathcal{B} = 200$	$\mathcal{B} = 250$
RS	0.1458	0.1826	0.2187
DiffATD	0.4362	0.4432	0.4929
GA	0.3250	0.5170	0.6257
EM-PTDM	**0.5561**	**0.6856**	**0.7875**

M.3 Analyzing the Exploration Strategies of EM-PTDM and DiffATD Under Increasing Task Complexity

In this section, we provide additional visualizations highlighting the exploration behavior of EM-PTDM and DiffATD across different stages of the active target discovery task. Using the task of discovering target balls with MNIST images as the prior, the visualizations in Figures 19, 20 clearly show that EM-PTDM consistently explores more effectively and identifies targets with higher accuracy, even under increasing task complexity while adhering to a strict budget and under an uninformative prior. These results further underscore the robustness and adaptability of EM-PTDM in challenging discovery scenarios, where efficient exploration of the search space is the key. **A striking emergent behavior is observed across both examples: in the early stages of the active discovery process, EM-PTDM engages in broader exploration of the search space compared to DiffATD. This initially results in fewer target discoveries (e.g., at step 60). However, this strategic exploration enables EM-PTDM to build a richer understanding of the environment, which it later exploits to surpass DiffATD, ultimately identifying a greater number of target regions before the observation budget is depleted (see step 250).**

https://doi.org/10.52202/085713-2163

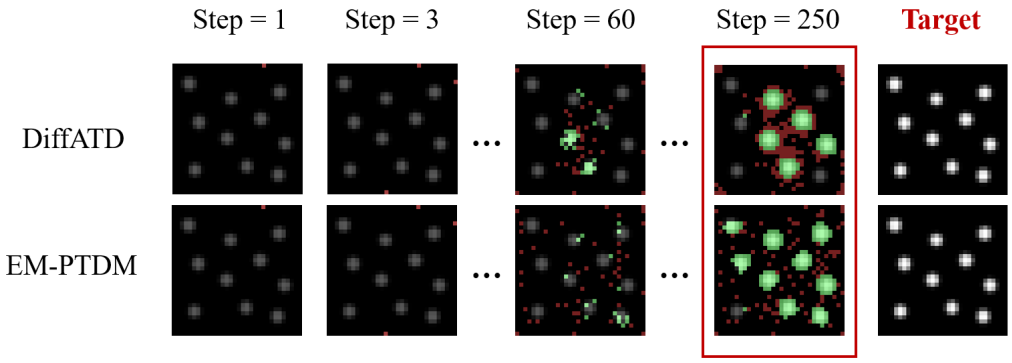


Figure 19: Visualizing the regions explored by each method at various stages of the ATD process. In these visualizations, patches corresponding to **successful queries are highlighted in Green**, while those resulting in **unsuccessful queries are highlighted in Red**. The task focuses on uncovering Disjoint Balls from MNIST digit images as the Prior. The results demonstrate that EM-PTDM discovers most target regions (i.e., Disjoint Balls).

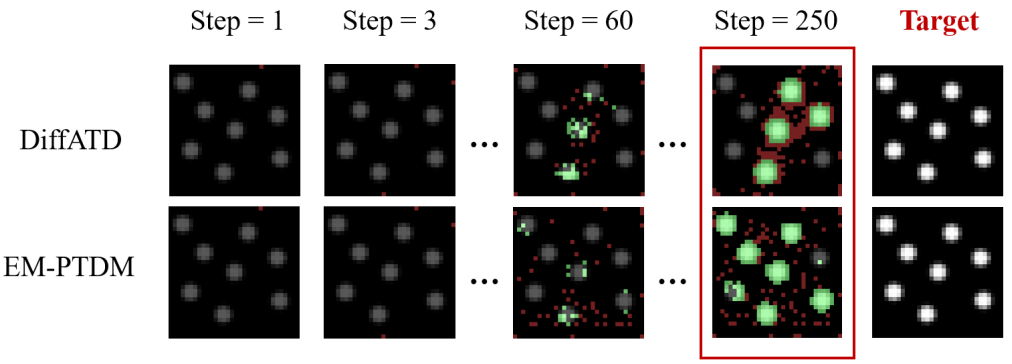


Figure 20: Visualizing the regions explored by each method at various stages of the ATD process. In these visualizations, patches corresponding to **successful queries are highlighted in Green**, while those resulting in **unsuccessful queries are highlighted in Red**. The task focuses on uncovering Disjoint Balls from MNIST digit images as the Prior. The results demonstrate that EM-PTDM discovers most target regions (i.e., Disjoint Balls).

N Architecture, Training Details: $h-$model, Pretrained Diffusion Model, and the Reward Model; and Computing Resources

N.1 Details of h-model

For the MNIST-to-Balls tasks, we employ 32-dimensional diffusion time-step embeddings and a single-block U-Net h-model with layer widths of [32, 64]. In the species discovery tasks, we also use 32-dimensional time-step embeddings, but with a single-block h-model featuring wider layers [32, 64, 128]. For the ImageNet-to-DOTA tasks, we increase the embedding dimensionality to 128, using a similar single-block h-model with widths [32, 64, 128]

N.2 Details of Reward Model

Our proposed method, EM-PTDM, utilizes a parameterized reward model, r_η, to steer the exploitation process. To this end, we employ a neural network consisting of a series of convolutional and fully connected layers, with non-linear ReLU activations as the reward model (r_η). The reward model's

https://doi.org/10.52202/085713-2163

goal is to predict a score ranging from 0 to 1, where a higher score indicates a higher likelihood that the measurement location corresponds to the target, based on its semantic features. Note that the size of the input semantic feature map for a given measurement location can vary depending on the downstream task. For instance, when working with DOTA, we use an 4×4 patch as the input feature size. After each measurement step, we update the model parameters (η) using the binary cross-entropy loss. Additionally, the training dataset is updated with the newly observed data point, refining the model's predictions over time. Naturally, as the search advances, the reward model refines its predictions, accurately identifying target-rich regions, which makes it progressively more dependable for informed decision-making. The reward model architecture consists of 1 convolutional layer with a 3×3 kernel, followed by 5 fully connected (FC) layers, each with its own weights and biases. The first FC layer maps an input of size $\frac{(input\ size)^2}{4}$ to an output of size 4 with weights and biases of size $[\frac{(input\ size)^2}{4}, 4]$ and $[4]$ respectively. The second FC layer transforms an input of dimension 4 to an output of size 32 with a 2-dimensional weight of size $[4, 32]$ and a bias of size $[32]$. The third FC layer maps 32 inputs to 16 outputs via a weight matrix of shape $[32, 16]$ and a bias vector of size $[16]$. The pre-final FC layer transforms inputs of size 16 to outputs of size 8 with $[16, 8]$ weights, and a bias of shape $[8]$. The final FC layer produces an output of size 2, with weights of size $[8, 2]$ and a bias of size $[2]$, representing the target and non-target scores. The reward model uses the leaky ReLU activation function after each layer. We update the reward model parameters after each measurement step based on the binary cross-entropy loss. The reward model is trained incrementally for 3 epochs after each measurement step using the gathered supervised dataset resulting from sequential observation, with a learning rate of 0.01.

N.3 Details of Primary Memory as Pretrained Diffusion Model

We use DDIM [21] as the diffusion model across datasets. The diffusion models used in different experiments are based on widely adopted U-Net-style architecture. For the MNIST dataset, we use 32-dimensional diffusion time-step embeddings, with the diffusion model consisting of 2 residual blocks. We select the time-step embedding vector dimension to match the input feature size, ensuring the diffusion model can process it efficiently. The block widths are set to $[32, 64, 128]$, and training involves 30 diffusion steps. For DOTA, we use the input feature size of $[128, 128, 3]$, the architecture featuring 128-dimensional time-step embeddings and a diffusion model with 2 residual blocks of width $[64, 128, 256, 256, 512]$. Finally, all experiments are implemented in Tensorflow and conducted on NVIDIA A100 40G GPUs. Our training and inference code will be made public.

O Statistical Significance Results of EM-PTDM

In order to strengthen our claim on EM-PTDM's superiority over the baseline methods, we have included the statistical significance results with different active target discovery settings, and present the results in Tables 10, 11. These results are based on 5 independent trials and further strengthen our empirical findings, reinforcing the stability and effectiveness of EM-PTDM in tackling active target discovery under an uninformative prior across diverse domains.

Table 10: Statistical Significance Results for Unknown Overhead Object Discovery.

Active Discovery of Overhead Objects with Ground Level ImageNet Images as the Prior.			
Method	$\mathcal{B} = 250$	$\mathcal{B} = 300$	$\mathcal{B} = 350$
RS	0.2325 ± 0.0190	0.2852 ± 0.0137	0.3207 ± 0.0168
DiffATD	0.5143 ± 0.0067	0.6391 ± 0.0102	0.7348 ± 0.0041
GA	0.4784 ± 0.0122	0.5659 ± 0.0096	0.6562 ± 0.0054
EM-PTDM	$\mathbf{0.5620 \pm 0.0073}$	$\mathbf{0.7013 \pm 0.0038}$	$\mathbf{0.8256 \pm 0.0093}$

https://doi.org/10.52202/085713-2163

Table 11: Statistical Significance Results for Unknown Species Discovery Task.

Active Discovery of Species CS with Species GG as the Prior.			
Method	$\mathcal{B} = 150$	$\mathcal{B} = 200$	$\mathcal{B} = 250$
RS	0.1624 ± 0.0133	0.2327 ± 0.0201	0.2775 ± 0.0154
DiffATD	0.3420 ± 0.0115	0.4365 ± 0.0057	0.4808 ± 0.0063
GA	0.4061 ± 0.0047	0.5067 ± 0.0079	0.5567 ± 0.0085
EM-PTDM	$\mathbf{0.4983 \pm 0.0060}$	$\mathbf{0.6495 \pm 0.0108}$	$\mathbf{0.6989 \pm 0.0056}$

P Impact of Weak Permanent Memory on Active Target Discovery Performance

When an extremely weak prior is used as the permanent memory, according to Equation (6), adapting to a new domain essentially reduces to learning the task from scratch through the transient memory alone. In this case, as indicated by Equation (7), the h-model no longer serves as a corrective mechanism; instead, under a partially observable environment, the entire responsibility for modeling the posterior shift falls on this lightweight module. However, this is beyond the capacity of such a lightweight module by design. In order to validate our hypothesis, we perform additional experiments under a remote sensing scenario, where we deliberately replaced the permanent memory with a diffusion model that could only output noise without any meaningful semantic structure. On top of this setup, we implemented EM-PTDM and observed the following phenomenon: even h-model was updated under partial observations, its limited capacity made it insufficient to compensate for the lack of a meaningful prior. As a result, all of the unexplored regions of the posterior estimation remained dominated by the weak prior—essentially resembling noise—leading to poor global environment estimation. We present our findings in the following Table 12. This supports the claim that **an extremely weak or non-semantic prior forces the transient memory to handle an unrealistically large modeling responsibility, which goes beyond its design capability.**

Table 12: Analysis of Weak Permanent Memory: Performance Comparison on DOTA

Method	$\mathcal{B} = 250$	$\mathcal{B} = 300$
EM-PTDM (**Random Noise** as Permanent Memory)	0.3117	0.3465
EM-PTDM (**ImageNet** as Permanent Memory)	**0.5620**	**0.7013**

Q More Details on Computational Cost across Search Space

We have conducted a detailed evaluation of sampling time and computational requirements of EM-PTDM across various search space sizes. We present the results in the following table. Our results (as reported in 13) show that **EM-PTDM remains efficient even as the search space scales, with sampling time per observation step ranging from 0.83 to 1.87 seconds, which is well within practical limits for most downstream applications.** This further reinforces EM-PTDM's scalability and real-world applicability.

Table 13: Details of Computation and Sampling Cost Across Varying Search Space Sizes

Active Discovery of Handwritten Digits		
Search Space	Computation Cost	Sampling Time per observation step (Seconds)
28×28	0.78 GB	0.83
128×128	1.51 GB	1.87

R Code Link

Our code and models are publicly available at this link.

https://doi.org/10.52202/085713-2163