Belief-Contraction-Driven Active Inverse Source Localization and Characterization

Yiwei Shi[1] , **Mengyue Yang**[1*] , **Qi Zhang**[2*] , **Cunjia Liu**[3*] , **Weinan Zhang**[4*] and **Weiru Liu**[1*]
[1]School of Engineering Mathematics and Technology, University of Bristol
[2]Department of Computer Science and Technology, Tongji University
[3]Department of Aeronautical and Automotive Engineering, Loughborough University
[4]School of Computer Science, Shanghai Jiao Tong University
yiwei.shi@bristol.ac.uk

Abstract

Active inverse source localization and characterization (ISLC) in dynamic fields requires sequential decision making under partial observability, where a mobile sensor must infer latent source parameters from sparse, noisy readings. We introduce a belief-contraction-driven approach that unifies inference, stopping, and control. An attention-augmented particle filter stabilizes Bayesian belief updates through ESS-based resampling, feature-aware sparse attention smoothing, and Metropolis–Hastings rejuvenation that preserves the filtering posterior. Belief contraction (posterior dispersion) defines both a termination rule and a goal-aligned intrinsic reward, enabling reinforcement learning without distance-to-source shaping. Across seven field modalities, spatial out-of-distribution tests, and nonstationary source shifts, our agent (ATT-PFRL) achieves higher completion, faster convergence, and more accurate localization than planning and RL+Bayes baselines under similar computation. Fixed-trajectory studies also show improved ESS and lower RMSE, isolating the benefit of the inference layer.

1 Introduction

Accurately and rapidly localizing unknown sources in spatially distributed fields, such as gas leaks, pollutant releases, or electromagnetic anomalies,is a critical capability for industrial safety, environmental monitoring, and emergency response. These tasks are commonly formulated as **Inverse Source Localization and Characterization (ISLC)**, where the objective is to infer latent source and environmental parameters (e.g., source position, intensity, and transport conditions) from *sparse, noisy, and localized measurements* collected by a mobile agent [Steiner and Bushe, 2001]. Active ISLC is inherently a **partially observable** sequential decision-making problem. Observations are local and often weakly informative, while the latent parameters remain unobserved; thus, effective behavior must be conditioned on a **belief** over those parameters rather than on raw measurements alone. This setting creates three practical challenges. First, *intermediate feedback is sparse and implicit*: the environment typically does not provide dense rewards that indicate progress toward successful localization, making reinforcement learning (RL) unstable or heavily dependent on hand-crafted shaping. Second, maintaining a reliable posterior under sparse/noisy sensing is non-trivial: standard particle filters can suffer from *weight degeneracy* and resampling-induced *particle impoverishment*, leading to overconfident yet unreliable beliefs. Third, real deployments demand robustness across varying field modalities and distribution shifts (e.g., changed wind or spatial regimes), where reactive heuristics or narrowly trained policies may fail to generalize.

Prior work approaches ISLC through Bayesian planning and information-theoretic search [Vergassola *et al.*, 2007; Hutchinson *et al.*, 2018a], bio-inspired plume tracing and casting strategies, or deep RL methods for adaptive exploration [Zhao *et al.*, 2022; Hu *et al.*, 2019; Park *et al.*, 2022]. However, existing solutions often face a trade-off between (i) *principled inference* and (ii) *goal-aligned control*. Information-theoretic planning can be computationally demanding and sensitive under partial observability, while RL approaches frequently rely on reward shaping that may be misaligned with the true objective (pinpointing the source) and can generalize poorly across conditions.

In this paper, we propose a *belief-contraction-driven framework* that couples Bayesian belief maintenance with belief-conditioned decision making for active ISLC. Our key idea is to treat *posterior contraction*—the reduction of uncertainty in the inferred belief—as the common objective that links inference, termination, and learning. Concretely, we (i) maintain a particle-based posterior over latent parameters and (ii) use a contraction criterion (based on posterior dispersion) as a *task-completion signal* and as a *sparse intrinsic reward* for learning. This converts implicit "mission completion" into an explicit, goal-aligned learning signal, avoiding distance-to-source shaping and directly reflecting whether the agent has localized the source with sufficient confidence.

To make belief contraction meaningful and robust under sparse/noisy sensing, we introduce an *attention-augmented particle inference layer* that mitigates degeneracy while controlling computational overhead. Starting from sequential importance sampling, we stabilize belief updates via (1) *ESS-*

*Corresponding author.

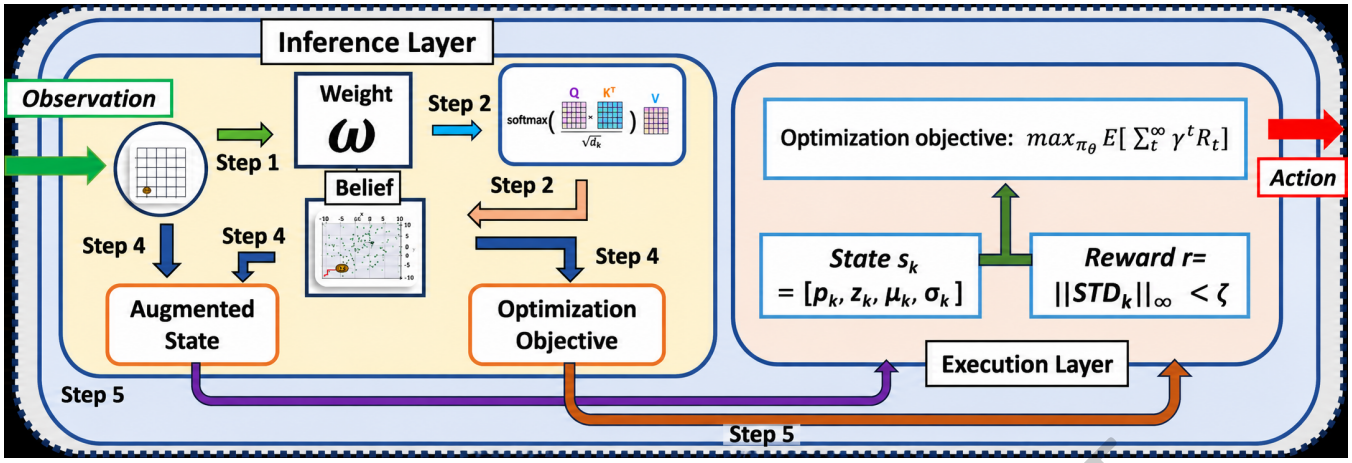


Figure 1: Attention-Based Inference Framework

triggered resampling when weight collapse is detected, (2) *feature-aware sparse attention smoothing* that redistributes weight mass among feature-similar hypotheses as a controlled perturbation on the simplex, and (3) *MCMC rejuvenation with Metropolis–Hastings correction* to restore diversity while preserving posterior invariance. The resulting belief state (posterior moments and uncertainty summaries) serves as the input to an execution policy. In this work, we instantiate the executor as an RL agent,*ATT-PFRL*,that learns exploration strategies conditioned on the belief, optimizing the sparse contraction-based intrinsic reward until the termination criterion is met. We validate the proposed framework in a suite of ISLC environments spanning diverse field modalities and challenging shifts. We further evaluate *out-of-distribution spatial shifts* and *nonstationary source changes*, highlighting robustness to distribution mismatch and dynamic regimes. To isolate inference improvements from policy-induced data collection, we additionally conduct *fixed-trajectory evaluations* that compare inference variants under identical observation sequences, and we perform dedicated analyses of *cessation threshold sensitivity* and *intrinsic reward design*, verifying that the contraction signal is calibrated against ground-truth localization quality.

Our main contributions are: 1.) A unified framework that aligns inference, termination, and learning by using posterior contraction as both a stopping criterion and a sparse intrinsic reward for belief-conditioned control. 2.) An efficient degeneracy-mitigation pipeline combining ESS-triggered resampling, feature-aware sparse attention smoothing, and MH-corrected rejuvenation to stabilize belief updates under sparse/noisy observations while preserving posterior correctness. 3.) Extensive experiments across diverse field types, out-of-distribution regimes, and nonstationary shifts, including inference decoupling and reward/cessation validations that directly support the proposed design choices.

2 Related Work

Traditional planning-based methods for ISLC typically fall into two categories: Information-Theoretic Approaches and Bio-Inspired Methods. The former strategies aim to reduce uncertainty about the source location, with *Infotaxis* [Vergassola *et al.*, 2007] minimizing posterior variance, *Entrotaxis* [Hutchinson *et al.*, 2018b] focusing on high-entropy regions, and *DCEE* [Chen *et al.*, 2021] balancing exploration and exploitation through uncertainty-driven decision-making. Although these methods have proven effective in controlled settings, they often rely on strong assumptions, leading to reduced efficiency and success rates when conditions deviate from their underlying assumptions. Meanwhile, the latter algorithms like *plume-tracing* [Farrell *et al.*, 2005] and *zigzag* [Balkovsky and Shraiman, 2002; Lochmatter *et al.*, 2008] mimic the *gradient-following behaviors* used by insects and animals—detecting changes in chemical concentration and adjusting movement to reacquire the plume if it is lost. Though effective in stable environments, these reactive approaches often fail or converge slowly in nonstationary, noisy, or complex fields, where their assumptions and simple local cues become inadequate. Unlike planning methods, **RL** [Mnih *et al.*, 2015; Schulman *et al.*, 2017; Lillicrap, 2015; Hu *et al.*, 2025] excels in high-dimensional tasks but faces two major hurdles in ISLC: sparse or implicit rewards, where locating a hidden source offers little immediate feedback, and inference, as RL inherently lacks the capacity to directly deduce hidden variables in noisy, evolving fields. Recent efforts combine RL with Bayesian updates to address this, but many approaches still rely on hand-crafted rewards or are domain-specific. PC-DQN [Zhao *et al.*, 2022] combines particle clustering and deep Q-networks for feature extraction, demonstrating high success rates in turbulent environments. LSTM-based RL [Hu *et al.*, 2019] enhances historical trajectory encoding for robust plume tracing under dynamic conditions, while DDPG with GMM [Park *et al.*, 2022] features facilitates real-time, efficient source estimation in stochastic scenarios. Among these methods, AGDC [Shi *et al.*, 2025] excels in handling sparse-feedback settings by detecting and halting upon goal completion while leveraging inference to improve decision-making under uncertainty. Its three variants extend its functionality: AGDC-KLD minimizes KL divergence for accurate localization [Filippi *et al.*, 2010], AGDC-ENT focuses on high-uncertainty areas[Haarnoja *et*

al., 2018], and AGDC-EE balances exploration and exploitation [Li *et al.*, 2022]. Despite its advancements, AGDC still faces challenges in efficiency, particularly in inference speed and scalability, for dynamic real-time environments.

3 Preliminaries

3.1 Convection-Diffusion Equation

Many seemingly disparate natural phenomena, such as *pollutant dispersion*, *gas diffusion*, and *electric field distributions*, share common physical mechanisms: **diffusion**, **convection**, and **external sources**. These mechanisms can be mathematically described by the general *convection-diffusion equation* (CDE) [Holley, 1969], which provides a versatile framework:

$$\alpha\nabla^2\phi - \vec{v}\cdot\nabla\phi + k_r\phi + S(x,y) = 0, \tag{1}$$

where $\phi(x,y)$ denotes the field variable (e.g., concentration or temperature), with $\alpha\nabla^2\phi$, $-\vec{v}\cdot\nabla\phi$, and $S(x,y)$ accounting for diffusion, convection, and external sources, respectively. The parameters, diffusion coefficient α, convection velocity $\vec{v}$, reaction rate k_r, and source term $S(x,y)$, can be adapted to describe various physical phenomena, from heat conduction to pollutant dispersion.

3.2 Gaussian Plume Model

The Gaussian plume model is derived as a steady-state analytical solution of the CDE, balancing simplicity and computational efficiency: $\phi(x,y) = \frac{q_s}{4\pi\alpha\|\boldsymbol{p}-\boldsymbol{p}_s\|}\exp\left(-\frac{\|\boldsymbol{p}-\boldsymbol{p}_s\|}{\lambda} - \frac{(x-x_s)u_x+(y-y_s)u_y}{2\alpha}\right)$, where q_s is the source strength, $\boldsymbol{p} = (x,y)$ and $\boldsymbol{p}_s = (x_s,y_s)$ are the coordinates of the observation point and source, respectively, u_x, u_y are the components of the convection velocity $\vec{v}$, $\lambda = \sqrt{\gamma\tau_s \Big/ (1 + u_s^2\tau_s/4\gamma)}$ is a decay parameter, τ_s is the time duration and α is the diffusion coefficient consistent with the general CDE notation. When wind is parameterized by speed and direction, we use $u_x = u_s\cos\varphi_s$ and $u_y = u_s\sin\varphi_s$. We parameterize wind as (u_x, u_y).

3.3 Partially Observable Markov Decision Process

Dynamic field modeling frequently encounters uncertainties arising from incomplete state information and noisy observations. These challenges can be systematically addressed through the *Partially Observable Markov Decision Process (POMDP)*, defined by the tuple $(S, \Omega, A, T, O, R, \gamma)$. At each discrete time step, the agent receives an observation $o_t \in \Omega$ according to $O(o_t|s_t, a_{t-1})$, takes an action $a_t \in A$, and the environment transitions to s_{t+1} via $T(s_{t+1}|s_t, a_t)$. The objective is to find a policy that maximizes the expected discounted return. Under partial observability, the policy conditions on a belief state. We model active ISLC as a POMDP where the *full* system state is $s_t = (\boldsymbol{p}_t, \Theta_t)$, consisting of the agent position $\boldsymbol{p}_t \in \mathbb{R}^2$ and latent source/environment parameters Θ_t (e.g., source location/strength, wind, diffusivity). The action is a discrete motion command $a_t \in \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$ that deterministically updates the position $\boldsymbol{p}_{t+1} = f(\boldsymbol{p}_t, a_t)$ (with boundary clipping). The observation is $o_t = (\boldsymbol{p}_t, z_t)$, where z_t is the noisy sensor reading at $\boldsymbol{p}_t$ with likelihood $p(z_t \mid \boldsymbol{p}_t, \Theta_t)$. We consider both (i) stationary settings where $\Theta_{t+1} = \Theta_t$ and (ii) nonstationary settings with a piecewise-smooth transition $p(\Theta_{t+1} \mid \Theta_t)$ (e.g., wind/source changes). Since Θ_t is unobserved, the policy conditions on the belief $b_t(\Theta) = p(\Theta_t \mid o_{1:t})$ maintained by particle filtering. Finally, instead of assuming dense external rewards, we define an *intrinsic termination reward* based on posterior uncertainty, aligning learning/planning with the ISLC objective. Objective: Therefore, our objective is to infer the posterior distribution $p(\Theta \mid o_{1:t})$ of the field-model parameters$\Theta = [x_s, y_s, q_s, u_x, u_y, \alpha, \lambda]^\top$ from sequential observations $o_{1:t}$ and to use this belief to guide action selection.

4 Methodology

Overview. ISLC is formulated as a POMDP in Sec. 3. We propose a belief-driven decision framework with the following loop: *(i) maintain a Bayesian belief over unknown source parameters using a particle approximation; (ii) turn belief contraction into a sparse intrinsic learning signal; (iii) learn a policy that actively queries informative locations.* In this paper, we present the RL-based execution strategy, **ATT-PFRL**, as the proposed method. Planning-style baselines are **evaluated** only in Sec. 5.

4.1 Bayesian Belief Approximation via SIS

State of uncertainty (belief). At step k, the agent observes $o_k = (\boldsymbol{p}_k, z_k)$, where $\boldsymbol{p}_k \in \mathbb{R}^2$ is the agent position and $z_k \in \mathbb{R}$ is the sensor reading. Let Θ denote the latent (unknown) source parameters (e.g., source location and emission strength; the exact parameterization is defined in Sec. 3). We maintain the *filtering posterior* (belief) $b_k(\Theta) := p(\Theta \mid o_{1:k})$, where $o_{1:k}$ is the observation history up to step k. The likelihood can be highly non-linear and multi-modal; keeping a posterior (instead of a single estimate) provides uncertainty information that is useful for termination and for belief-conditioned control. **Particle approximation.** We approximate b_k using N_k weighted particles $\{(\Theta_k^i, w_k^i)\}_{i=1}^{N_k}$, $w_k^i \geq 0$, $\sum_{i=1}^{N_k} w_k^i = 1$, via $b_k(\Theta) \approx \sum_{i=1}^{N_k} w_k^i\,\delta(\Theta - \Theta_k^i)$, where $\delta(\cdot)$ is the Dirac delta. Here, Θ_k^i is a hypothesis of the source parameters and w_k^i is its posterior mass. Unless otherwise stated, N_k can be fixed; we later allow adaptive N_k for efficiency. **Sequential importance sampling (SIS).** The filtering posterior satisfies $p(\Theta_{k+1} \mid o_{1:k+1}) \propto p(o_{k+1} \mid \Theta_{k+1})\,p(\Theta_{k+1} \mid \Theta_k)\,p(\Theta_k \mid o_{1:k})$, where $p(o_{k+1} \mid \Theta_{k+1})$ is the likelihood and $p(\Theta_{k+1} \mid \Theta_k)$ is a (possibly trivial) parameter transition model. Given a proposal distribution $q(\Theta_{k+1} \mid \Theta_k, o_{1:k+1})$, SIS updates the unnormalized weights by

$$\bar{w}_{k+1}^i \;\propto\; w_k^i \;\times\; \frac{p\big(o_{k+1} \mid \Theta_{k+1}^i\big)\,p\big(\Theta_{k+1}^i \mid \Theta_k^i\big)}{q\big(\Theta_{k+1}^i \mid \Theta_k^i, o_{1:k+1}\big)}, \tag{2}$$

followed by normalization $w_{k+1}^i = \bar{w}_{k+1}^i / \sum_{j=1}^{N_k} \bar{w}_{k+1}^j$. This correction accounts for sampling from q rather than directly from the posterior.

Stationary parameters (default in ISLC). In many ISLC episodes, Θ is stationary within an episode. We set $\Theta_{k+1}^i = \Theta_k^i, p(\Theta_{k+1} \mid \Theta_k) = \delta(\Theta_{k+1} - \Theta_k)$, and choose $q(\Theta_{k+1}^i \mid$

$\Theta_k^i, o_{1:k+1}) = \delta(\Theta_{k+1}^i - \Theta_k^i)$. Then (2) reduces to pure Bayesian reweighting: $\bar{w}_{k+1}^i = w_k^i \cdot p\big(o_{k+1} \mid \Theta_k^i\big)$. This is the simplest and most efficient update when parameters do not drift within an episode. **Belief moments used downstream.** We summarize b_k using the weighted mean and covariance:$\boldsymbol{\mu}_k = \sum_{i=1}^{N_k} w_k^i \Theta_k^i, \boldsymbol{\Sigma}_k = \sum_{i=1}^{N_k} w_k^i (\Theta_k^i - \boldsymbol{\mu}_k)(\Theta_k^i - \boldsymbol{\mu}_k)^\top$. We also define a compact uncertainty vector $\mathrm{STD}_k = \sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_k)}$. $\boldsymbol{\mu}_k$ provides a point summary; $\boldsymbol{\Sigma}_k$ (and STD_k) quantifies uncertainty for cessation and for belief-conditioned control. **Convergence and finite-particle issues.** Under standard assumptions (e.g., bounded likelihood and proposal support coverage), SIS converges to the true posterior as $N_k \to \infty$. In practice, with finite N_k, weights may become highly concentrated (degeneracy), and resampling may reduce diversity. This motivates the degeneracy mitigation block in Sec. 4.2.

4.2 Degeneracy Mitigation

Goal of this block. The SIS update is repeated at every decision step; thus we need a stabilizing mechanism that (i) reduces weight degeneracy, (ii) preserves useful particle diversity, and (iii) keeps per-step overhead manageable. We implement a unified *stabilize–accelerate* block consisting of: **(1)** ESS-triggered resampling, **(2)** feature-aware *sparse* attention smoothing as a controlled perturbation on weights, and **(3)** MH-corrected MCMC rejuvenation whose Markov kernel leaves the filtering posterior invariant. We also allow the particle budget N_k to be adapted in later stages for efficiency.
(1) ESS-triggered resampling. Given normalized weights $\{w_k^i\}_{i=1}^{N_k}$, we compute effective sample size $\mathrm{ESS}_k = \frac{1}{\sum_{i=1}^{N_k}(w_k^i)^2}$. If $\mathrm{ESS}_k < \eta N_k$, where $\eta \in (0,1)$ is a threshold ratio, we resample (systematic or multinomial) and reset weights to $1/N_k$. ESS quantifies the degree of degeneracy; resampling re-allocates particles toward high-likelihood regions when only a few particles carry most of the mass.
(2) Feature-aware sparse attention smoothing. Resampling alone can reduce diversity, and using scalar weights as $Q{=}K{=}V{=}w$ ignores particle content. We therefore compute attention using particle features. For each particle i, define a feature vector $\boldsymbol{f}_k^i = \big[\Theta_k^i;\ \log p(o_k \mid \Theta_k^i);\ \log(w_k^i + \epsilon_{\mathrm{num}})\big]$, where $\epsilon_{\mathrm{num}} > 0$ is a small constant for numerical stability. We embed it via a shared projection $g(\cdot)$ (e.g., an MLP): $\boldsymbol{e}_k^i = g(\boldsymbol{f}_k^i) \in \mathbb{R}^d$, where d is the embedding dimension.
Sparse attention construction. For each particle i, we build a neighbor set $\mathcal{S}_i$ using approximate m-nearest neighbors (mNN) in the embedding space, where m is a sparsity budget. We compute softmax attention only over $\mathcal{S}_i$: $\alpha_{ij}^{(s)} \propto \exp\Big((\boldsymbol{e}_k^i)^\top \boldsymbol{e}_k^j / \sqrt{d}\Big)$ s.t. $j \in \mathcal{S}_i, \alpha_{ij}^{(s)} = 0$ if $j \notin \mathcal{S}_i$, and $\sum_{j=1}^{N_k} \alpha_{ij}^{(s)} = 1$. We optionally expand $\mathcal{S}_i$ until the retained probability mass exceeds $1-\delta$, where $\delta \in (0,1)$ bounds the dropped tail mass per row. Let $A_k^{(s)} \in \mathbb{R}^{N_k \times N_k}$ denote the resulting row-stochastic attention matrix with $(A_k^{(s)})_{ij} = \alpha_{ij}^{(s)}$.
Attention smoothing on the simplex. We apply a convex smoothing operator on weights:$\tilde{\boldsymbol{w}}_k = (1-\varepsilon)\boldsymbol{w}_k + \varepsilon\big(A_k^{(s)}\big)^\top \boldsymbol{w}_k, \varepsilon \in [0,1]$, followed by normalization. The update mixes weight mass along feature-similar particles; this can reduce extreme concentration while remaining a controlled perturbation of the original weights.
Proposition 1 (Simplex preservation). If $\boldsymbol{w}_k$ lies on the probability simplex and $A_k^{(s)}$ is row-stochastic, then $\tilde{\boldsymbol{w}}_k$ lies on the probability simplex.
Proposition 2 (Controlled sparse approximation). If the dropped tail mass per row is at most δ, then the deviation between dense and sparse smoothing satisfies $\|\tilde{\boldsymbol{w}}_k^{\mathrm{dense}} - \tilde{\boldsymbol{w}}_k^{\mathrm{sparse}}\|_1 \le 2\varepsilon\delta$ (see App. for details).
(3) MH-corrected MCMC rejuvenation. Using $\tilde{\boldsymbol{w}}_k$, we compute $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\Sigma}_k \leftarrow \boldsymbol{\Sigma}_k + \epsilon_{\mathrm{num}}\mathbf{I}$ for numerical stability. To mitigate resampling-induced impoverishment while keeping the update statistically principled, we apply an MH-corrected *move* step that targets the filtering posterior $\pi_k(\Theta) = p(\Theta \mid o_{1:k}) \propto p(\Theta)\prod_{t=1}^k p(o_t \mid \Theta)$. In implementation, rejuvenation is invoked when degeneracy handling is active (e.g., after ESS-triggered resampling), and we run N_{ESS} MH moves per particle, where N_{ESS} is a small fixed integer. Resampling duplicates particles; MH moves can restore diversity while maintaining the target posterior as the stationary distribution. **Target evaluation.** We evaluate the MH ratio in log-space using the unnormalized log-density $\log\tilde{\pi}_k(\Theta) = \log p(\Theta) + \sum_{t=1}^k \log p(o_t \mid \Theta)$, by replaying the stored within-episode history $o_{1:k}$; thus each proposal requires $O(k)$ likelihood evaluations and the total overhead is $O(N_k N_{ESS})$ per rejuvenation call. **Proposals.** We use a Gaussian random-walk proposal $\Theta_{\mathrm{new}}^i = \Theta_k^i + h_{\mathrm{rw}}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}, \boldsymbol{\xi} \sim \mathcal{N}(0,\mathbf{I})$, where $h_{\mathrm{rw}} > 0$ is a step size. When $\nabla_\Theta \log p(o_t \mid \Theta)$ is available, we also consider a Langevin-style proposal with an attention-weighted direction $x_k^i = \sum_{j=1}^{N_k} \alpha_{ij}^{(s)} \nabla_\Theta \log p(o_k \mid \Theta_k^j)$, $\Theta_{\mathrm{new}}^i = \Theta_k^i + \frac{h^2}{2}\boldsymbol{\Sigma}_k x_k^i + h\,\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}, \boldsymbol{\xi} \sim \mathcal{N}(0,\mathbf{I})$, where $h > 0$ is a step size. We monitor the MH acceptance rate and tune (h_{rw}, h) to avoid degenerate accept/reject behavior. **MH correction.** Each proposal is accepted with probability $a\big(\Theta_k^i \to \Theta_{\mathrm{new}}^i\big) = \min\Big\{1,\ \frac{\tilde{\pi}_k(\Theta_{\mathrm{new}}^i)\, q(\Theta_k^i \mid \Theta_{\mathrm{new}}^i)}{\tilde{\pi}_k(\Theta_k^i)\, q(\Theta_{\mathrm{new}}^i \mid \Theta_k^i)}\Big\}$, where $q(\cdot \mid \cdot)$ is the proposal density. The resulting Markov kernel leaves π_k invariant.
Proposition 3 (Posterior invariance). Assuming exact evaluation of $\tilde{\pi}_k(\Theta)$ up to a normalizing constant, the MH acceptance rule above enforces detailed balance with respect to $\pi_k(\Theta) = p(\Theta \mid o_{1:k})$; therefore, the rejuvenation kernel preserves posterior invariance.

4.3 Cessation and Belief-to-Policy Learning

Cessation and intrinsic reward. We stop an episode when posterior uncertainty is sufficiently small. We compute $\mathrm{STD}_k = \sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_k)}$. Termination occurs when $\|\mathrm{STD}_k\|_\infty < \zeta$, where $\zeta > 0$ is a user-defined threshold. Because ISLCENV does not provide shaped external rewards, we define a sparse, goal-aligned intrinsic signal: $r_{k+1} = \mathbb{I}\left[\|\mathrm{STD}_{k+1}\|_\infty < \zeta\right]$. This reward directly reflects the ISLC objective (localize by reducing posterior uncertainty) without introducing additional hand-crafted shaping terms. **Belief-conditioned control.** We define a belief-augmented state for control: $s_k = [\boldsymbol{p}_k, z_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k], \boldsymbol{\sigma}_k = \sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_k)}$. ATT-

PFRL learns a stochastic policy $\pi_\theta(a_k \mid s_k)$ that maximizes the discounted return $\max_{\pi_\theta} \mathbb{E}[\sum_{t=0}^{\infty} \gamma_{RL}^t r_{k+t+1}]$, where $\gamma_{RL} \in (0,1)$ is the discount factor. We employ an actor-critic algorithm for discrete actions: the critic estimates $V_\psi(s_k)$ and the TD error is $\delta_k = r_{k+1} + \gamma_{RL} V_\psi(s_{k+1}) - V_\psi(s_k)$, with standard gradient updates for θ and ψ. Pseudocode is provided in Alg 1.

Algorithm 1 ATT-PFRL Core Loop (Inference + Execution)

Initialize agent position p_0, step $k \leftarrow 0$
Initialize particles $\{(\Theta_i^0, w_i^0)\}_{i=1}^N \sim p(\Theta)$, $w_i^0 \leftarrow 1/N$
while $k < K_{\max}$ **and** $\|\sigma_k\|_\infty \geq \zeta$ **do**
  Observe $o_k = (p_k, z_k)$
  for $i = 1, \ldots, N$ **do**
    $\Theta_i^k \leftarrow \Theta_i^{k-1}$ (or $\sim p(\Theta|\Theta_i^{k-1})$)
    $\bar{w}_i^k \leftarrow w_i^{k-1}\, p(z_k|p_k, \Theta_i^k)$
  end for
  Normalize w^k
  Compute ESS; if low then resample ($w^k \leftarrow 1/N$)
  (ATT) $w^k \leftarrow (1-\varepsilon)w^k + \varepsilon(A^{(s)})^\top w^k$; normalize
  (MH) if resampled then rejuvenate particles
  Compute μ_k, Σ_k; $\sigma_k \leftarrow \sqrt{\mathrm{diag}(\Sigma_k)}$
  $s_k \leftarrow [p_k, z_k, \mu_k, \sigma_k]$, sample $a_k \sim \pi_\theta(\cdot|s_k)$, move
end while
Output μ_k (and Σ_k) as the estimated source parameters.

5 Experiments

Source parameter	Distribution	Source parameter	Distribution
x_s, y_s (location)	$U(5, 20)$	Release strength q_s	$U(10, 3000)$
Wind velocity (u_x, u_y)	$U(0, 6)$	Decay parameter λ	$U(0, 8)$
Diffusivity α	$U(1, 5)$	Wind direction φ_s	$U(0, 2\pi)$

Table 1: Parameter distributions for training scenarios.

In this paper, we utilize the ISLC environments (ISLCenv) to study active source localization under sparse feedback. ISLCenv [Shi *et al.*, 2024] uses a Gaussian model to simulate field distributions and a sensor model to produce intensity observations. The environment provides only positional and intensity observations (no shaped rewards), stressing (i) belief inference under particle degeneracy and (ii) generalizable information-seeking policies.

5.1 Setup: environment, baselines, and metrics

Scenario generation (ID vs. OOD). Training scenarios are defined in a 25×25 area and generated by randomly sampling source/environment parameters at the start of each episode from Table 1. Unless otherwise stated, we set the ESS resampling threshold ratio $\eta = 0.6$ (Sec. 4.2) and the default cessation threshold $\zeta = 0.5$ (Sec. 4.3). The agent starts uniformly in $(0, 5) \times (0, 5)$ and moves with step size 1. We evaluate each setting on 1,000 unseen test scenarios.

For out-of-distribution (OOD) experiments, the training region is confined to $(10, 15) \times (10, 15)$, while testing regions are $(5, 10) \times (15, 20)$ and $(15, 20) \times (15, 20)$ (Fig. 2b). This spatially disjoint split evaluates whether policies rely on belief-conditioned decision rules rather than memorizing region-specific trajectories. **Fields.** We test seven field modalities: Temperature (Temp.), Concentration (Conc.), Magnetic (Mag.), Electric (Elec.), Gas (Gas), Energy (En.), and Noise (Noise), which differ in signal characteristics and noise patterns. **Baselines.** We compare against two groups: (i) RL+Bayes methods: AGDC-KLD [Shi *et al.*, 2025] and its variants AGDC-ENT and AGDC-EE; (ii) planning+Bayes methods: Infotaxis [Vergassola *et al.*, 2007], Entrotaxis [Hutchinson *et al.*, 2018b], and DCEE [Chen *et al.*, 2021]. A Random policy is included as a lower bound. **Metrics.** We report: **(1) OCE** (Operational Completion Efficacy), the stop rate under each method's stopping rule. For contraction-based methods, k_{stop} is the first step satisfying $\|\mathrm{STD}_{k_{\text{stop}}}\|_\infty < \zeta$ (Sec. 4.3), otherwise $k_{\text{stop}} = K_{\max}$, and $\mathrm{OCE} = \Pr[k_{\text{stop}} < K_{\max}]$. **(2) ADE** (Average Deployment Efficiency), the traveled path length up to k_{stop}, $\mathrm{ADE} = \mathbb{E}\big[\sum_{k=0}^{k_{\text{stop}}-1} \|p_{k+1} - p_k\|_2\big]$. **(3) LPS** (Localization Precision Score), the terminal localization error of the posterior mean, $\mathrm{LPS} = \mathbb{E}\big[\|(\mu^x_{k_{\text{stop}}}, \mu^y_{k_{\text{stop}}}) - (x^*, y^*)\|_2\big]$. We additionally report **REV** (runtime proxy; lower is faster) for completeness, but emphasize OCE/ADE/LPS for task performance.

5.2 Main results: in-distribution performance

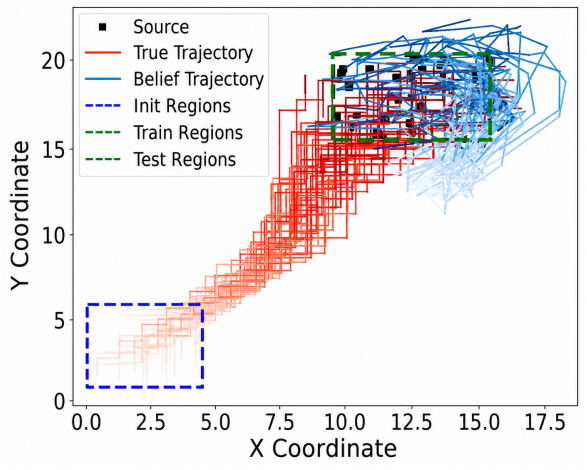
Method	Temp.	Conc.	Mag.	Elec.	Gas	En.	Noise
OCE							
ATT-PFRL	0.95±0.05	0.94±0.05	0.94±0.05	0.82±0.04	0.96±0.05	0.63±0.03	0.94±0.05
AGDC-KLD	0.90±0.05	0.91±0.05	0.89±0.04	0.77±0.04	0.92±0.05	0.61±0.03	0.91±0.05
AGDC-ENT	0.80±0.04	0.81±0.04	0.81±0.04	0.68±0.03	0.79±0.04	0.51±0.03	0.80±0.04
AGDC-EE	0.87±0.04	0.88±0.04	0.86±0.04	0.74±0.04	0.86±0.04	0.57±0.03	0.86±0.04
Infotaxis	0.85±0.04	0.86±0.04	0.85±0.04	0.75±0.04	0.84±0.04	0.55±0.03	0.80±0.04
Entrotaxis	0.24±0.01	0.23±0.01	0.25±0.01	0.15±0.01	0.22±0.01	0.14±0.01	0.23±0.01
DCEE	0.58±0.03	0.59±0.03	0.58±0.03	0.43±0.02	0.56±0.03	0.36±0.02	0.57±0.03
Random	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05
ADE							
ATT-PFRL	20±1.0	19±1.0	18±0.9	19±0.8	17±0.9	19±0.5	19±1.0
AGDC-KLD	23±1.2	22±1.1	22±1.1	23±0.9	20±1.0	22±0.6	21±1.1
AGDC-ENT	25±1.3	24±1.2	25±1.2	20±1.0	22±1.1	24±0.7	23±1.2
AGDC-EE	25±1.3	24±1.2	24±1.2	19±1.0	21±1.1	23±0.7	22±1.1
Infotaxis	50±2.5	48±2.4	51±2.4	58±1.9	43±2.2	47±1.4	45±2.3
Entrotaxis	62±3.1	60±3.0	59±3.0	61±2.5	56±2.8	55±1.8	58±2.9
DCEE	57±2.9	55±2.8	54±2.7	55±2.3	51±2.6	57±1.6	53±2.7
Random	>150	>150	>150	>150	>150	>150	>150
REV							
ATT-PFRL	0.15±0.08	0.14±0.07	0.14±0.07	0.12±0.06	0.14±0.07	0.09±0.05	0.13±0.07
AGDC-KLD	0.10±0.05	0.10±0.05	0.10±0.05	0.09±0.05	0.10±0.05	0.07±0.04	0.10±0.05
AGDC-ENT	0.10±0.05	0.10±0.05	0.10±0.05	0.09±0.05	0.10±0.05	0.07±0.04	0.10±0.05
AGDC-EE	0.10±0.05	0.10±0.05	0.10±0.05	0.09±0.05	0.10±0.05	0.07±0.04	0.10±0.05
Infotaxis	1.50±0.08	1.40±0.07	1.40±0.07	1.20±0.06	1.30±0.07	0.80±0.04	1.30±0.07
Entrotaxis	1.40±0.07	1.30±0.07	1.30±0.07	1.10±0.06	1.20±0.06	0.70±0.04	1.20±0.06
DCEE	1.40±0.07	1.30±0.07	1.30±0.07	1.10±0.06	1.20±0.06	0.70±0.04	1.20±0.06
Random	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
LPS							
ATT-PFRL	0.05±0.01	0.05±0.01	0.05±0.01	0.08±0.01	0.05±0.01	0.06±0.01	0.05±0.01
AGDC-KLD	0.20±0.01	0.20±0.01	0.20±0.01	0.17±0.01	0.20±0.01	0.13±0.01	0.20±0.01
AGDC-ENT	0.25±0.01	0.24±0.01	0.24±0.01	0.20±0.01	0.22±0.01	0.14±0.01	0.23±0.01
AGDC-EE	0.23±0.01	0.22±0.01	0.22±0.01	0.18±0.01	0.20±0.01	0.12±0.01	0.21±0.01
Infotaxis	0.60±0.03	0.60±0.03	0.60±0.03	0.51±0.02	0.60±0.03	0.39±0.02	0.60±0.03
Entrotaxis	0.70±0.04	0.70±0.04	0.70±0.04	0.60±0.03	0.70±0.04	0.46±0.02	0.70±0.04
DCEE	0.60±0.03	0.60±0.03	0.60±0.03	0.51±0.02	0.60±0.03	0.39±0.02	0.60±0.03
Random	5.0±0.25	5.0±0.25	5.0±0.25	5.0±0.25	5.0±0.25	5.0±0.25	5.0±0.25

Table 2: Comparison under in-distribution (ID) scenarios.

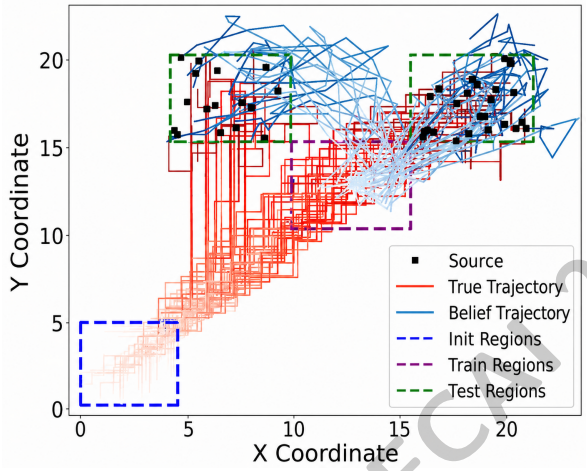
Table 2 provides end-to-end evidence that ATT-PFRL improves both task completion and localization accuracy across all field modalities. High OCE indicates that the learned policy can reliably trigger termination, while low LPS indicates that termination typically occurs near the true source rather than due to accidental contraction.The large ADE gap vs. planning baselines suggests that hand-crafted heuristics

can be myopic under noise/model mismatch, while belief-conditioned RL learns more efficient information gathering.

5.3 Generalization: out-of-distribution



(a) In-distribution (ID)



(b) Out-of-distribution (OOD)

Figure 2: Visualization of test trajectories (100 episodes).

Method	Temp.	Conc.	Mag.	Elec.	Gas	En.	Noise
ATT-PFRL (ours)	0.95	0.94	0.94	0.82	0.96	0.63	0.94
AGDC-KLD	0.448	0.453	0.446	0.382	0.458	0.307	0.452
AGDC-ENT	0.271	0.268	0.273	0.225	0.265	0.168	0.269
AGDC-EE	0.288	0.291	0.285	0.249	0.289	0.192	0.285

Table 3: OOD generalization across fields (spatially shifted regions)

OOD performance drops substantially for RL+Bayes baselines, while ATT-PFRL maintains high completion rates across fields (Table 3). Fig. 2b further shows that belief trajectories continue to converge near the source under spatial shifts, providing a process-level explanation for the aggregated OCE results. Since OOD amplifies inference errors, we next test inference robustness via fixed-trajectory evaluation.

5.4 Uniform mixing lacks attention

Uniform mixing is a special case of convex smoothing on the weight simplex: $\tilde{w}_i = (1-\rho)w_i + \rho\frac{1}{N}, \rho \in [0,1]$, which corresponds to replacing feature-aware attention with a uniform row-stochastic matrix. We test whether the gain of attention is merely due to generic smoothing by sweeping ρ.

ρ	ADE (Basic)	OCE (Basic)	ADE (OOD)	OCE (OOD)
0.0	22±1.1	0.93±0.05	28±1.4	0.88±0.04
0.1	19±1.0	0.95±0.05	24±1.2	0.91±0.05
0.2	18±0.9	0.96±0.05	23±1.2	0.92±0.05
0.3	20±1.0	0.94±0.05	26±1.3	0.89±0.04
0.4	25±1.3	0.90±0.05	32±1.6	0.84±0.04
0.5	32±1.6	0.85±0.04	40±2.0	0.78±0.04

Table 4: Sensitivity to ρ in uniform mixing (convex smoothing).

Convex				Non-convex			
Dist.	OCE	ADE	LPS	Dist.	OCE	ADE	LPS
Beta	0.95±0.05	19±1.0	0.05±0.01	Star Shape	0.92±0.05	22±1.1	0.07±0.01
Gaussian	0.96±0.05	18±0.9	0.04±0.01	Quarter Ring	0.90±0.05	24±1.2	0.08±0.01
Dirichlet	0.94±0.05	20±1.0	0.06±0.01	Half Ring	0.88±0.04	26±1.3	0.09±0.01
Uniform	0.95±0.05	19±1.0	0.05±0.01	Full Ring	0.86±0.04	28±1.4	0.10±0.01

Table 5: Robustness to prior mismatch in the initial belief.

Table 4 shows a sweet-spot behavior: mild uniform mixing ($\rho \approx 0.1$–0.2) slightly improves OCE/ADE by mitigating extreme collapse, but larger ρ biases the posterior toward uniformity and degrades both ID and OOD performance. This control rules out the alternative explanation that "any smoothing" would work; the benefit comes from feature-aware mass transfer that preserves informative modes. Table 5 further demonstrates robustness to prior-shape mismatch (including non-convex priors), supporting that attention regularization does not rely on a fortunate prior geometry and remains stable under diverse belief initializations.

5.5 Adaptation under nonstationary source shifts

We evaluate adaptability in a nonstationary setting where the source location undergoes an abrupt change during inference: after 15 steps, the source position shifts to a new location while the agent continues operating without reset. This tests whether the inference layer can rapidly reallocate posterior mass after a regime change.

Field	OCE		ADE		LPS	
	Ours	AGDC-KLD	Ours	AGDC-KLD	Ours	AGDC-KLD
Temperature	**0.95±0.05**	0.90±0.05	**20±1.0**	23±1.2	**0.05±0.01**	0.20±0.01
Concentration	**0.94±0.05**	0.91±0.05	**19±1.0**	22±1.1	**0.05±0.01**	0.20±0.01
Magnetic	**0.94±0.05**	0.89±0.04	**18±0.9**	22±1.1	**0.05±0.01**	0.20±0.01
Electric	**0.82±0.04**	0.77±0.04	**19±0.8**	23±0.9	**0.08±0.01**	0.17±0.01
Gas	**0.96±0.05**	0.92±0.05	**17±0.9**	20±1.0	**0.05±0.01**	0.20±0.01
Energy	**0.63±0.03**	0.61±0.03	**19±0.5**	22±0.6	**0.06±0.01**	0.13±0.01
Noise	**0.94±0.05**	0.91±0.05	**19±1.0**	21±1.1	**0.05±0.01**	0.20±0.01

Table 6: Adaptability to nonstationary conditions.

Nonstationarity is a stricter stress test than spatial OOD: the agent must abandon an already contracting belief and relocalize without resetting exploration. ATT-PFRL maintains higher completion and lower localization error than AGDC-KLD, suggesting that degeneracy control and rejuvenation help preserve particle diversity, which in turn enables faster

posterior reallocation after regime changes. This observation is consistent with the fixed-trajectory results in Sec. 5.6: higher ESS correlates with stronger recovery from weak or shifting observations in Table 6.

5.6 Inference ablation: fixed-trajectory evaluation

Method	RMSE@20 (m)↓	ESS@20↑	Likelihood Evals↓	Wall-time (ms)↓
PF-baseline	5.2±0.8	120±35	10000	45±8
PF+Attention (R=0)	3.6±0.6	240±50	10500	52±9
PF+MH (ε=0)	4.1±0.7	180±42	15000	68±12
PF+Att+MH (Ours)	**2.8±0.4**	**285±48**	15800	72±11

Table 7: Inference quality on fixed random-walk trajectories.

To isolate the contribution of the *inference layer* (ESS-triggered resampling, feature-aware attention smoothing, and MH rejuvenation) from policy-induced data collection, we compare inference variants under identical observation sequences $o_{1:K}$. We evaluate on the most challenging setting: fixed random-walk trajectories where observations are weakly informative. Table 7 shows that attention smoothing substantially increases ESS and reduces RMSE, indicating that it directly mitigates weight collapse. MH rejuvenation alone yields a smaller gain, while combining attention and MH achieves the best RMSE/ESS, supporting the complementarity claimed in Sec. 4.2: attention prevents extreme collapse, and MH restores diversity after resampling. This provides a causal explanation for the end-to-end improvements in Table 2 and the robustness observed under OOD tests.

5.7 Cessation & reward validation

Our framework uses belief contraction both as a termination criterion and as an intrinsic learning signal: termination occurs when $\|\mathrm{STD}_k\|_\infty < \zeta$, and $r_{k+1} = \mathbb{I}[\|\mathrm{STD}_{k+1}\|_\infty < \zeta]$. This subsection validates (i) threshold sensitivity, (ii) calibration against ground-truth accuracy, and (iii) reward design.

Threshold sensitivity with GT-success. To avoid circular evaluation (defining success using the same threshold being swept), we evaluate termination using an independent ground-truth criterion: let $e_{\mathrm{stop}} = \|(\mu_x, \mu_y) - (x^\star, y^\star)\|_2$ at termination and define GT-success as $e_{\mathrm{stop}} < \epsilon_{\mathrm{loc}}$ with $\epsilon_{\mathrm{loc}} = 1.0$m. We sweep ζ only at test time, keeping policy and inference fixed.

ζ	Stop Rate ↑	GT-Success ↑	Avg. Steps ↓	ADE ↓	False Stop ↓
0.2	0.88	**0.96**	31±6	22±2	0.02
0.3	0.92	**0.95**	28±5	21±2	0.04
0.4	0.94	0.94	24±4	20±1	0.06
0.5	**0.95**	0.92	**22±4**	**20±1**	**0.08**
0.6	0.96	0.89	20±3	19±1	0.12
0.7	0.97	0.86	18±3	19±2	0.15
0.8	0.98	0.81	16±3	18±2	0.21
1.0	0.99	0.75	14±2	18±2	0.28

Table 8: Cessation threshold sweep on Temperature (ID).

As ζ increases, Stop Rate rises monotonically, but GT-Success decreases and False Stop increases, revealing a clear efficiency–accuracy trade-off. The default $\zeta = 0.5$ is a balanced operating point: it achieves strong GT-Success while avoiding excessive exploration (Table 8). This directly supports the claim that our stopping signal is not only effective (high OCE) but also controllable and transferable.

Calibration against ground-truth error. On 1,000 test episodes per field, we record contraction $c_k = \|\mathrm{STD}_k\|_\infty$ and ground-truth error e_k over time. We report Spearman rank correlation, false termination rate at $\zeta = 0.5$, and precision/recall of $\mathbb{I}[c_k < \zeta]$. The contraction signal is consistently and strongly negatively correlated with true localization error across all fields, indicating that posterior contraction is a reliable proxy for accuracy rather than an arbitrary confidence measure. Importantly, this calibration becomes meaningful in light of Sec. 5.6: since the inference layer demonstrably improves RMSE/ESS under controlled observations, the contraction signal is grounded in improved posterior quality, mitigating "confident but wrong" failure modes.

Criterion	GT-Success ↑	Avg. Steps ↓	False Term. ↓	ρ_S ↓
$\|\mathrm{STD}\|_\infty < 0.5$ (ours)	**0.92**	22±4	**0.08**	**-0.78**
Entropy $< \tau_H$	0.87	24±5	0.14	-0.68
$\max_i w_i > \tau_w$	0.84	20±4	0.18	-0.61
Fixed steps (k=25)	0.79	25	0.23	N/A

Table 9: Comparison of termination criteria (Temperature, ID).

We compare contraction-based cessation with entropy thresholding, max-weight thresholding, and fixed-step stopping on Temperature (ID) in Table 9. This comparison shows that contraction-based cessation achieves the best overall balance: it yields higher GT-Success and lower premature termination than entropy/max-weight criteria.

Reward Type	OCE (ID) ↑	GT-Success (ID) ↑	ADE ↓	OCE (OOD) ↑	Sample Eff. ↓
R1: Indicator (ours)	**0.95**	**0.92**	**20±1**	**0.88**	1200 eps to 0.90
R2: Dense contraction	0.91	0.87	21±1	0.84	**980** eps to 0.90
R3: Entropy reduction	0.88	0.83	23±1	0.80	1450 eps to 0.90

Table 10: Reward design ablation for ATT-PFRL.

Reward formulation ablation. We train ATT-PFRL under three intrinsic reward variants with identical budgets: (R1) indicator (ours), (R2) dense contraction shaping, and (R3) entropy reduction in Table 10. Dense shaping (R2/R3) can speed up early learning but degrades asymptotic performance and OOD generalization, suggesting that optimizing a dense proxy (contraction reduction per step) can misalign the policy with the actual stopping objective. In contrast, the indicator reward (R1) directly matches the cessation goal and therefore preserves goal alignment under shifts, closing the loop between inference quality (Sec. 5.6), and robust end-to-end behavior (Secs. 5.2–5.3).

6 Conclusion

We present ATT-PFRL for active source localization under sparse feedback by coupling belief inference with RL. The attention-augmented particle filter mitigates degeneracy and improves posterior quality. Experiments across diverse fields show higher completion and better localization than RL+Bayes and planning baselines, including OOD. Belief contraction is calibrated as a stopping signal and provides an intrinsic reward without hand-crafted distance shaping.

References

[Balkovsky and Shraiman, 2002] Eugene Balkovsky and Boris I Shraiman. Olfactory search at high reynolds number. *Proceedings of the national academy of sciences*, 99(20):12589–12593, 2002.

[Chen *et al.*, 2021] Wen-Hua Chen, Callum Rhodes, and Cunjia Liu. Dual control for exploitation and exploration (dcee) in autonomous search. *Automatica*, 133:109851, 2021.

[Farrell *et al.*, 2005] Jay A Farrell, Shuo Pang, and Wei Li. Chemical plume tracing via an autonomous underwater vehicle. *IEEE Journal of Oceanic Engineering*, 30(2):428–442, 2005.

[Filippi *et al.*, 2010] Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE, 2010.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[Holley, 1969] Edward R Holley. Unified view of diffusion and dispersion. *Journal of the Hydraulics division*, 95(2):621–632, 1969.

[Hu *et al.*, 2019] Hangkai Hu, Shiji Song, and CL Phillip Chen. Plume tracing via model-free reinforcement learning method. *IEEE transactions on neural networks and learning systems*, 30(8):2515–2527, 2019.

[Hu *et al.*, 2025] Kun Hu, Muning Wen, Xihuai Wang, Shao Zhang, Yiwei Shi, Minne Li, Minglong Li, and Ying Wen. Pmat: Optimizing action generation order in multi-agent reinforcement learning. *arXiv preprint arXiv:2502.16496*, 2025.

[Hutchinson *et al.*, 2018a] Michael Hutchinson, Cunjia Liu, and Wen-Hua Chen. Information-based search for an atmospheric release using a mobile robot: Algorithm and experiments. *IEEE Transactions on Control Systems Technology*, 27(6):2388–2402, 2018.

[Hutchinson *et al.*, 2018b] Michael Hutchinson, Hyondong Oh, and Wen-Hua Chen. Entrotaxis as a strategy for autonomous search and source reconstruction in turbulent conditions. *Information Fusion*, 42:179–189, 2018.

[Li *et al.*, 2022] Zhongguo Li, Wen-Hua Chen, and Jun Yang. Concurrent active learning in autonomous airborne source search: Dual control for exploration and exploitation. *IEEE Transactions on Automatic Control*, 68(5):3123–3130, 2022.

[Lillicrap, 2015] TP Lillicrap. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[Lochmatter *et al.*, 2008] Thomas Lochmatter, Xavier Raemy, Loïc Matthey, Saurabh Indra, and Alcherio Martinoli. A comparison of casting and spiraling algorithms for odor source localization in laminar flow. In *2008 IEEE International Conference on Robotics and Automation*, pages 1138–1143. IEEE, 2008.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Park *et al.*, 2022] Minkyu Park, Pawel Ladosz, and Hyondong Oh. Source term estimation using deep reinforcement learning with gaussian mixture model feature extraction for mobile sensors. *IEEE Robotics and Automation Letters*, 7(3):8323–8330, 2022.

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

[Shi *et al.*, 2024] Yiwei Shi, Kevin McAreavey, Cunjia Liu, and Weiru Liu. Reinforcement learning for source location estimation: A multi-step approach. In *2024 IEEE International Conference on Industrial Technology (ICIT)*, pages 1–8. IEEE, 2024.

[Shi *et al.*, 2025] Yiwei Shi, Muning Wen, Qi Zhang, Weinan Zhang, Cunjia Liu, and Weiru Liu. Autonomous goal detection and cessation in reinforcement learning: A case study on source term estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 738–745, 2025.

[Steiner and Bushe, 2001] H Steiner and WK Bushe. Large eddy simulation of a turbulent reacting jet with conditional source-term estimation. *Physics of Fluids*, 13(3):754–769, 2001.

[Vergassola *et al.*, 2007] Massimo Vergassola, Emmanuel Villermaux, and Boris I Shraiman. 'infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409, 2007.

[Zhao *et al.*, 2022] Yong Zhao, Bin Chen, XiangHan Wang, Zhengqiu Zhu, Yiduo Wang, Guangquan Cheng, Rui Wang, Rongxiao Wang, Ming He, and Yu Liu. A deep reinforcement learning based searching method for source localization. *Information Sciences*, 588:67–81, 2022.