Adapting Actively on the Fly: Relevance-Guided Sequential Learning with Latent Concepts for Geospatial Discovery

Jowaria Khan[1] **Anindya Sarkar**[2] **Yevgeniy Vorobeychik**[2] **Elizabeth Bondi-Kelly**[1]

[1]University of Michigan, Ann Arbor, MI, USA
[2]Washington University in St. Louis, St. Louis, MO, USA

Abstract

In environmental monitoring, data collection is often costly, sparse, and shaped by urgent public-health needs. This is particularly true for cancer-causing PFAS (Per- and polyfluoroalkyl substances) contamination, where discussions with domain experts and environmental organizations highlight the need to strategically identify high-risk, under-observed regions under tight sampling budgets. More broadly, similar challenges arise in disaster response and public health settings, where dynamic environments make it essential to efficiently uncover hidden targets from limited ground truth. Yet sparse and biased geospatial labels limit the applicability of existing learning-based methods, such as reinforcement learning. To address this, we propose a unified geospatial discovery framework that integrates active learning, online adaptation, and concept-guided reasoning to support strategic data collection under limited budgets in changing environments. Our approach introduces two key innovations built on a shared notion of *concept relevance*, which captures how domain-specific factors influence target presence: (i) a *concept-weighted uncertainty sampling strategy*, where uncertainty is modulated by learned relevance from readily available concepts such as land cover and source proximity; and (ii) a *relevance-aware meta-batch formation strategy* that promotes semantic diversity during online updates, improving generalization in dynamic environments. We evaluate our framework on PFAS contamination discovery as a real-world-inspired environmental monitoring task, with additional experiments on sparse-supervision land-cover prediction to examine cross-domain applicability. Together, these demonstrate effective geospatial discovery under limited data and changing conditions.

1 Introduction

Discovering targets of interest in large, costly-to-sample, dynamic spaces is a core challenge across environmental monitoring, disaster response, public health, and other geospatial tasks, where targets may include pollution hotspots, damaged regions, or areas at risk of disease [Bondi et al., 2022, Sarkar et al., 2024b]. This challenge is central to PFAS contamination monitoring, where discussions with domain experts motivate the need to identify high-risk, under-observed regions despite sparse and expensive ground truth. Moreover, PFAS contamination, like socioeconomic conditions such as poverty, is not directly observable in conventional satellite imagery and must instead be inferred from indirect environmental or spatial indicators [Leung et al., 2023, Ayush et al., 2020]. Decision-making in such environments must therefore be guided by uncertain and sparse prior data, balancing exploration to gather new information with exploitation to focus on regions most likely to contain areas of interest, like pollution hotspots.

Prior geospatial search methods [Sarkar et al., 2023a, 2024b] often rely on RL or POMDP-based approaches [Kaelbling et al., 1998, Ross et al., 2008], which can require large numbers of interactions

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

or extensive labeled data [Lattimore et al., 2013, Zhang et al., 2023, Schulman et al., 2017]. This is impractical when field labels are costly, and acquisition budgets are small, e.g., $\sim$100 samples per year in environmental monitoring [US EPA, 2018]. Although non-stationary bandit methods exist, adapting to changing conditions while exploiting high-dimensional spatial and semantic structure remains challenging [Li et al., 2010, Bubeck and Cesa-Bianchi, 2012, Ban et al., 2024, Garivier and Moulines, 2011]. Related lifelong learning methods [Zhu et al., 2025, Vogelstein et al., 2025, Mendez-Mendez et al., 2023] address non-stationarity but often assume repeated access to prior examples, replay, or explicit task boundaries [Kirkpatrick et al., 2017, Shin et al., 2017]. However, as environments evolve, older observations may become less relevant, making indefinite replay both impractical and potentially counterproductive. Conventional episodic meta-learning methods [Finn et al., 2017, 2018, 2019] also commonly rely on predefined task structures and support/query splits, which are difficult to establish when labels are acquired sequentially and task boundaries are not explicit. Unlike static pool-based active learning, candidate regions and their features may be visible, but labels must be acquired sequentially as the underlying environment changes. Therefore, identifying high-utility samples requires novel adaptive strategies. We formalize these challenges through our newly proposed *Open-World Learning for Geospatial Prediction and Sampling (OWL-GPS)* problem setting.

OWL-GPS Assumptions

(OWL-GPS) operates under three key constraints: **(i)** labels are collected sequentially through querying, while candidate regions and their input features are available beforehand, **(ii)** the environment is non-stationary, such that the relevance of previously collected observations may change over time, and **(iii)** sampling budgets are strictly limited during both training and inference.

These constraints motivate strategies that jointly guide data acquisition and adaptation as new observations become available.

In light of this challenging setting, our key contributions are summarized as follows:

- We define the **Open-World Learning for Geospatial Prediction and Sampling (OWL-GPS)** setting, characterized by policy-driven sampling over a visible candidate pool, strict budgets, evolving distributions, and bounded-memory online adaptation.
- We propose a **concept-guided relevance encoder** based on a Conditional Variational Auto-Encoder conditioned on domain-informed concept variables.
- We develop a **relevance-aware meta-training strategy** that forms high-utility meta-batches online using uncertainty and conceptual dissimilarity, enabling adaptation in spatially continuous, non-stationary settings without static buffers or fixed episodes.
- We validate our framework on a **real-world discovery task**, *PFAS hotspot detection* using a **novel dataset**, and a **generalization experiment**, *rare land cover identification*, demonstrating robustness under sparse supervision and distribution shift, setting an OWL-GPS benchmark.

2 Problem Formulation

We consider sequential target discovery under budget-constrained label acquisition. Let $\mathcal{U}_t$ denote the set of currently unobserved candidate regions available at acquisition step t. Candidate inputs are available for scoring before acquisition, while their ground-truth labels are revealed only after they are queried. Thus, the sequential constraint applies to label acquisition and model adaptation rather than to candidate visibility.

Each candidate region $x \in \mathcal{U}_t$ contains P spatial locations. A predictive model

$$f_{\theta_t}(x)_i \in [0, 1] \tag{1}$$

estimates the probability that location i belongs to the target class, where θ_t denotes the model parameters at acquisition step t.

At each step, an acquisition policy π_t^{acq} selects one of the currently unobserved candidates:

$$q_t = \pi_t^{\text{acq}}\left(\mathcal{U}_t, f_{\theta_{t-1}}, \mathcal{M}_{t-1}, B_t\right), \tag{2}$$

where $\mathcal{M}_{t-1}$ denotes previously acquired information available to the learner and B_t is the remaining acquisition budget.

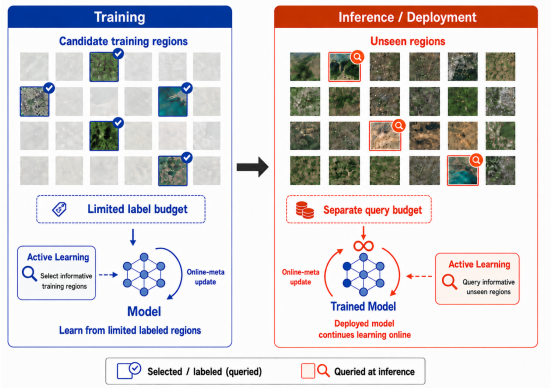


Figure 1: Overview of the OWL-GPS problem formulation.

Querying region q_t incurs acquisition cost $c(q_t)$ and reveals its available ground-truth observation $y^{(q_t)}$. The newly acquired observation can then be used to update the predictive model and the learner's memory before the next acquisition step.

In non-stationary environments, previously acquired observations may become less representative of current conditions as the underlying distribution evolves. Consequently, acquisition decisions must account for both immediate discovery and the value of newly collected supervision for subsequent adaptation.

Budget-constrained discovery objective. Let $\Omega(q_t)$ denote the set of locations in queried region q_t for which ground-truth labels are revealed. For example, in PFAS monitoring, q_t may represent a geospatial image patch, while $\Omega(q_t)$ contains only the pixel locations associated with available field measurements. We distinguish between *acquiring supervision* and *successfully discovering a target*. A queried observation may improve future model updates regardless of the current prediction; however, a target is counted as successfully discovered only when it is truly positive and the model correctly identifies it as positive at the time of acquisition.

We therefore define the discovery objective as

$$\max_{\{\pi_t^{\mathrm{acq}}\}_{t=1}^{T}} \sum_{t=1}^{T} \sum_{i \in \Omega(q_t)} y_i^{(q_t)} \, \mathbb{I}\left[f_{\theta_{t-1}}\left(x^{(q_t)}\right)_i \geq \tau \right] \quad \text{s.t.} \quad \sum_{t=1}^{T} c(q_t) \leq C, \tag{3}$$

where C is the total acquisition budget and τ is the prediction threshold. A positive queried location that is misclassified at step t can still contribute to subsequent learning once its label is revealed, but it does not contribute to the discovery objective at that step.

This objective creates an inherent coupling between acquisition and adaptation. The queries selected by π_t^{acq} determine which supervision becomes available for future updates, while each update of f_{θ_t} changes the information available to subsequent acquisition decisions. The problem becomes particularly challenging when the data distribution also changes across acquisition steps, spatial regions, or environments.

2.1 The OWL-GPS Setting

We refer to this regime as Open-World Learning for Geospatial Prediction and Sampling (OWL-GPS). It combines sequential budget-constrained acquisition, prediction-aware target discovery, and possible spatial or temporal distribution shift. These constraints create a feedback loop: acquisition determines the supervision available for adaptation, while the adapted model determines which observations are most valuable to acquire next.

OWL-GPS specifies the learning problem rather than a particular solution. In Sec. 4, we introduce a relevance-guided approach that couples acquisition and adaptation through a shared learned representation.

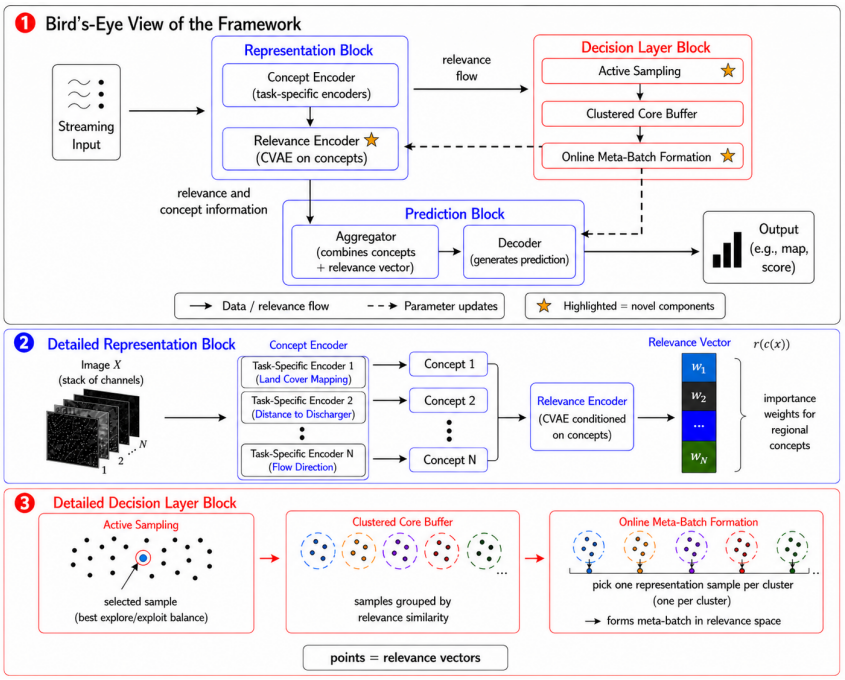


Figure 2: Overview of the proposed relevance-guided sequential learning framework. A shared concept-relevance representation structures both sequential acquisition and online adaptation.

3 Related Work

Active Learning and Sequential Acquisition. Active learning selectively acquires informative labels under limited annotation budgets [Ren et al., 2021, Cacciarelli and Kulahci, 2023]. OWL-GPS similarly scores unlabeled candidates before sequential acquisition, but also adapts the predictive model online as labels are obtained. Our focus is therefore on coupling acquisition with continued model adaptation rather than treating label selection as an isolated decision.

Online, Meta, and Continual Learning. Online learning studies adaptation from sequential observations [Hoi et al., 2021, Shalev-Shwartz, 2012], while continual and lifelong learning address changing data distributions and forgetting [Kirkpatrick et al., 2017, Shin et al., 2017, Vogelstein et al., 2025]. Recent work also considers constrained regimes with limited or no replay and without explicit task boundaries [Jung et al., 2020, Aghasanli et al., 2025]. Meta-learning instead focuses on rapid adaptation from limited observations [Finn et al., 2017, Gharoun et al., 2024], often through explicit task distributions or episodic training. Our setting differs in that the observations available for adaptation are themselves acquired sequentially. Motivated by online meta-learning, our method uses the same learned relevance geometry to guide both acquisition and the construction of subsequent meta-batches.

RL-based Active Geospatial Search. Active geospatial search (AGS) has used reinforcement-learning and related sequential decision-making approaches to optimize search policies in spatial domains [Sarkar et al., 2023a, 2024a,b]. Such methods can learn effective exploration strategies by simulating search episodes, but doing so can be difficult when labeled observations or realistic training trajectories are scarce. Our setting shares the sequential discovery objective of AGS, but additionally requires the predictive model itself to adapt online from newly queried observations under bounded replay memory. Thus, the acquisition policy and the evolving predictive model influence one another throughout deployment.

4 Relevance-Guided Sequential Learning

Concept relevance captures how informative domain-specific factors, such as land cover or proximity to potential pollution sources, are for identifying targets in a given region. Our framework learns a region-specific relevance representation $r(x)$ from domain-informed concepts $c(x)$ and uses the

same relevance geometry to guide both acquisition and online adaptation. Figure 2 summarizes the framework.

4.1 Representation Layer: Learning Concept and Relevance

Concept encoder. Let K denote the number of domain-informed concept channels. We pretrain an encoder to obtain concept-specific representations $\tilde{c}_k(x)$ [Blumenfeld, 2023, Kingma et al., 2013] and concatenate them as

$$\tilde{c}(x) = [\tilde{c}_1(x), \ldots, \tilde{c}_K(x)].$$

The concept channels encode structured factors relevant to the underlying discovery task. For example, in environmental monitoring, they may capture land cover, hydrological structure, or proximity to potential sources. To reduce redundancy between concept representations, we apply Gram–Schmidt orthogonalization [Leon et al., 2013],

$$c(x) = \mathrm{GS}(\tilde{c}(x)),$$

and use $c(x)$ as input to the relevance model. Architecture and pretraining details are in Appendix P.

Probabilistic relevance encoder. Because the importance of individual concepts may vary across regions, we model region-specific relevance as a latent variable using a conditional variational autoencoder (CVAE) [Kingma et al., 2013, Sohn et al., 2015]. Specifically,

$$q_\zeta(r \mid c(x)) = \mathcal{N}\left(\mu_\zeta(c(x)), \mathrm{diag}\left(\sigma_\zeta^2(c(x))\right)\right), \tag{4}$$

where $r(x) \in \mathbb{R}^K$ denotes the region-specific relevance latent aligned with the K concept representations. Using the reparameterization trick,

$$r(x) = \mu_\zeta(c(x)) + \epsilon \odot \sigma_\zeta(c(x)), \qquad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

The target predictor marginalizes over the relevance latent:

$$f_\theta(x) = \mathbb{E}_{r \sim q_\zeta(r \mid c(x))}\left[p_\phi(y \mid c(x), r)\right], \qquad \theta = (\zeta, \phi), \tag{6}$$

where $p_\phi(y \mid c(x), r)$ denotes the decoder distribution over the target conditioned on the concept representation and sampled relevance latent.

We place a standard isotropic Gaussian prior ($p(r) = \mathcal{N}(0, I)$) on the relevance latent variable, and train the relevance model using the conditional-VAE objective

$$\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{KL}} D_{\mathrm{KL}}\left(q_\zeta(r \mid c(x)) \,\|\, p(r)\right), \tag{7}$$

where λ_{KL} governs the expressive capacity of the latent space (where an excessively high value degrades prediction performance, whereas an insufficiently low value distorts the latent geometry), and $\mathcal{L}_{\mathrm{pred}}$ uses the pixel-wise focal loss introduced by [Lin et al., 2018].

4.2 Decision and Adaptation Layer

The learned relevance geometry structures bounded memory, acquisition, and subsequent online updates, coupling the learner's sampling and adaptation decisions.

4.2.1 Relevance-Aware Bounded-Memory Meta-Batches

At acquisition step t, the learner maintains a fixed-capacity memory $\mathcal{M}_t$ consisting of a *core buffer* and a *reservoir buffer*. Observations may be reused while retained in memory, but each buffer follows a finite-lifespan eviction policy, preventing indefinite replay.

Rather than constructing meta-batches uniformly from memory, we organize retained observations in the learned relevance space. This encourages updates to cover distinct relevance regimes while prioritizing observations that have been underused in previous online updates.

For each example x in the core buffer, let $\mathrm{duration}_x$ denote the time since insertion and count_x the number of times it has been selected for online updates. After clustering retained examples in relevance space, we select one representative from each cluster:

$$\mathcal{D}_t^{\mathrm{core}} = \left\{ x_c^\star \;\middle|\; x_c^\star = \arg\max_{x \in \mathcal{E}_c} \exp\left(\frac{\mathrm{duration}_x}{\mathrm{count}_x + 1}\right),\ c = 1, \ldots, N_c \right\}, \tag{8}$$

where $\mathcal{E}_c$ denotes the examples assigned to cluster c and N_c is the number of clusters.

When the core buffer reaches capacity, eligible observations are transferred to a reservoir buffer. We sample K_r reservoir examples without replacement according to

$$\mathcal{D}_t^{\text{res}} \sim \text{Sample}_{K_r}^{\text{w/o replace}} \left(\left\{ \frac{\exp\left(\frac{\text{duration}_i}{\text{count}_i+1}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\text{duration}_j}{\text{count}_j+1}\right)} \right\}_{i=1}^{N} \right). \tag{9}$$

The final meta-batch is constructed as follows:

$$\mathcal{D}_t = \mathcal{D}_t^{\text{core}} \cup \mathcal{D}_t^{\text{res}}. \tag{10}$$

The reservoir itself is also capacity-limited and applies lifespan-based eviction.

4.2.2 Relevance-Guided Acquisition

Training-time acquisition. Let $\mathcal{C}_{t-1}$ denote the core buffer at the beginning of acquisition step t. We measure a candidate's novelty relative to this buffer as

$$w_t(x) = \sum_{x' \in \mathcal{C}_{t-1}} \|\mu_\zeta(c(x)) - \mu_\zeta(c(x'))\|_2^2, \tag{11}$$

where novelty is computed against the core buffer only. Training-time acquisition combines this relevance-space novelty with predictive uncertainty:

$$q_t = \arg\max_{x \in \mathcal{U}_t} [\exp(w_t(x)) \sum_{i=1}^{P} \exp\big(-\big|f_{\theta_{t-1}}(x)_i - 0.5\big|\big)]. \tag{12}$$

Deployment-time acquisition. At deployment, we first compute an exploitation score, which prioritizes regions likely to contain targets for an unobserved region x at step t as follows:

$$S_{\text{exploit}}(x) = \exp(-w_t(x)) \sum_{i=1}^{P} \exp\big(f_{\theta_{t-1}}(x)_i\big), \tag{13}$$

We also compute an exploration score, which promotes both diversity and uncertainty:

$$S_{\text{explore}}(x) = \exp(w_t(x)) \sum_{i=1}^{P} \exp\big(-\big|f_{\theta_{t-1}}(x)_i - 0.5\big|\big). \tag{14}$$

To guide selection, we combine these scores using a budget-aware trade-off:

$$a_t(x) = \kappa_t S_{\text{explore}}(x) + (1 - \kappa_t) S_{\text{exploit}}(x), \tag{15}$$

with $\kappa_t = \frac{C-t}{C+t}$, which allows exploration to decrease as the query budget is consumed. We select

$$q_t = \arg\max_{x \in \mathcal{U}_t} a_t(x). \tag{16}$$

4.2.3 Relevance-Structured Online Adaptation

We consider a general online meta-learning formulation [Finn et al., 2019] for adapting the predictive model after sequential acquisitions, with practical considerations under sparse supervision discussed in Appendix U. The difference from uniform online meta-updating is that the update set $\mathcal{D}_t$ is explicitly structured using the relevance geometry defined above. In this bilevel formulation, $\mathcal{D}_t$ is split as $\mathcal{D}_t^{\text{tr}} \cup \mathcal{D}_t^{\text{val}}$. The inner update takes the form

$$\theta_t' = \theta_{t-1} - \alpha \nabla_\theta \mathcal{L}\left(\theta_{t-1}; \mathcal{D}_t^{\text{tr}}\right). \tag{17}$$

The corresponding meta-update is expressed as

$$\tilde{\theta}_t = \theta_{t-1} - \beta \nabla_{\theta_{t-1}} \mathcal{L}\left(\theta_t'; \mathcal{D}_t^{\text{val}}\right). \tag{18}$$

Adaptation to the newly acquired observation can then be expressed as

$$\theta_t = \tilde{\theta}_t - \eta \nabla_\theta \mathcal{L}\left(\tilde{\theta}_t; (x^{(q_t)}, y^{(q_t)})\right). \tag{19}$$

The queried observation is incorporated into the core buffer, with memory transitions and eviction governed by the bounded-memory mechanism in Sec. 4.2.1. Consequently, acquisition alters the observations available for future updates, while adaptation changes the predictions and relevance estimates used by future acquisition decisions.

Optimization Motivation Although the full objective is non-convex and nonstationary, a local analysis under L-smoothness and a Polyak–Łojasiewicz condition yields

$$\mathbb{E}[\mathcal{L}_T(\theta_T) - \mathcal{L}_T(\theta_T^\star)] \leq \mathcal{O}\left(\frac{\sigma_{\text{rel}}^2}{T} + \frac{G_1}{T^2} + \frac{V_T}{T}\right), \tag{20}$$

where θ_T^* is the optimal predictive model parameter at the T-th step, σ_{rel}^2 is the variance of stochastic update gradients, G_1 is the initial local suboptimality, and V_T is the cumulative drift of the local optimum. This motivates relevance-structured batches as a mechanism for reducing noisy or redundant online updates. Full assumptions and derivation are provided in the Appendix section E; Sec. 5.5 evaluates this motivation through empirical gradient variance.

5 Experiments

5.1 Experimental Protocol

We evaluate on PFAS contamination discovery as the primary benchmark and a sparsified land-cover task as a controlled cross-domain portability test.[1] Table 1 summarizes the protocol; additional dataset, preprocessing, and optimization details are provided in the Appendix.

Table 1: Experimental protocol for PFAS and land-cover evaluation.

	PFAS	Land Cover
Ground truth	704 measured sites	sparsified NLCD labels
Patch size	256×256	256×256
Spatial resolution	30 m	30 m
Input channels	33 (2019), 41 (2022)	4 Sentinel-2 bands
Spatial evaluation	disjoint regions	disjoint regions
Temporal evaluation	$2019 \rightarrow 2022$	–
Training budget	150	150
Deployment budget	50	50
Evaluation supervision	measured points only	retained sparse pixels

Sparse PFAS supervision. PFAS labels are point measurements. During training, we expand each label over nearby surface-water pixels, excluding non-water pixels; all evaluation remains restricted to measured locations. Sec. 5.4 tests point-only supervision to assess dependence on this expansion.

Metrics. We evaluate budgeted *discovery* using Success Rate (SR) and *predictive performance* using accuracy, macro F-score, precision, and recall. Let $\Omega(q_t)$ denote the valid evaluation locations in queried region q_t. We compute

$$\text{SR} = \frac{1}{C}\sum_{t=1}^{C} \frac{\sum_{i\in\Omega(q_t)} y_i^{(q_t)} \mathbb{I}\big[f_{\theta_{t-1}}(x^{(q_t)})_i \geq \tau\big]}{\max\Big(1, \sum_{i\in\Omega(q_t)} y_i^{(q_t)}\Big)}. \tag{21}$$

SR measures correctly identified targets under the acquisition budget, whereas macro F-score balances prediction across target and non-target classes. PFAS metrics are evaluated only at measured locations.

Baselines and controlled variants. We compare against methods spanning greedy acquisition, uncertainty-based active learning, bandit-style exploration, geospatial foundation-model transfer, and online/meta-learning in Tables 3 and 4 . We additionally include controlled variants of our framework that replace the proposed acquisition or adaptation mechanisms while keeping the remaining components fixed.

All learned methods are evaluated under the same training and deployment query budgets, PFAS supervision protocol, and number of training optimization steps. Framework-controlled variants use the same concept representation and optimization settings. Prithvi retains its own pretrained representation so that the comparison evaluates the published foundation model rather than a reimplementation using our representation.

5.2 PFAS Discovery

We first evaluate spatial generalization on geographically disjoint 2019 PFAS regions.

[1] Code: `https://github.com/jowariak/OWL-GPS-NeurIPS-2026`.

Table 2: Comparison methods used in our experiments.

Method	Description
GA	Greedy acquisition that queries the highest-confidence candidate regions.
AL	Exploration-focused, uncertainty-based active learning that prioritizes candidates with uncertain predictions [Cacciarelli and Kulahci, 2023].
Prithvi	A pretrained geospatial foundation model used for prediction with online updates during acquisition [Blumenfeld, 2023].
UCB	A controlled variant using an **U**pper **C**onfidence **B**ound acquisition policy [Auer et al., 2002].
AML	A controlled **A**ctive **M**eta-**L**earning variant in which meta-batches are formed from latent representations of previously queried examples [Kaddour et al., 2020].
OML	A controlled **O**nline **M**eta-**L**earning variant using continual single-sample meta-adaptation, replacing episodic batch updates [Finn et al., 2019].

Table 3: PFAS discovery on spatially disjoint 2019 regions (mean $\pm$ std. over three seeds).

Method	Acc.	F-score	Prec.	Recall	SR
GA	84 ± 1.5	46 ± 1.3	42 ± 1.4	50 ± 1.7	92 ± 2.0
AL	81 ± 1.8	50 ± 1.5	42 ± 1.3	49 ± 1.6	93 ± 2.1
Prithvi	86 ± 1.6	47 ± 1.3	43 ± 1.4	50 ± 1.5	94 ± 1.8
UCB	79 ± 1.3	61 ± 1.1	61 ± 1.2	61 ± 1.4	94 ± 1.6
AML	86 ± 2.0	46 ± 1.4	43 ± 1.5	50 ± 1.3	95 ± 2.0
OML	83 ± 1.7	54 ± 1.2	56 ± 1.4	53 ± 1.5	94 ± 1.9
Ours	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{78 \pm 2.0}$	$\mathbf{63 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

Our method consistently outperforms all baselines, achieving the highest F-score and SR, with particularly strong recall, indicating improved target recovery under the fixed query budget. Figure 3 illustrates the sequential acquisition behavior: as acquisition proceeds, SR increases while the number of non-target regions queried decreases, consistent with the policy shifting from broader exploration toward target-rich regions.

5.3 Distribution Shift

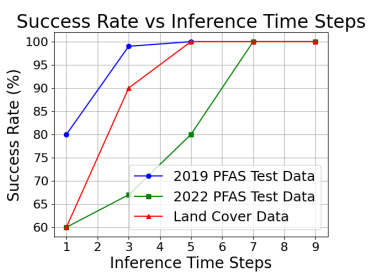
Temporal transfer: 2019 $\rightarrow$ 2022. We next train using 2019 PFAS observations and deploy on 2022 observations in Table 4. This constitutes a challenging temporal-transfer setting because site coverage, environmental conditions, and available input channels differ between years. We evaluate our method against GA, AL, Prithvi, UCB, AML, and OML over five matched seeds and perform paired two-sided tests against UCB, with Holm correction across reported metrics.

Table 4: Temporal transfer from 2019 to 2022 over five matched seeds (mean $\pm$ standard deviation). The final column reports Holm-adjusted p-values against UCB.

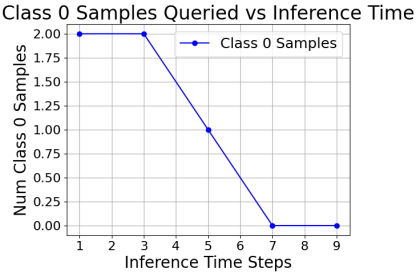
Metric	Ours	UCB	Prithvi	GA	AL	AML	OML	p_{Holm}
SR	$\mathbf{86 \pm 1.4}$	81 ± 1.6	78 ± 1.7	77 ± 2.2	76 ± 1.7	80 ± 2.1	84 ± 1.9	.001
F-score	$\mathbf{55 \pm 1.7}$	52 ± 1.6	52 ± 1.7	46 ± 1.3	44 ± 1.2	46 ± 1.5	51 ± 1.2	.011
Accuracy	$\mathbf{77 \pm 2.0}$	74 ± 2.0	71 ± 1.9	72 ± 1.6	72 ± 1.3	68 ± 2.0	75 ± 1.7	.020
Precision	$\mathbf{56 \pm 1.2}$	53 ± 1.1	53 ± 1.3	43 ± 1.5	39 ± 1.0	45 ± 1.7	52 ± 1.4	.015
Recall	56 ± 2.4	$\mathbf{57 \pm 1.0}$	54 ± 1.1	50 ± 1.4	50 ± 1.5	48 ± 1.4	51 ± 1.5	.420

Our method achieves the highest SR, F-score, accuracy, and precision under temporal shift, with statistically significant improvements over UCB after Holm correction ($p \leq .020$). OML remains competitive in SR (84 vs. 86), while recall is statistically comparable to UCB ($p = .420$). Overall, these results demonstrate improved budgeted discovery and predictive performance under temporal distribution shift.

Controlled geographic shift. To isolate spatial shift while holding the year, channels, and preprocessing fixed, we partition 2019 observations into eastern and western regions. We evaluate East$\rightarrow$West transfer, using one initial labeled observation and a budget of 30 source-region acquisitions for training. We reserve a fixed, class-balanced holdout of 20 western observations (10 per class) for evaluation, which remains excluded from acquisition and adaptation. Before target-region adaptation, the model achieves 58 ± 2.5 F-score, 62 ± 1.8 precision, and 60 ± 1.0 recall. After acquiring five



(a) SR over acquisition steps.



(b) Non-target queries.

Figure 3: Sequential acquisition behavior on the 2019 PFAS benchmark. As the acquisition budget is consumed, discovery improves while queries increasingly concentrate on target-rich regions.

additional western observations and performing one round of online adaptation, these improve to 70 ± 2.5, 71 ± 2.0, and 70 ± 1.4, respectively. For reference, the standard spatial split achieves 65 ± 2.0 F-score, 78 ± 2.0 precision, and 62 ± 2.0 recall. We restrict this analysis to East→West because the eastern region contains only two class-0 observations, making a meaningful class-balanced holdout and evaluation of reverse transfer particularly challenging. Overall, these results indicate that limited target-region adaptation can improve predictive performance under controlled geographic shift.

5.4 Sensitivity to Sparse-Label Expansion

Because our primary PFAS training setup expands point measurements over nearby surface-water pixels, we test whether the observed gains depend on this assumption by retraining with measured point labels only. Evaluation uses measured observations in both settings.

Table 5: Sensitivity to PFAS training supervision.

Training supervision	Ours F1	Best baseline F1	Ours SR	Best baseline SR
Point-to-patch expansion	$\mathbf{65 \pm 2.0}$	61 ± 1.1 (UCB)	$\mathbf{98 \pm 1.0}$	95 ± 2.0 (AML)
Point labels only	$\mathbf{65 \pm 2.0}$	56 ± 2.0 (UCB)	$\mathbf{90 \pm 1.4}$	84 ± 2.0 (AML)

Point-only supervision reduces our SR from 98 to 90, but the relative advantage persists: our method achieves 90 SR versus 84 for the strongest baseline, while its F-score remains 65 compared with 56 for the strongest baseline. Thus, label expansion improves absolute discovery performance but does not explain the gains over competing methods.

Table 6: Sparse-supervision land-cover results (mean $\pm$ std., three seeds).

Method	Acc.	F-score	Prec.	Recall	SR
GA	46 ± 0.8	35 ± 1.0	48 ± 0.9	40 ± 1.3	47 ± 1.4
AL	58 ± 1.8	41 ± 1.5	50 ± 1.2	47 ± 1.9	60 ± 1.0
Prithvi	61 ± 1.2	40 ± 1.4	51 ± 1.5	56 ± 1.7	61 ± 1.8
UCB	55 ± 2.1	45 ± 1.7	53 ± 1.3	58 ± 1.5	55 ± 1.3
AML	91 ± 1.3	48 ± 1.3	46 ± 1.6	50 ± 1.8	90 ± 1.1
OML	92 ± 1.7	48 ± 1.6	47 ± 1.8	50 ± 1.9	92 ± 1.5
Ours	$\mathbf{93 \pm 1.4}$	$\mathbf{71 \pm 1.1}$	$\mathbf{70 \pm 1.3}$	$\mathbf{72 \pm 1.2}$	$\mathbf{94 \pm 1.0}$

5.5 Component Ablations

We replace one component at a time while holding the framework fixed. All replacements reduce F-score, while the largest SR drops occur without relevance-guided acquisition or relevance-aware meta-batching, supporting the use of a shared relevance geometry for both decisions.

Gradient variance. Consistent with Sec. 4.2.3, relevance-aware meta-batches reduce empirical gradient variance from 4.8×10^{-3} under uniform batching to 2.9×10^{-3}. This complements the random-meta-batch ablation in Table 7 and supports the motivation for relevance-structured updates.

5.6 Validating the Learned Relevance Representation

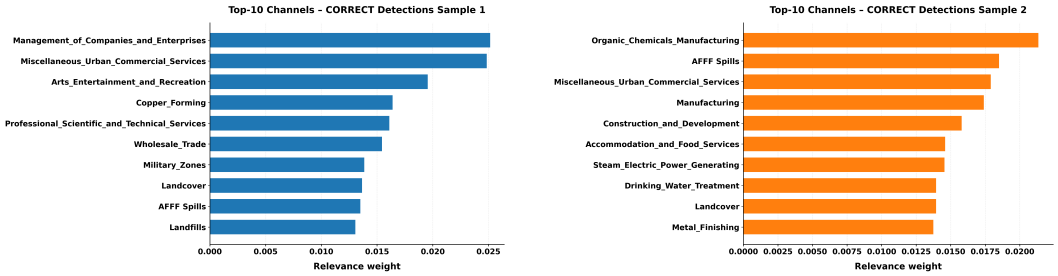
We validate relevance through interventions and deliberately corrupted concepts; qualitative visualizations are provided in the Appendix Q.

Table 7: Component ablations on 2019 PFAS.

Variant	F-score	SR
No relevance encoder	56 ± 1.7	95 ± 2.0
No relevance-guided acquisition	55 ± 1.7	77 ± 1.5
Random meta-batch	56 ± 2.1	69 ± 1.4
GIA $\rightarrow$ DBSCAN	57 ± 1.6	94 ± 1.1
CVAE relevance $\rightarrow$ MLP	56 ± 1.8	94 ± 2.3
No concept orthogonalization	56 ± 1.5	86 ± 0.7
Full framework	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

Intervention by relevance rank. To test whether model predictions depend more strongly on concepts assigned greater relevance, we zero one relevance channel at a time without performing an online update. Zeroing the highest-, median-, and lowest-ranked channels changes the predicted target probability by 0.0058, 0.0045, and 0.0037, respectively. The monotonic decrease with relevance rank indicates that the learned ordering is aligned with predictive dependence.

Context-dependent relevance. Figure 4 compares the highest-ranked concept channels for two correctly predicted regions. The rankings differ across regions despite both predictions being correct, providing qualitative evidence that the learned relevance representation is region-specific rather than a single global concept-importance vector.



(a) Region 1.

(b) Region 2.

Figure 4: Context-dependent relevance profiles for two correctly predicted regions. Different concept rankings across regions illustrate the region-specific nature of the learned relevance representation.

Robustness to uninformative concepts. We spatially permute selected concept channels, preserving their marginal distributions while breaking target alignment. Corrupted channels receive substantially lower relevance than genuine channels, and are ranked near the bottom of the 33 channels (2.9×10^{-4} vs. 3.3×10^{-2}; median rank $28/33$), while performance changes only from 65 to 63 F-score and 98 to 95 SR. These results indicate that the relevance mechanism largely identifies and suppresses concept channels whose spatial information has been made uninformative.

5.7 Cross-Domain Portability: Land Cover

We test cross-domain portability using four-band Sentinel-2 imagery and the water land-cover class as target under artificially sparsified supervision. This controlled experiment changes both the sensor representation and prediction target while retaining the sparse-supervision and sequential-acquisition protocol. Table 6 reports the results. Our method performs best across all metrics, with the largest gain in F-score (71 vs. 48), indicating that the acquisition–adaptation mechanism transfers beyond the PFAS setting.

Conclusion and Limitations We introduce a relevance-guided framework for sequential geospatial discovery under strict acquisition budgets, coupling sampling and online adaptation through a shared concept-relevance representation. Across PFAS monitoring and a controlled sparse-supervision land-cover task, the method improves target discovery under limited labels and distribution shift. The framework relies on meaningful domain concepts, which may limit its applicability when such structure is unavailable.

Acknowledgments and Disclosure of Funding

This work was supported in part by the University of Michigan, the Univeristy of Michigan e-HAIL, National Science Foundation (NSF) grants IIS-2214141 and CCF-2403758, and Office of Naval Research (ONR) award N000142412663.

We thank David Andrews, Sydney Evans, Alexa Friedman, and Kaley Beins from the Environmental Working Group (EWG) for their valuable insights and discussions that helped motivate and inform the PFAS monitoring application of this research.

Competing Interests. The authors declare no competing interests.

References

Agil Aghasanli, Yi Li, and Plamen Angelov. Prototype-based continual learning with label-free replay buffer and cluster preservation loss. In *Proceedings of the CVPR Workshop on Dynamic Graphs, Embeddings, Belief and Feedback (DG-EBF) 2025*, 2025.

Mateus Henrique Amaral, Lira Luz Benites-Lazaro, Paulo Antonio de Almeida Sinisgalli, Humberto Prates da Fonseca Alves, and Leandro Luiz Giatti. Environmental injustices on green and blue infrastructure: Urban nexus in a macrometropolitan territory. *Journal of Cleaner Production*, 289:125829, 2021.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Generating interpretable poverty maps using object detection in satellite images. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, page 4410–4416, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/608. URL `https://www.ijcai.org/proceedings/2020/608`.

Yikun Ban, Yunzhe Qi, and Jingrui He. Neural contextual bandits for personalized recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1246–1249, 2024.

Josh Blumenfeld. Nasa and ibm openly release geospatial ai foundation model for nasa earth observation data — earthdata, August 2023. URL `https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model`.

Elizabeth Bondi, Haipeng Chen, Christopher D. Golden, Nikhil Behari, and Milind Tambe. Micronutrient deficiency prediction via publicly available satellite data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12454–12460, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i11.21512.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Davide Cacciarelli and Murat Kulahci. Active learning for data streams: A survey. *Machine Learning*, 2023. doi: 10.1007/s10994-023-06454-2.

Su Cao, Xuesheng Zhao, and Shouhang Du. Multi-type and fine-grained urban green space function mapping based on bert model and multi-source data fusion. *International Journal of Digital Earth*, 17(1):2308723, 2024.

Ylenia Casali, Nazli Yonca Aydin, and Tina Comes. Machine learning for spatial analyses in urban areas: a scoping review. *Sustainable cities and society*, 85:104050, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International conference on machine learning*, pages 1920–1930. PMLR, 2019.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd international conference on Algorithmic learning theory*, ALT'11, page 174–188, Berlin, Heidelberg, October 2011. Springer-Verlag. ISBN 978-3-642-24411-7.

Luise Ge, Michael Lanier, Anindya Sarkar, Bengisu Guresti, Chongjie Zhang, and Yevgeniy Vorobeychik. Learning policy committees for effective personalization in mdps with diverse tasks. *arXiv preprint arXiv:2503.01885*, 2025.

Anne Gharaibeh, Abdulrazzaq Shaamala, Rasha Obeidat, and Salman Al-Kofahi. Improving land-use change modeling by integrating ann with cellular automata-markov chain model. *Heliyon*, 6(9), 2020.

Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H Gandomi. Meta-learning approaches for few-shot learning: A survey of recent advances. *ACM Computing Surveys*, 56(12):1–41, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021. doi: 10.1016/j.neucom.2021.04.112.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 7990–8001, 2020.

Jean Kaddour, Steindór Sæmundsson, et al. Probabilistic active meta-learning. *Advances in Neural Information Processing Systems*, 33:20813–20822, 2020.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

MD Kamruzzaman, Julian Hine, and Tan Yigitcanlar. Investigating the link between carbon dioxide emissions and transport-related social exclusion in rural northern ireland. *International journal of environmental science and technology*, 12(11):3463–3478, 2015.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

CY Lam, K Tai, and AM Cruz. Topological network and gis approach to modeling earthquake risk of infrastructure systems: a case study in japan. *Applied geography*, 127:102392, 2021.

Tor Lattimore, Marcus Hutter, and Peter Sunehag. The sample-complexity of general reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 28–36. PMLR, 2013.

Steven J Leon, Åke Björck, and Walter Gander. Gram-schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 20(3):492–532, 2013.

Shui Cheung Edgar Leung, Dushanthi Wanninayake, Dechao Chen, Nam-Trung Nguyen, and Qin Li. Physicochemical properties and interactions of perfluoroalkyl substances (pfas) - challenges and opportunities in sensing and remediation. *The Science of the Total Environment*, 905:166764, December 2023. ISSN 1879-1026. doi: 10.1016/j.scitotenv.2023.166764.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Yaoxing Liao, Zhaoli Wang, Xiaohong Chen, and Chengguang Lai. Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model. *Journal of Hydrology*, 624: 129945, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. (arXiv:1708.02002), February 2018. doi: 10.48550/arXiv.1708.02002. URL `http://arxiv.org/abs/1708.02002`. arXiv:1708.02002 [cs].

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

Luiz A Manfré, Eliane Hirata, Janaína B Silva, Eduardo J Shinohara, Mariana A Giannotti, Ana Paula C Larocca, and José A Quintanilha. An analysis of geospatial technologies for risk and natural disaster management. *ISPRS International Journal of Geo-Information*, 1(2):166–185, 2012.

Raveena Marasinghe, Tan Yigitcanlar, Severine Mayere, Tracy Washington, and Mark Limb. Computer vision applications for urban planning: A systematic review of opportunities and constraints. *Sustainable Cities and Society*, 100:105047, 2024.

Jorge Mendez-Mendez, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Embodied lifelong learning for task and motion planning. pages 2134–2150. PMLR, 2023.

Carles Milà, Joan Ballester, Xavier Basagana, Mark J Nieuwenhuijsen, and Cathryn Tonne. Estimating daily air temperature and pollution in catalonia: A comprehensive spatiotemporal modelling of multiple exposures. *Environmental Pollution*, 337:122501, 2023.

Sergio Alberto Monjardin-Armenta, Wenseslao Plata-Rocha, Carlos Eduardo Pacheco-Angulo, Cuauhtémoc Franco-Ochoa, and Jesus Gabriel Rangel-Peraza. Geospatial simulation model of deforestation and reforestation using multicriteria evaluation. *Sustainability*, 12(24):10387, 2020.

Chulwoong Park, Wonjun No, Junyong Choi, and Youngchul Kim. Development of an ai advisor for conceptual land use planning. *Cities*, 138:104371, 2023.

Massimo Regona, Tan Yigitcanlar, Carol Hon, and Melissa Teo. Artificial intelligence and sustainable development goals: Systematic literature review of the construction industry. *Sustainable Cities and Society*, 108:105499, 2024.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021. doi: 10.1145/3472291.

Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-draa. Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research*, 32:663–704, 2008.

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, 2023a.

Anindya Sarkar, Michael Lanier, Scott Alfeld, Jiarui Feng, Roman Garnett, Nathan Jacobs, and Yevgeniy Vorobeychik. A visual active search framework for geospatial exploration. (arXiv:2211.15788), October 2023b. doi: 10.48550/arXiv.2211.15788. URL `http://arxiv.org/abs/2211.15788`. arXiv:2211.15788 [cs].

Anindya Sarkar, Alex DiChristofano, Sanmay Das, Patrick J. Fowler, Nathan Jacobs, and Yevgeniy Vorobeychik. Active geospatial search for efficient tenant eviction outreach. (arXiv:2412.17854), December 2024a. doi: 10.48550/arXiv.2412.17854. URL `http://arxiv.org/abs/2412.17854`. arXiv:2412.17854 [cs].

Anindya Sarkar, Michael Lanier, Scott Alfeld, Jiarui Feng, Roman Garnett, Nathan Jacobs, and Yevgeniy Vorobeychik. A visual active search framework for geospatial exploration. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 8301–8310, January 2024b. doi: 10.1109/WACV57701.2024.00813. URL https://ieeexplore.ieee.org/document/10484136/?arnumber=10484136.

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

Mahyat Shafapourtehrany, Maryna Batur, Farzin Shabani, Biswajeet Pradhan, Bahareh Kalantar, and Haluk Özener. A comprehensive review of geospatial technology applications in earthquake preparedness, emergency management, and damage assessment. *Remote Sensing*, 15(7):1939, 2023.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Tim Heinrich Son, Zack Weedon, Tan Yigitcanlar, Thomas Sanchez, Juan M Corchado, and Rashid Mehmood. Algorithmic urban planning for smart and sustainable development: Systematic review of the literature. *Sustainable Cities and Society*, 94:104562, 2023.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022.

Asha Thapliyal, Sanjeev Kimothi, Ajay Kumar Taloor, Mahendra Pratap Singh Bisht, Pankaj Mehta, and Girish Chandra Kothyari. Glacier retreat analysis in the context of climate change impact over the satopanth (spg) and bhagirathi-kharak (bkg) glaciers in the mana basin of the central himalaya, india: a geospatial approach. *Geosystems and Geoenvironment*, 2(1):100128, 2023.

OW US EPA. Pfas resources, data and tools, January 2018. URL https://www.epa.gov/pfas/pfas-resources-data-and-tools.

OW US EPA. Our current understanding of the human health and environmental risks of pfas, October 2021. URL https://www.epa.gov/pfas/our-current-understanding-human-health-and-environmental-risks-pfas.

Joshua T Vogelstein, Jayanta Dey, Hayden S Helm, Will LeVine, Ronak D Mehta, Tyler M Tomita, Haoyin Xu, Ali Geisa, Qingyang Wang, Gido M Van De Ven, et al. Simple lifelong learning machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zhiyuan Yan, Junxi Li, Xuexue Li, Ruixue Zhou, Wenkai Zhang, Yingchao Feng, Wenhui Diao, Kun Fu, and Xian Sun. Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, et al. Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023.

Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023.

Guilin Zhu, Dongyue Wu, Changxin Gao, Runmin Wang, Weidong Yang, and Nong Sang. Adaptive prototype replay for class incremental semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10932–10940, 2025.

Adapting Actively on the Fly: Relevance-Guided Sequential Learning with Latent Concepts for Geospatial Discovery (Appendix)

Overview of the Contents

A. Effect of $\kappa(\mathcal{C})$ on Search Performance 17

B. Connection between Active Sampling and Variance in Relevance Space 17

C. Evolution of Discriminative Power in Relevance Space as Search Progress 17-18

D. Search Performance Across Varying Search Budget 17-18

E. Derivation of the Local Optimization Bound 18-19

F. On the Robustness of the Concept Encoder 20

G. Influence of Core and Reservoir Buffer Sizes on Search Performance 20

H. Mathematical Motivation for the Training Sampling Objective 20-21

I. Mathematical Motivation for the Inference Sampling Objective 21-22

J. Details of Greedy Intersection Clustering Algorithm 22-24

K. Additional Analysis on the Importance of Relevance Guided Sampling 24

L. Importance of Orthogonalization in the Concept Space 24

M. Dataset and Split Details 24-25

N. Details of Pseudolabel Generation Procedure to Induce Stability During Training . . 25-27

O. Details of Training and Inference Hyperparameters 27

P. Architecture Details 27-28

P.1 Details of Concept Encoder 27

P.2 Details of Relevance Encoder 27-28

P.3 Details of the Decoder 28

P.4 Modeling Spatial and Hydrological Structure 28

Q. Additional Relevance Interpretability Analysis 28-29

R. Additional Details about the Baselines 29-30

S. Extended Related Work 30-31

S.1 Environment Monitoring 30

S.2 Geospatial Foundational Model 30-31

T. Time and Memory Complexity Details of the Proposed Framework 31

T.1 Time Complexity Details 31

T.2 Details about Compute Resource and Code 31

U. Implementation Details and Training/Inference Pseudocode 31-33

V. Experimental Setup of Domain-Informed Greedy Baseline 33-34

W. Impact Statement 34-35

A Effect of $\kappa(\mathcal{C})$ on Search Performance

We conduct experiments to assess the impact of $\kappa(\mathcal{C})$ on the search policy's active search performance. Specifically, we investigate how amplifying the exploration weight by setting $\kappa(\mathcal{C}) = \max\{0, \kappa(\alpha \cdot \mathcal{C})\}$ with $\alpha > 1$, and enhancing the exploitation weight by setting $\alpha < 1$, influence the overall effectiveness of the approach. We present SR results for $\alpha \in \{0.2, 1, 5\}$ as shown in Table 8. Performance improves when moving from an exploitation-heavy setting (α = 0.2) to balanced exploration-exploitation (α = 1.0). Increasing α further yields comparable performance within variance, indicating diminishing returns from overly aggressive exploration. These results suggest that while insufficient exploration degrades performance, the method remains robust to larger α values, with α = 1 providing a strong and stable operating point.

Table 8: Effect of $\kappa(\mathcal{C})$

SR Performance across varying α with $\mathcal{C} = 50$		
$\alpha = 0.2$	$\alpha = 1.0$	$\alpha = 5.0$
$94\% \pm 1.0$	$\mathbf{98\% \pm 1.0}$	$98\% \pm 1.9$

B Connection between Active Sampling and Variance in Relevance Space

In this analysis, we aim to assess the effectiveness of the active sampling strategy. To achieve this, we compute the component-wise variance of the relevance vectors corresponding to the set of samples selected via the proposed active sampling strategy at two different query steps during the active sampling process at inference time. Our observations reveal a progressive decline in batch variance as the search advances, indicating a shift in the model's behavior from exploration to increased exploitation. This behavior is illustrated in Figure 5, 6, which highlight the effectiveness of the proposed sampling strategy.

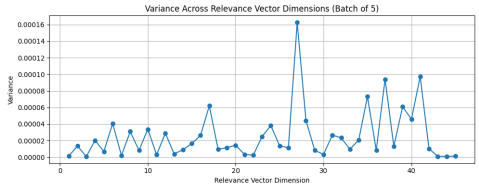


Figure 5: Variance across dimensions of the relevance vector during the initial Active Discovery Phase.

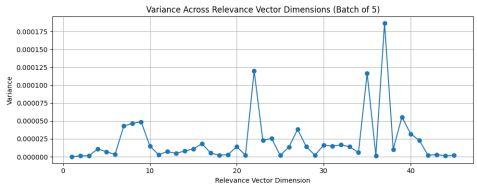


Figure 6: Variance across dimensions of the relevance vector during the later Active Discovery Phase.

C Evolution of Discriminative Power in Relevance Space as Search Progress

To assess the effectiveness of our training paradigm, we investigate how the relevance space evolves throughout the active search process. We project the learned relevance vectors into two dimensions using t-SNE at two distinct training stages. Initially, the embeddings are densely entangled, showing little class-specific structure. Note that in our case, we only have two distinct classes, namely class 0 and class 1, where class 0 refers to the absence of the target region, and class 1 indicates the presence of the target region. However, as more informative samples are acquired through active querying, we observe a clear separation in embeddings corresponding to different input classes. This progressive disentanglement highlights the growing discriminative capacity of the relevance space, validating our approach's ability to refine representations in a data-efficient manner and ensuring increasingly reliable and robust search performance. We present the illustrative visualizations in Figure 7, 8 using two randomly selected samples from two distinct classes.

D Search Performance Across Varying Search Budget

In this section, we present the search performance of our proposed approach under different search budgets, summarized in Table 9. Our results show that increasing the search budget consistently improves performance. As the search budget increases, additional ground truth data becomes

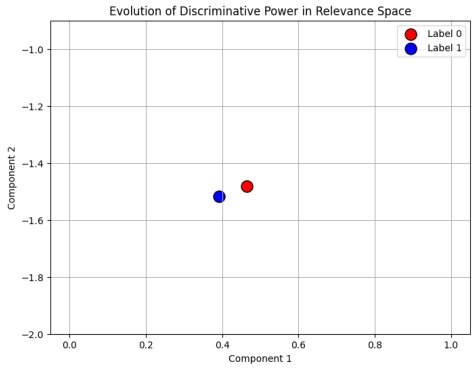


Figure 7: TSNE Visualization of Relevance Vectors during the initial Active Discovery Phase.

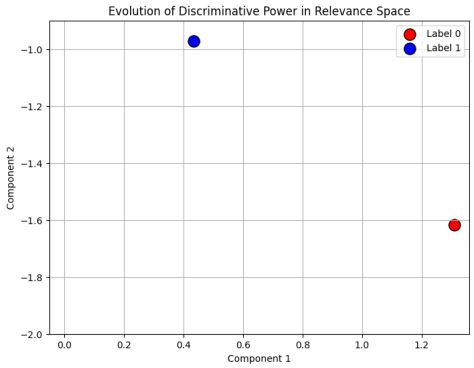


Figure 8: TSNE Visualization of Relevance Vectors during the later Active Discovery Phase.

available, enabling more effective model parameter optimization. This improved optimization enhances predictive accuracy, which in turn strengthens the overall search performance.

Table 9: **Search performance across varying search budgets $\mathcal{C}$.**

Search Budget ($\mathcal{C}$)	Accuracy (%)	F-score (%)	SR (%)
10	80 ± 1.5	50 ± 1.1	43 ± 1.3
30	81 ± 1.3	56 ± 1.2	58 ± 1.6
50	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

E Derivation of the Local Optimization Bound

The analysis in Sec. 4.2.3 is intended as a local optimization motivation rather than a global convergence guarantee for the full non-convex, nonstationary model. We formalize here how stochastic gradient variance and distributional drift affect the local tracking error.

Let $\mathcal{L}_t(\theta)$ denote the online optimization objective at acquisition step t, and let $\theta_t^\star$ denote a local minimizer of $\mathcal{L}_t$. Define the expected local suboptimality

$$G_t = \mathbb{E}[\mathcal{L}_t(\theta_t) - \mathcal{L}_t(\theta_t^\star)]\,, \tag{22}$$

and the movement of the local optimum

$$\delta_t = \left\|\theta_{t+1}^\star - \theta_t^\star\right\|_2, \qquad V_T = \sum_{t=1}^{T-1} \delta_t. \tag{23}$$

Local assumptions. We assume that, within the local optimization region traversed by the online updates:

1. $\mathcal{L}_t$ is L-smooth:
$$\mathcal{L}_t(\theta') \le \mathcal{L}_t(\theta) + \langle \nabla \mathcal{L}_t(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|_2^2.$$
2. $\mathcal{L}_t$ satisfies a local Polyak–Łojasiewicz (PL) condition with parameter $\mu > 0$:
$$\frac{1}{2}\|\nabla \mathcal{L}_t(\theta)\|_2^2 \ge \mu\,[\mathcal{L}_t(\theta) - \mathcal{L}_t(\theta_t^\star)]\,.$$
3. The stochastic gradient g_t obtained from the online meta-batch is conditionally unbiased and has bounded variance:
$$\mathbb{E}_t[g_t] = \nabla \mathcal{L}_t(\theta_t), \qquad \mathbb{E}_t\left[\|g_t - \nabla \mathcal{L}_t(\theta_t)\|_2^2\right] \le \sigma_{\text{rel}}^2.$$
4. Local changes in the objective induced by distributional drift are bounded proportionally to the movement of the local optimum. In particular, for some constant $c_d > 0$, the one-step tracking perturbation is at most $c_d \eta_t \delta_t$.

One-step recursion. Consider the stochastic update

$$\theta_{t+1} = \theta_t - \eta_t g_t. \tag{24}$$

By L-smoothness,

$$\mathbb{E}_t[\mathcal{L}_t(\theta_{t+1})] \leq \mathcal{L}_t(\theta_t) - \eta_t \|\nabla \mathcal{L}_t(\theta_t)\|_2^2 + \frac{L\eta_t^2}{2}\mathbb{E}_t[\|g_t\|_2^2] \tag{25}$$

$$\leq \mathcal{L}_t(\theta_t) - \frac{\eta_t}{2}\|\nabla \mathcal{L}_t(\theta_t)\|_2^2 + \frac{L\eta_t^2}{2}\sigma_{\text{rel}}^2, \tag{26}$$

where the second inequality holds for $\eta_t \leq 1/L$.

Applying the PL condition gives

$$\mathbb{E}_t\left[\mathcal{L}_t(\theta_{t+1}) - \mathcal{L}_t(\theta_t^\star)\right] \leq (1 - \mu\eta_t)\left[\mathcal{L}_t(\theta_t) - \mathcal{L}_t(\theta_t^\star)\right] + \frac{L\eta_t^2}{2}\sigma_{\text{rel}}^2. \tag{27}$$

Accounting for movement of the local optimum then yields the tracking recursion

$$G_{t+1} \leq (1 - \mu\eta_t)G_t + \frac{L\eta_t^2}{2}\sigma_{\text{rel}}^2 + c_d\eta_t\delta_t. \tag{28}$$

Unrolling the recursion. Take

$$\eta_t = \frac{2}{\mu(t+2)},$$

clipped at $1/L$ if necessary; clipping affects only constants in the following asymptotic bound. Then

$$1 - \mu\eta_t = \frac{t}{t+2}.$$

Using

$$\prod_{j=s+1}^{T-1} \frac{j}{j+2} = \frac{(s+1)(s+2)}{T(T+1)},$$

unrolling Eq. (28) gives

$$G_T \leq \frac{2G_1}{T(T+1)} + \frac{2L\sigma_{\text{rel}}^2}{\mu^2 T(T+1)}\sum_{s=1}^{T-1}\frac{s+1}{s+2} \tag{29}$$

$$+ \frac{2c_d}{\mu T(T+1)}\sum_{s=1}^{T-1}(s+1)\delta_s. \tag{30}$$

Since

$$\sum_{s=1}^{T-1}\frac{s+1}{s+2} \leq T$$

and

$$\sum_{s=1}^{T-1}(s+1)\delta_s \leq T\sum_{s=1}^{T-1}\delta_s = TV_T,$$

we obtain

$$\boxed{\mathbb{E}\left[\mathcal{L}_T(\theta_T) - \mathcal{L}_T(\theta_T^\star)\right] \leq \mathcal{O}\left(\frac{\sigma_{\text{rel}}^2}{T} + \frac{G_1}{T^2} + \frac{V_T}{T}\right).} \tag{31}$$

This bound should be interpreted locally: it does not establish global convergence of the full neural model or prove that relevance-aware meta-batches universally achieve smaller gradient variance. Rather, it shows that, all else equal, lower stochastic gradient variance yields a tighter local tracking error. Sec. 5.5 evaluates this mechanism empirically, where relevance-aware batching reduces measured gradient variance from 4.8×10^{-3} to 2.9×10^{-3}.

F On the Robustness of the Concept Encoder

A practical challenge in real-world environmental monitoring is that the set of available input channels may vary across datasets or time periods. In our PFAS setting, the 2022 data includes a more comprehensive set of domain-specific channels compared to 2019, reflecting improved data availability over time.

OWL-GPS accommodates such variation through its concept-wise encoding and relevance-based fusion architecture. Each input channel is independently processed by a pretrained concept encoder, producing representations in a shared embedding space. The relevance encoder uses shared parameters to estimate concept importance, allowing it to operate over different numbers of available channels. Specifically, for temporal transfer, we align the 33 concepts available in 2019 with a canonical 41-concept space, assigning zero relevance to unavailable concepts. During 2022 deployment, all 41 concepts are processed using their corresponding pretrained encoders, while the learned relevance estimation and prediction components are adapted online using acquired observations.

This design enables OWL-GPS to accommodate changes in feature availability while maintaining a consistent downstream representation. Empirically, the 2019-to-2022 results demonstrate its ability to operate under temporal distribution shift involving changes in both environmental conditions and available input features.

G Influence of Core and Reservoir Buffer Sizes on Search Performance

In this section, we analyze the impact of the core and reservoir buffer sizes on the search performance. We present the results in Table 10. Results in Table 10 show that performance drops when the core buffer dominates the reservoir, while balanced or reservoir-heavy configurations achieve stable and consistently high performance across all metrics. This indicates robustness to buffer size selection provided adequate reservoir capacity is preserved.

Table 10: **Search performance across varying buffer configurations.**

Buffer configuration	Accuracy (%)	F1-score (%)	SR (%)
Core $>$ Reservoir	77 ± 1.9	50 ± 1.1	88 ± 0.5
Core = Reservoir	87 ± 2.2	56 ± 1.0	98 ± 1.1
Core $<$ Reservoir (Ours)	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

H Mathematical Motivation for the Training Sampling Objective

We motivate the relevance-guided exploration objective using Gaussian similarity in the learned relevance space. Let $z(x) = \mu_\zeta(c(x)) \in \mathbb{R}^d$ denote the mean relevance representation of a candidate region x, and let $\mathcal{C}_{t-1}$ denote the core buffer before acquisition step t.

For two isotropic Gaussian densities with covariance $\sigma^2 I$, their overlap satisfies

$$\int \mathcal{N}(u; z, \sigma^2 I)\mathcal{N}(u; z', \sigma^2 I)\, du = (4\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|z - z'\|_2^2}{4\sigma^2}\right). \tag{32}$$

Normalizing by the overlap of a Gaussian with itself removes the constant factor and gives the similarity kernel

$$K_\sigma(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{4\sigma^2}\right). \tag{33}$$

To measure how similar an unobserved candidate is to previously acquired observations, we aggregate these similarities over the core buffer using their product:

$$R_t(x) = \prod_{x' \in \mathcal{C}_{t-1}} K_\sigma(z(x), z(x')) \tag{34}$$

$$= \exp\left(-\frac{1}{4\sigma^2} \sum_{x' \in \mathcal{C}_{t-1}} \|z(x) - z(x')\|_2^2 \right) \tag{35}$$

$$= \exp\left(-\frac{w_t(x)}{4\sigma^2} \right), \tag{36}$$

where $w_t(x)$ is the relevance-space novelty score defined in Eq. (11). Thus, the aggregate similarity decreases as the candidate becomes more distant from previously acquired observations. Its inverse, $R_t(x)^{-1}$, provides a corresponding novelty-promoting factor.

Using unit temperature ($4\sigma^2 = 1$) yields the novelty multiplier $\exp(w_t(x))$ used in our acquisition policy. Here, σ specifies the bandwidth of the auxiliary similarity kernel rather than the learned variance of the relevance encoder.

Novelty alone does not determine how informative a new label will be. We therefore combine it with predictive uncertainty. For pixel predictions $p_i(x) = f_{\theta_{t-1}}(x)_i$, define

$$U_t(x) = \sum_{i=1}^{P} \exp\left(-|p_i(x) - 0.5|\right). \tag{37}$$

This quantity assigns greater weight to predictions near the binary decision boundary.

Combining novelty and uncertainty gives the training-time acquisition objective

$$q_t = \arg\max_{x \in \mathcal{U}_t} \exp(w_t(x)) \sum_{i=1}^{P} \exp\left(-|f_{\theta_{t-1}}(x)_i - 0.5|\right), \tag{38}$$

which matches Eq. (12).

The Gaussian-overlap calculation provides an exact relationship between relevance-space distance and the proposed similarity measure. The resulting acquisition score is a novelty-weighted uncertainty heuristic, rather than an exact expression for the differential entropy of a Gaussian mixture.

I Mathematical Motivation for the Inference Sampling Objective

During deployment, the acquisition policy must balance exploration of unfamiliar regions with exploitation of regions predicted to contain targets. We motivate these two components using the relevance-space similarity measure developed in Appendix H and the predictive probabilities of the target model.

Let $p_i(x) = f_{\theta_{t-1}}(x)_i$ denote the predicted probability that pixel i in candidate region x belongs to the target class. If Y_i denotes its binary target label, then under the predictive distribution,

$$\mathbb{E}\left[\sum_{i=1}^{P} Y_i \mid x \right] = \sum_{i=1}^{P} p_i(x). \tag{39}$$

This motivates favoring regions with high predicted target probabilities.

We use the smooth, increasing per-pixel transformation $\exp(p_i(x))$ to construct a target-richness score:

$$T_t(x) = \sum_{i=1}^{P} \exp(p_i(x)). \tag{40}$$

This is a monotone-in-each-pixel surrogate for target richness, not an exact expected target count.

For exploitation, we additionally favor candidates similar to previously acquired observations in the learned relevance space. From Appendix H, using unit temperature, the aggregate similarity is

$$R_t(x) = \exp(-w_t(x)). \tag{41}$$

We therefore define the exploitation score as

$$S_{\text{exploit}}(x) = R_t(x)T_t(x) \tag{42}$$

$$= \exp(-w_t(x)) \sum_{i=1}^{P} \exp(f_{\theta_{t-1}}(x)_i), \tag{43}$$

which is Eq. (13).

For exploration, we instead favor relevance-space novelty and uncertain predictions. Using the inverse similarity factor from Appendix H and the uncertainty score $U_t(x)$, we obtain

$$S_{\text{explore}}(x) = R_t(x)^{-1}U_t(x) \tag{44}$$

$$= \exp(w_t(x)) \sum_{i=1}^{P} \exp\left(-|f_{\theta_{t-1}}(x)_i - 0.5|\right), \tag{45}$$

which matches Eq. (14).

Finally, we combine the two scores using the budget-dependent coefficient κ_t:

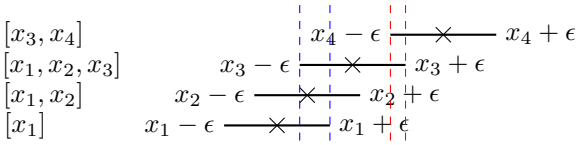
$$a_t(x) = \kappa_t S_{\text{explore}}(x) + (1 - \kappa_t)S_{\text{exploit}}(x), \tag{46}$$

and select $q_t = \arg\max_{x \in \mathcal{U}_t} a_t(x)$, as in Eqs. (15)–(16).

The resulting policy promotes exploratory acquisitions when the available budget is large and increasingly emphasizes target-rich regions as the budget is consumed. Both components are explicitly constructed from relevance-space similarity and model predictions. They serve as acquisition surrogates rather than exact joint log-likelihood expressions.

J Details of Greedy Intersection Clustering Algorithm

This section details the greedy-intersection clustering algorithm used to cluster the relevance vectors in the core buffer, as highlighted in the meta-batch formation subsection in the main paper. The Greedy Intersection clustering algorithm was originally proposed in Ge et al. [2025]. We present it in detail for completeness. Consider a set $\theta \in T$ of relevance vectors, fix K (i.e., the number of clusters), and suppose ϵ is the radius of each cluster. The key intuition behind the Greedy Intersection Algorithm is that for any $\theta \in T$, an ϵ-hypercube centered at θ characterizes all possible θ' that can cover θ in the sense that $||\theta - \theta'||_\infty \leq \epsilon$. Consequently, if two ϵ-hypercubes centered at θ and θ' overlap, any point within their intersection can simultaneously cover both. We demonstrate this with the following straightforward example:



Each cross marks a target parameter to be covered, and each line segment shows the range within which an ϵ-close representative of that parameter may lie. Selecting a point within the overlapping region of these intervals allows us to simultaneously cover multiple parameters.

The Greedy Intersection algorithm builds on this intuition by generalizing it as follows. In the first stage, it constructs an intersection tree independently for each dimension. For the s-th dimension, the data points are sorted in ascending order based on their s-th coordinate, denoted as $x_1 < x_2 \cdots < x_n$. For each point x_i, we initialize a list containing only $[x_i]$, which will be used to track how many other points can be jointly covered along with it.

Starting from the second smallest data point x_2, we check if $x_2 - \epsilon \leq x_1 + \epsilon$, i.e., if $x_2 \leq x_1 + 2\epsilon$. Since $x_2 - \epsilon > x_1 - \epsilon$ due to our sorting, any point inside $[x_2 - \epsilon, x_1 + \epsilon]$ can cover both x_1, x_2. Consequently, if this interval is valid, we append x_1 to x_2's list to indicate that both points can be covered simultaneously. More generally, for any point x_i, we iterate backward through the preceding points x_j (from $j = i - 1$ down to 1) to verify if $x_i \leq x_j + 2\epsilon$. If the condition is met, we add x_j to x_i's list. Because the set is sorted, as soon as the condition fails, we can stop checking; if

$x_i > x_j + 2\epsilon$, it is guaranteed that $x_i > x_{j'} + 2\epsilon$ for all $j' < j$. Ultimately, the common covering interval for all points in x_i's list is $[x_i - \epsilon, x_j + \epsilon]$, where j represents the smallest index successfully added to the list. There are $1 + 2 + \cdots + n - 1 = \mathcal{O}(n^2)$ comparisons in total. We form a set of these lists, and call it $\mathcal{A}_s$ for the s-th dimension. The figure above illustrates how the algorithm works to find $\mathcal{A}_1 = \{[x_1], [x_1, x_2], [x_1, x_2, x_3], [x_3, x_4]\}$.

The second stage of the algorithm aims to identify a hypercube that covers the maximum number of points, selecting one axis-aligned interval from each dimension. By the geometry of Euclidean space, two points θ_1 and θ_2 are within an ϵ-distance in ℓ_∞ norm if and only if they appear together in each other's lists across all dimensions. Thus, to find a hypercube whose center is within ℓ_∞-distance of the most data points, we search for the combination of lists $l_1, \ldots, l_d$, where each l_s is chosen from the corresponding set $\mathcal{A}_s$, such that their intersection has the maximum cardinality. In our illustrative example, we observe that the groups $[x_1, x_2, x_3]$ and $[x_3, x_4]$ must be covered separately, each by a different point positioned between the red or blue vertical lines.

A key limitation of the Greedy Intersection Algorithm (GIA) is its exponential complexity with respect to the number of dimensions in the data. To address the challenge of high dimensionality, we apply standard dimensionality reduction techniques—specifically, Principal Component Analysis (PCA) - to project each data point into a lower-dimensional space that is computationally more tractable. We then apply the Greedy Intersection Algorithm to these PCA-transformed points. In our experiments, we use a reduced dimensionality of 7, which we found to yield strong performance in our problem setting. The complete algorithm is presented below in Algorithm 1.

Algorithm 1 Greedy Intersection

Input: $T = \{\theta_i\}_{i=1}^N$, $\epsilon > 0$, $K \geq 1$
Output: Parameter cover C

$C \leftarrow []$
for $round\ k$ = 1 to K **do**
  for $dimension\ m$ = 1 to d **do**
    Sort T in ascending order based on their m-th coordinates
    $lists_m \leftarrow []$
    for $indiviual\ i = 2$ to N **do**
      $S_i \leftarrow [\theta_i]$
      for $j = i - 1$ to 1 **do**
        if θ_i's m-th coordinate $<$ θ_j's m-th coordinate $+2\epsilon$ **then**
          Add θ_j to S_i
        else
          if $lists_m[-1] \subseteq S_i$ **then**
            $lists_m[-1] \leftarrow S_i$
          else
            Add S_i to $lists_m$
          end if
          break
        end if
      end for
    end for
  end for
  $S^{1*}, \ldots, S^{m*} \leftarrow \text{argmax}_{S^1 \in lists_1, \ldots, S^m \in lists_m} |S^1 \cap \cdots \cap S^m|$
  $covered \leftarrow S^{1*} \cap \cdots \cap S^{m*}$
  $\hat{\theta}_k \leftarrow$ average of the $covered$
  $T \leftarrow T - covered$
  C.adds($\hat{\theta}_k$)
end for
return C

Additionally, to evaluate the effectiveness of the Greedy Intersection clustering algorithm, we conduct a comparative analysis by replacing it with a standard clustering method (i.e., DBSCAN) and measuring the resulting performance on the downstream task. The outcomes of this comparison are summarized in the following table. Our empirical results indicate that incorporating the Greedy Intersection clustering approach within the framework achieves improved performance on accuracy

and SR metrics, while yielding a significant improvement in F-score. This improvement indicates enhanced robustness of the framework, reflecting a better balance between precision and recall under the proposed clustering strategy, thereby underscoring the effectiveness of the Greedy Intersection clustering approach under the proposed framework.

Table 11: **Effectiveness of the clustering approach.**

Clustering approach	Accuracy (%)	F-score (%)	SR (%) ($\mathcal{C} = 50$)
DBSCAN	81 ± 2.2	57 ± 1.6	94 ± 1.1
Ours	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

K Additional Analysis on the Importance of Relevance Guided Sampling

Our adoption of a Conditional Variational Autoencoder (CVAE) as the backbone of our relevance estimator is driven by the inherent dynamism of open-world geospatial search environments, where the relevance of concepts can shift unpredictably across both spatial and temporal dimensions. The CVAE framework enables the model not only to learn complex relevance patterns but also to quantify associated uncertainties - empowering the system to make decisions that are both adaptive and robust to changing conditions. This uncertainty-aware modeling is especially crucial in large-scale and data-sparse geospatial scenarios, where assumptions of static or simple relevance easily break down.

In rigorous empirical comparisons, the CVAE-based estimator consistently outperforms a simpler multi-layer perceptron (MLP) attention-based relevance estimation baseline (denoted as Ours (RG $\rightarrow$ MLP)), delivering improvements in performance, as demonstrated in the results in Table 12 and also detailed in Section 5 (Tables 7). These findings underscore the substantial practical advantage of leveraging probabilistic deep generative modeling for relevance estimation, affirming that the CVAE's ability to jointly represent relevance values and their uncertainties translates directly to superior real-world search effectiveness.

Table 12: **Analyzing the impact of relevance-guided sampling.**

Method	Accuracy (%)	F-score (%)	SR (%)
Ours (RG $\rightarrow$ MLP)	85 ± 1.7	56 ± 1.8	94 ± 2.3
Ours	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

L Importance of Orthogonalization in the Concept Space

We assess the effectiveness of the orthogonalization layer by comparing the performance of the proposed framework with a variant that only excludes orthogonalization. The results are presented in Table 13. The empirical results indicate a noticeable decline in the metrics when the orthogonalization layer is omitted from the proposed framework, highlighting the importance of the orthogonalization of the concepts within the proposed framework.

Table 13: **Importance of concept orthogonalization.**

Orthogonalization	Accuracy (%)	F-score (%)	SR (%) ($\mathcal{C} = 50$)
No	80 ± 1.0	56 ± 1.5	86 ± 0.7
Yes (our framework)	$\mathbf{88 \pm 2.0}$	$\mathbf{65 \pm 2.0}$	$\mathbf{98 \pm 1.0}$

M Dataset and Split Details

We evaluate our approach on two geospatial prediction tasks: PFAS contamination prediction and land cover classification, both of which use satellite data aligned with sparsely sampled ground truth.

For the PFAS contamination prediction task, each sample consists of a multi-channel raster patch centered on a known PFAS measurement site from EPA datasets, including the National Rivers and Streams Assessment (NRSA) [2] and the National Lakes Assessment Fish Tissue Study [3]. This dataset

[2] NRSA

[3] National Lakes Assessment Fish Tissue Study

is inherently sparse, with only 704 labeled sample points across the US. Each patch is of size 256 × 256 pixels at 30-meter resolution and contains a set of multi-channel geospatial features, with the number of channels varying across datasets (33 for PFAS 2019 and 41 for PFAS 2022). These channels encode diverse geospatial features, such as land cover rasters from the National Land Cover Database (NLCD) [4], flow direction rasters capturing hydrological connectivity obtained from the ArcGis Pro software, and distance transforms to known PFAS discharger sites obtained from the U.S EPA database [5]. This rich set of environmental and spatial layers provides the model with meaningful context to predict contamination.

For the land cover classification task, we use imagery from the Sentinel-2 surface reflectance dataset [6]. For each sample region, we extract 4-channel image patches using the Near-Infrared (NIR), Red, Green, and Blue bands. Each patch is of size 256 × 256 pixels at 30-meter resolution. These patches are aligned with the same spatial footprint as the PFAS patches to ensure consistency in coverage.

All datasets used are publicly available and used in accordance with their respective licenses (e.g., U.S. EPA datasets, NLCD), which permit research use.

Dataset Splits and Query Budgets. We construct disjoint training and test splits under both spatial and temporal generalization settings as shown in Figure 9. In the spatial setting, training and test samples are drawn from non-overlapping geographic regions within the same year (2019), while in the temporal setting, models are trained on 2019 data and evaluated on 2022 data to capture distribution shift. For each setting, we ensure that training and test sets contain no overlapping sample locations or spatial patches. During training, we use a fixed query budget of 150 samples to simulate limited labeled data acquisition. At inference, we evaluate performance under a primary query budget of 50 samples, with additional budgets considered in ablations. For the land cover (LC) task, we follow the same protocol but simulate sparse supervision by randomly subsampling labels, treating it as a controlled generalization experiment rather than a real-world acquisition setting.

N Details of Pseudolabel Generation Procedure to Induce Stability During Training

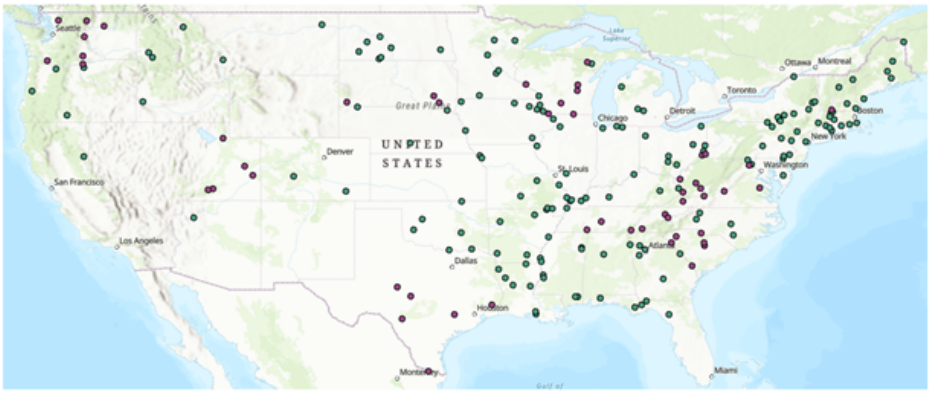
Given the scarcity of ground truth data in PFAS monitoring, typically limited to sparse point-based measurements, we employ a point-based label expansion strategy to generate dense supervision signals suitable for segmentation training. Specifically, each labeled point is used to generate a pseudo-label mask over its surrounding patch, enabling the model to learn from broader spatial context [7]. Special "noise masks" are used in the loss computation to reduce the influence of uncertainty introduced due to the pseudo-labeling assumptions.

For each georeferenced point with a PFAS contamination label (1 for above health advisory threshold, 0 for below), we extract a raster patch centered on that point. All surface water pixels within this patch are assigned the same label as the central point, reflecting the assumption, guided by environmental science, that contamination may be correlated in hydrologically connected regions. Non-surface water pixels are assigned a special label (2) and are excluded from loss computation. This strategy effectively converts sparse point annotations into dense training masks, allowing the model to be trained as a segmentation network despite the absence of full pixel-wise labels.

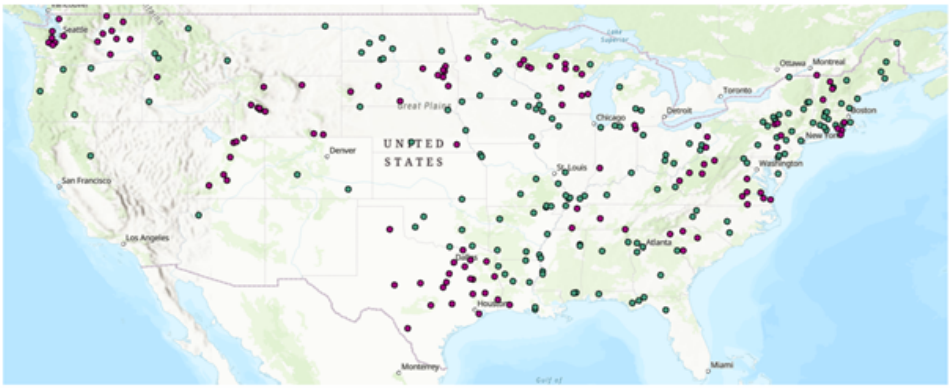
Crucially, model evaluation is performed exclusively on ground truth point data from the U.S. EPA PFAS datasets, ensuring that performance metrics reflect real-world observations and are not influenced by the assumptions underlying pseudo-label generation.

For the land cover dataset, dense per-pixel land cover annotations are used as supervision only during training, and are derived from public land cover products that include standard classes such as water, forest, urban, and agricultural land [8]. During testing, however, we adopt a sparse label evaluation setup: only one or two labeled pixels are retained in each test mask, simulating the limited label availability often encountered in practice. The model is expected to make predictions over the entire

[4] NLCD
[5] US EPA
[6] Sentinel-2
[7] FOCUS
[8] NLCD



(a) Spatial generalization (2019): training and test samples are drawn from disjoint geographic regions within the same year.



(b) Temporal generalization: training samples from 2019 and test samples from 2022, capturing distribution shift over time.

Figure 9: **PFAS dataset splits.** Green denotes training samples and purple denotes test samples. Splits are geographically disjoint; apparent overlap arises from visualization at this scale.

patch, but evaluation is conducted only at these sparsely labeled locations. This setup mirrors the PFAS task and encourages the model to generalize beyond isolated supervision.

N.1 Additional Discussions on the Label Cost

The OWL-GPS setting models constraints on supervision, not imagery. For instance, **Sentinel imagery is abundant; however, the labels required for the discovery tasks in our datasets are not. In real scientific monitoring problems like PFAS contamination, labeled examples are scarce, expensive, and slow to acquire.** Moreover, OWL-GPS assumes an online, memory-limited regime in which the agent cannot store or replay past inputs beyond a small fixed buffer. These are the core constraints the framework is designed to capture.

1. PFAS (2019 & 2022): Each PFAS label corresponds to an actual field measurement collected by environmental agencies. Such measurements require field deployment, sample collection, laboratory chemical analysis, and administrative coordination. As a result, very few labeled samples exist, and expanding coverage is constrained by budget, personnel, and lab throughput. This is the source of "costly" and "resource-constrained" operations. In addition, under OWL-GPS assumptions, inputs are non-replayable: once a geospatial region has been processed and leaves the limited memory buffer, it cannot be revisited for training or queried again. The agent selects regions for labeling from the visible candidate pool, while previously acquired observations can only be retained for reuse within bounded memory.
2. Land Cover (LC): For LC, labels are not inherently difficult to acquire, but the purpose of including LC is to demonstrate that the framework generalizes beyond PFAS. To place LC

into a discovery setting comparable to PFAS, we **sparsify the supervision as described in Appendix N**. This creates a regime where only a small subset of tiles are labeled, matching the limited-supervision assumption of OWL-GPS. We clarify that this sparsification is simulated to evaluate generalization, not a claim about real-world LC label difficulty.

O Details of Training and Inference Hyperparameters

We used a patch size of 256×256 pixels for all geospatial regions. Model training was performed using the Adam optimizer, with learning rate dynamically adjusted using a polynomial decay scheduler with warmup. Training used focal loss with a KL-divergence regularization term as part of a conditional variational autoencoder objective for relevance learning. Online-meta updates were performed using meta-batches of size 10, formed by selecting one representative sample per cluster from the core buffer, and three samples from the reservoir buffer. All reported results are averaged over three independent runs, unless otherwise specified, with different random seeds to ensure robustness.

P Architecture Details

P.1 Details of Concept Encoder

This section details the implementation of the concept encoder as highlighted in the concept encoder subsection in the main paper. The Concept Encoder is implemented using a modified Vision Transformer (ViT) architecture, tailored to handle geospatial imagery possibly with temporal structure. The encoder is structured as a deep transformer-based module that ingests high-dimensional multi-spectral or multi-temporal satellite imagery and maps it into a low-dimensional concept space that captures the essential semantics of the input.

The encoder uses a patch-based embedding mechanism to divide the input tensor into non-overlapping spatial patches. Each patch is linearly projected into an embedding vector, to which a learnable class token and fixed sinusoidal 3D positional encodings are added. These embeddings are then processed through a deep stack of transformer blocks, each consisting of multi-head self-attention and MLP layers, followed by a final normalization step.

The encoder is initialized using a sin-cos positional embedding and can optionally load pretrained weights. To improve representation disentanglement and mitigate redundancy, we apply Gram-Schmidt orthogonalization to the learned latent vectors, which promotes diversity among the learned concept axes. The output of the encoder is a tuple of patch-wise embeddings enriched with contextual information, which serve as the base representations for relevance inference and downstream target prediction.

Training Objective. The concept encoder is pretrained in a masked autoencoding (MAE) framework following prior geospatial foundation models. Given an input patch $x \in \mathbb{R}^{H\times W\times K}$, a subset of patches is randomly masked, and the encoder processes only the visible tokens. A lightweight decoder then reconstructs the masked regions.

We optimize a reconstruction loss over masked patches:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{(h,w)\in\mathcal{M}} \|x_{h,w} - \hat{x}_{h,w}\|_2^2,$$

where $\mathcal{M}$ denotes the set of masked patches, $\hat{x}$ represents the reconstructed masked patch. This encourages the latent representation to capture the underlying spatial and semantic structure of the multi-channel geospatial input.

P.2 Details of Relevance Encoder

This section details the implementation of the relevance encoder as highlighted in the relevance encoder and decoder subsection in the main paper. The Relevance Encoder estimates a latent relevance distribution over concepts to capture their varying importance in influencing target presence across geospatial regions. It is modeled as the probabilistic encoder component in a Conditional Variational Autoencoder (CVAE), and takes as input the region-specific concept embeddings from the concept encoder.

This module is designed to process batches of concept embeddings organized per concept, and leverages a lightweight yet expressive convolutional backbone consisting of three convolutional layers with ReLU activations and batch normalization, followed by spatial downsampling via max-pooling. The resulting feature maps are flattened and passed through two separate linear layers to produce the mean and standard deviation vectors, one for each concept. These vectors parameterize a diagonal Gaussian distribution from which a sample relevance vector is drawn during training via the reparameterization trick.

The output relevance vector quantifies the relative contribution and uncertainty of each concept with respect to the target prediction task. This facilitates not only accurate modeling but also reasoning about the influence of different domain factors. Empirically, the relevance encoder was shown to significantly boost predictive accuracy and search efficiency.

P.3 Details of the Decoder

This section details the implementation of the decoder as highlighted in the relevance encoder and decoder subsection in the main paper. The Decoder is responsible for producing pixel-level predictions of target presence by conditioning on the fused representation of concepts and their corresponding relevance scores. It is designed using a dual-head FCN (Fully Convolutional Network) architecture to enhance segmentation robustness during learning.

Each head is a configurable FCN variant that processes the fused feature map via a sequence of convolutional layers, followed by feature concatenation and a final classification layer. The first head is lightweight with a single convolution, while the second includes a deeper convolutional stack to model more complex interactions.

The decoder architecture aligns with semantic segmentation principles and supports variable numbers of output classes, dropout for regularization, and batch normalization to stabilize training. This component plays a critical role in translating abstract latent representations into actionable geospatial predictions.

P.4 Modeling Spatial and Hydrological Structure.

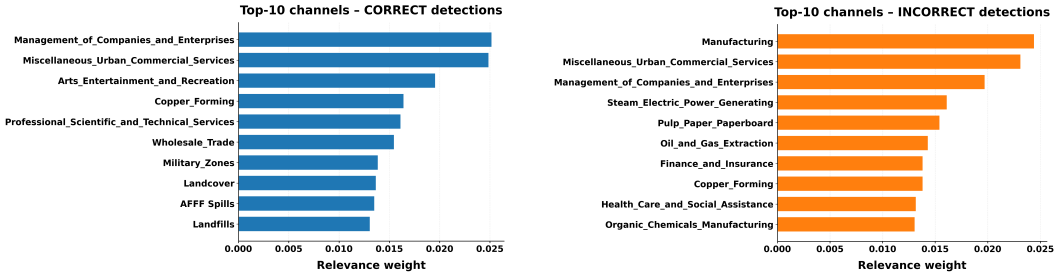
While OWL-GPS does not explicitly incorporate graph-based message passing over river networks, spatial and hydrological dependencies are encoded through the input representation. Specifically, the model ingests domain-informed channels such as flow direction rasters, distance-to-source features, and land cover, which capture hydrological connectivity and spatial context. As a result, regions that are hydrologically linked (e.g., upstream/downstream) exhibit similar patterns in the concept space, enabling the model to implicitly learn these dependencies. Importantly, OWL-GPS operates in a sequential, budget-constrained setting where the full spatial graph is not known a priori and evolves as new regions are queried. This makes explicit graph construction and propagation impractical, as it would require access to the entire network structure and labeled data across nodes. Instead, our approach leverages learned relevance over concept representations to generalize across regions with similar hydrological characteristics, allowing information from observed locations to influence decision-making in unobserved areas without explicit message passing.

Q Additional Relevance Interpretability Analysis

We provide an additional qualitative analysis of the learned relevance representation, complementing the quantitative and context-dependent relevance analyses in Sec. 5.6.

Q.1 Relevance Profiles for Correct and Incorrect Predictions

We compare representative correct and incorrect predictions. Figure 10 shows that a correct high-confidence prediction emphasizes plausible PFAS-related environmental factors, including proximity to landfills, Aqueous Film-Forming Foam (AFFF) spills, and military sites [US EPA, 2021], whereas the incorrect prediction exhibits a less concentrated relevance profile across these drivers. This provides complementary qualitative evidence that prediction outcomes are associated with different relevance structures.



(a) Correct prediction.

(b) Incorrect prediction.

Figure 10: Highest-ranked concept channels for representative correct and incorrect predictions.

R Additional Details about the Baselines

R.1 Details about the Baselines:

OWL-GPS introduces a new constrained, streaming, non-replayable problem formulation. Therefore, each baseline must be carefully adapted to comply with the OWL-GPS constraints, such as streaming inputs, no revisitation, limited memory and query budget, and updates only through queried labels. Below, we provide explicit descriptions of how every baseline is trained and evaluated under this unified setup.

1. **Active Learning (AL)**: Successful performance in OWL-GPS requires a delicate balance between exploration and exploitation during inference. AL methods primarily focus on exploration by querying areas of highest model uncertainty. While this ensures thorough coverage, it overlooks exploiting accumulated knowledge about promising regions. Consequently, AL struggles to efficiently navigate the trade-off necessary for high performance in OWL-GPS. For training with the active learning baseline, we adhere strictly to the standard maximum uncertainty strategy for selecting samples to query during a standard supervised training with a ViT. For evaluation, we apply our sampling strategy with one key modification: the selection of queried samples is driven exclusively by the exploration score, with no consideration given to the exploitation score in the sampling process.
2. **Greedy Approach (GA)**: GA represents the contrasting extreme by prioritizing exploitation — querying locations where the model predicts the highest likelihood of the target. Purely exploitation-driven strategies like GA risk premature convergence to suboptimal regions and fail to sufficiently explore the environment. Our experimental findings corroborate this limitation, showing inferior performance of GA relative to our method. GA conducts a single forward pass over all incoming regions during evaluation after a standard supervised training phase with a ViT, assigning confidence scores to each. It then selects the top-K regions based on the query budget. GA does not perform online updates, store past regions, or revisit inputs, making it a strict single-pass, non-adaptive baseline within OWL-GPS.
3. **Prithvi**: Although Prithvi is a powerful and general framework, it is not specifically designed to handle the multi-faceted challenges of the OWL-GPS problem, such as continual adaptation in dynamic environments. These complex constraints underscore the need for a tailored approach like ours. Specifically, the geospatial foundation model Prithvi is trained using a standard supervised approach. During evaluation, it assigns confidence scores to unobserved samples, selects the highest scoring sample, updates its parameters with the latest observations, and repeats this process for subsequent batches until the query budget is exhausted.
4. **UCB (Bandit exploration)**: At first glance, the OWL-GPS problem may appear similar to a Multi-Armed Bandit (MAB) scenario, making UCB a natural baseline. However, unlike typical MAB settings where each arm is independent, geospatial environments exhibit strong spatial correlations; observing one location informs us about neighboring regions. UCB does not leverage this spatial structure, treating each arm independently. This lack of spatial modeling fundamentally limits its performance in our problem setting, as reflected in Table 2. We adapt UCB to the OWL-GPS setting by treating each candidate region in the visible unlabeled pool as an arm, assigning it a score based on its predicted reward (model confidence) and an exploration bonus. At each acquisition round, the highest-scoring

candidates are selected under the query budget, and the model is updated online using the newly acquired labels. Thus, UCB operates under the same candidate visibility and query budget as our framework, while differing in its acquisition strategy.

5. **Online Meta-Learning (OML)**: OML in our setup follows the online meta-update rule. One key difference is that OML does not use our meta-batch formation strategy with GIA for constructing diverse update sets. Instead, OML directly applies the update using the examples present in our buffer at that moment.
6. **Active Meta-Learning (AML)**: AML follows a training approach similar to ours but differs in two key ways: (a) it forms the meta-batch using standard random sampling from the buffer, and (b) it lacks an online adaptation mechanism. Our experiments demonstrate that AML is ineffective in the OWL-GPS setting, highlighting the critical role of an online update mechanism combined with a coverage-based meta-batch formation strategy.

S Extended Related Work

S.1 Environment Monitoring:

Prior studies such as Sarkar et al. [2023b] and Sarkar et al. [2023a] have investigated active sampling in related environmental monitoring contexts. However, several fundamental constraints, including the need for strictly online operation, adherence to tight budget limits, memory constraints, and the handling of non-stationary and evolving distributions, distinguish our setting and make these previous approaches unsuitable for OWL-GPS. More generally, there is a rich literature on environmental monitoring methods, but the vast majority focus on settings with static or replayable datasets, abundant labeling, and stationary environments, without addressing the unique challenges posed by our real-world, online discovery scenario. Below, we briefly review key environmental monitoring works to further contextualize the distinctiveness of our approach. Environmental monitoring relies heavily on geospatial analysis, an interdisciplinary approach combining geography, computer science, statistics, and engineering to extract valuable insights from spatial data such as satellite imagery, GPS, and historic datasets. This enables critical applications including understanding land use patterns (Park et al. [2023]), infrastructure planning (Amaral et al. [2021]; Regona et al. [2024]), and assessing environmental impacts (Milà et al. [2023]). These geospatial tools support urban planning for sustainable expansion (Gharaibeh et al. [2020]) and transportation systems (Kamruzzaman et al. [2015]). Recent technological advances, including improved satellite resolution, enhanced GPS accuracy, and increased drone deployment, have transformed data collection and analysis capabilities, facilitating real-time monitoring of key phenomena such as deforestation (Monjardin-Armenta et al. [2020]), glacier retreat (Thapliyal et al. [2023]), and urban heat islands. Additionally, geospatial analysis supports disaster risk management, including wildfire (Shafapourtehrany et al. [2023]), flood (Liao et al. [2023]), and earthquake prediction (Lam et al. [2021]), improving evacuation and relief efforts (Manfré et al. [2012]).

Machine learning (ML), deep learning (DL), computer vision (CV), and natural language processing (NLP) have been increasingly integrated with geospatial tools to enhance data processing and predictive power (Casali et al. [2022]). These methods facilitate large-scale pattern recognition and predictive modeling, crucial for environmental risk assessment and urban sustainability (Son et al. [2023]). CV enables automated interpretation of satellite and street-view imagery for land classification and change detection (Marasinghe et al. [2024]), whereas NLP analyses research articles and reports for insights on environmental changes and biodiversity trends (Cao et al. [2024]). Overall, this integration of geospatial sciences with advanced computational methods paves the way for innovative environmental monitoring solutions, addressing challenges from urbanization to climate change with greater precision and scale. However, our current OWL-GPS problem setting is fundamentally distinct from previously studied problems, capturing the complex nuances of real-world environmental monitoring challenges in a unified framework.

S.2 Geospatial Foundational Model

The development of specialized foundation models has been driven by the need for precision and contextual sensitivity within scientific domains. In geospatial analysis, Geospatial Foundation Models (GFMs) have emerged as powerful tools tailored to interpreting complex Earth surface and atmospheric patterns (Mai et al. [2023]). GFMs address critical challenges such as spatial heterogeneity (Sun et al. [2022]), temporal dynamics (Yao et al. [2023]), and the multidimensional

nature of geospatial data (Jakubik et al. [2023]). Transformer architectures, especially Vision Transformers (ViT; Dosovitskiy [2020]) and hierarchical designs like Swin Transformers (Liu et al. [2021]), have become foundational components due to their ability to model long-range dependencies and dynamic attention. Recent innovations enhance GFMs capabilities for spatiotemporal data, for instance, incorporating temporal information as channels (Jakubik et al. [2023]) or designing multi-branch networks to capture spatial affinity and temporal continuity (Yao et al. [2023]). Adaptations of pretrained models such as SAM (Kirillov et al. [2023]) exemplify how conventional image analysis methods are refined for specific geospatial tasks like SAR imagery segmentation (Yan et al. [2023]). Furthermore, Masked Autoencoders (MAE; He et al. [2022]) are widely employed for self-supervised training, enabling scalable learning from unlabeled geospatial imagery, while supervised fine-tuning is applied for specialized downstream tasks requiring category-specific outputs (Yan et al. [2023]; Yao et al. [2023]). Among Geospatial Foundation Models, IBM's Prithvi (Jakubik et al. [2023]) is distinguished by its innovative approach to GeoAI and geospatial data analysis, making it a prime candidate for detailed evaluation in this work. Prithvi typically supports six spectral bands: Red, Green, Blue, NIR, SWIR 1, and SWIR 2, enabling richer spectral information capture beyond conventional RGB imagery. This multi-band capability enhances model versatility and effectiveness across diverse geospatial applications, including land cover mapping and environmental monitoring.

T Time and Memory Complexity Details of the Proposed Framework:

T.1 Time Complexity Details

Let K denote the bounded memory size, D the relevance embedding dimension, N the number of available unlabeled candidates, and B the online update batch size. Let F and G denote the computational costs of a forward evaluation and a forward-backward update, respectively.

For a fixed number of clusters and reduced PCA dimensions, the fast Greedy Intersection clustering procedure costs $O(K^2D)$ in the worst case, excluding the PCA projection. Computing relevance-space novelty for one candidate requires $O(KD)$ operations, while evaluating its predictions costs $O(F)$. Scoring the entire unlabeled candidate pool therefore requires $O(NKD + NF)$. A gradient update over the relevance-structured batch costs $O(BG)$.

The overall computational cost per acquisition round is thus

$$O(K^2D + NKD + NF + BG),$$

in addition to PCA computation when the projection is recomputed. Since the memory capacity and meta-batch size are bounded, the dependence on the unlabeled pool size N is linear, rather than quadratic, for fixed model evaluation costs.

Furthermore, we present the time and memory costs associated with our proposed framework as follows:

- On an NVIDIA A100 GPU, sampling step during inference takes on average approximately 2.1 minutes (128.8 seconds). The model's decision-making step is lightweight, requiring only 4 GB of GPU memory.

T.2 Details about Compute Resource and Code

All experiments were conducted using PyTorch on an NVIDIA A100 GPU.

U Implementation Details and Training/Inference Pseudocode

The proposed bilevel meta-update is motivated by the need to adapt to newly acquired observations while maintaining generalization across previously encountered data. Such updates remain relevant in sparse-supervision and costly-acquisition settings, provided the available observations offer sufficient supervision to construct meaningful meta-training and meta-validation subsets. However, our experiments operate under particularly limited ground-truth availability, with only 290 and 414 labeled pixels in the entire 2019 and 2022 datasets, respectively. Although pseudo-labeling provides additional training supervision, independently verified labels remain scarce, and sequential acquisition budgets further restrict their availability. Combined with bounded-memory constraints, dividing the resulting relevance-structured batches into separate meta-training and meta-validation subsets leaves very limited supervision for each update. In our experiments, this resulted in unstable

bilevel optimization. To address this, we relax the bilevel structure by incorporating newly acquired observations directly into the relevance-structured meta-batch $\mathcal{D}_t$, alongside representative historical observations. Model parameters are then updated over this unified batch using a gradient-based update analogous to Eq. 17, without separate meta-validation or latest-observation-based updates. This retains relevance-guided online adaptation while providing more stable optimization under the limited supervision available in our experiments.

We present training and inference pseudocode in Algorithm 2 and Algorithm 3, respectively.

Algorithm 2 Relevance-Guided Sequential Training with Latent Concepts for Geospatial Discovery (Training)

Require: Training pool $\mathcal{X}_{\text{train}}$, Query budget $\mathcal{C}_{\text{train}}$
Require: Concept encoder CE_ψ, relevance encoder RE_ζ, decoder RD_ϕ
Require: Core buffer $\mathcal{C}_0 = \emptyset$, reservoir buffer $\mathcal{R}_0 = \emptyset$
Initialize parameters $\theta = \{\zeta, \phi\}$, step $t \leftarrow 1$, and ψ is frozen.
while $t \leq \mathcal{C}_{\text{train}}$ **do**
  (Training-time sampling)
  for each unlabeled region $x \in \mathcal{X}_{\text{train}}$ **do**
    $\tilde{c}(x) \leftarrow \text{CE}_\psi(x)$
    $c(x) \leftarrow \text{GS}(\tilde{c}(x))$
    $(\mu_\zeta(c(x)), \sigma_\zeta(c(x))) \leftarrow \text{RE}_\zeta(c(x))$
    Sample $r(c(x)) \sim \mathcal{N}(\mu_\zeta(c(x)), \text{diag}(\sigma^2_\zeta(c(x))))$
    $f_\theta(x) \leftarrow \text{RD}_\phi(c(x), r(c(x)))$
    $w_t(x) \leftarrow \sum_{x' \in \mathcal{C}_{t-1}} \|\mu_\zeta(c(x)) - \mu_\zeta(c(x'))\|_2^2$
    $\exp(w_t(x))$ denotes the relevance-space novelty weight of region x
    $u_{\text{dec}}(x) \leftarrow$ *prediction uncertainty* of region x using $\sum_{i=1}^{P} \exp^{\left(-|f_\theta(x_i)-0.5|\right)}$
    $\text{score}_{\text{train}}(x) \leftarrow \exp^{(w_x)}\ u_{\text{dec}}(x)$
  end for
  $x^{(q_t)} \leftarrow \arg\max_{x \in \{\text{Set of Unobserved regions}\}} \text{score}_{\text{train}}(x)$
  $y^{(q_t)} \leftarrow$ Observe ground truth labels corresponds to the region $x^{(q_t)}$.
  (Buffer update)
  $\mathcal{C}_t \leftarrow \mathcal{C}_{t-1}$, $\mathcal{R}_t \leftarrow \mathcal{R}_{t-1}$
  Update durations of retained samples in $\mathcal{C}_t$ and $\mathcal{R}_t$
  Insert $(x^{(q_t)}, y^{(q_t)})$ into $\mathcal{C}_t$ with initialized duration and count
  Move expired core samples to $\mathcal{R}_t$, evict expired or excess reservoir samples
  (Meta-batch formation)
  Cluster elements of $\mathcal{C}_t$ in relevance space
  From each cluster, select a sample maximizing $\exp(\text{duration}/(\text{count}+1))$ to form $\mathcal{D}_t^{\text{core}}$
  Select K_r samples from $\mathcal{R}_t$ with weights $\propto \exp(\text{duration}/(\text{count}+1))$ to form $\mathcal{D}_t^{\text{res}}$
  $\mathcal{D}_t \leftarrow \mathcal{D}_t^{\text{core}} \cup \mathcal{D}_t^{\text{res}}$
  $D_t^{\text{comb}} \leftarrow D_t \cup \{(x^{(q_t)}, y^{(q_t)})\}$
  (Unified Update Step)
  $\mathcal{L}_{\text{comb}} \leftarrow \sum_{(x,y) \in \mathcal{D}_t^{\text{comb}}} [\mathcal{L}_{\text{pred}}(f_\theta(x), y) + \lambda_{\text{KL}} D_{\text{KL}}(q_\zeta(r \mid c(x)) \| p(r))]$
  $\theta \leftarrow \text{AdamStep}(\theta, \nabla_\theta \mathcal{L}_{\text{comb}})$
  Increment counts for all samples used in $\mathcal{D}_t^{\text{comb}}$
  $t \leftarrow t + 1$
end while
Return $\theta = \{\zeta, \phi\}$.

Algorithm 3 Relevance-Guided Sequential Learning with Latent Concepts for Geospatial Discovery (**Inference**)

Require: Test pool $\mathcal{X}_{\text{test}}$, Query budget $\mathcal{C}_{\text{test}}$
Require: Concept encoder CE_ψ, relevance encoder RE_ζ, decoder RD_ϕ
Require: Trained model parameters θ_{train}, core buffer $\mathcal{C}_{\text{train}}$, reservoir buffer $\mathcal{R}_{\text{train}}$

Initialize $\theta \leftarrow \theta_{\text{train}}$, $\mathcal{C}_0 \leftarrow \mathcal{C}_{\text{train}}$, $\mathcal{R}_0 \leftarrow \mathcal{R}_{\text{train}}$, $\mathcal{U} \leftarrow \mathcal{X}_{\text{test}}$, and $t \leftarrow 1$; ψ remains frozen
while $t \leq \mathcal{C}_{\text{test}}$ **and** $\mathcal{U} \neq \emptyset$ **do**
    (Scoring)
    for each $x \in \mathcal{U}$ **do**
        $\tilde{c}(x) \leftarrow \text{CE}_\psi(x)$
        $c(x) \leftarrow \text{GS}(\tilde{c}(x))$
        $(\mu_\zeta(c(x)), \sigma_\zeta(c(x))) \leftarrow \text{RE}_\zeta(c(x))$
        Sample $r(c(x)) \sim \mathcal{N}(\mu_\zeta(c(x)), \text{diag}(\sigma_\zeta^2(c(x))))$
        $f_\theta(x) \leftarrow \text{RD}_\phi(c(x), r(c(x)))$
        $w_t(x) \leftarrow \sum_{x' \in \mathcal{C}_{t-1}} \|\mu_\zeta(c(x)) - \mu_\zeta(c(x'))\|_2^2$
        $\exp(w_t(x))$ denotes the relevance-space novelty weight of region x
        $u_{\text{dec}}(x) \leftarrow$ *prediction uncertainty* of region x using $\sum_{i=1}^{P} \exp^{\left(-|f_\theta(x_i)-0.5|\right)}$
        $S_{\text{exploit}}(x) \leftarrow \exp^{(-w_x)} \sum_{i=1}^{P} \exp^{\left(f_\theta(x_i)\right)}$
        $S_{\text{explore}}(x) \leftarrow \exp^{(w_x)}\ u_{\text{dec}}(x)$
        $\kappa_t \leftarrow \frac{C_{\text{test}}-t}{C_{\text{test}}+t}$
        $a_t(x) \leftarrow \kappa_t S_{\text{explore}}(x) + (1-\kappa_t) S_{\text{exploit}}(x)$
    end for
    $x^{(q_t)} \leftarrow \arg\max_{x \in \{\text{Set of Unobserved regions}\}} a_t(x)$
    $y^{(q_t)} \leftarrow$ Observe ground truth labels corresponds to the region $x^{(q_t)}$.
    (Compute Evaluation Metrics)
    RECORDTARGETS$(x^{(q_t)}, y^{(q_t)})$ update SR and other metrics, such as F-Score, Accuracy, Precision, Recall.
    (Buffer update)
    Insert $(x^{(q_t)}, y^{(q_t)})$ into $\mathcal{D}^{\text{core}}$ with $(\text{duration} = 0, \text{count} = 0)$
    Update durations in $\mathcal{D}^{\text{core}}$ and $\mathcal{D}^{\text{reservoir}}$, evict by lifespan
    Move evicted core samples to $\mathcal{D}^{\text{reservoir}}$, drop excess reservoir samples
    (Meta-batch formation)
    Cluster elements of $\mathcal{C}_t$ in relevance space
    From each cluster, select a sample maximizing $\exp(\text{duration}/(\text{count}+1))$ to form $\mathcal{D}_t^{\text{core}}$
    Select K_r samples from $\mathcal{R}_t$ with weights $\propto \exp(\text{duration}/(\text{count}+1))$ to form $\mathcal{D}_t^{\text{res}}$
    $\mathcal{D}_t \leftarrow \mathcal{D}_t^{\text{core}} \cup \mathcal{D}_t^{\text{res}}$
    $D_t^{\text{comb}} \leftarrow D_t \cup \{(x^{(q_t)}, y^{(q_t)})\}$
    (Unified Update Step)
    $\mathcal{L}_{\text{comb}} \leftarrow \sum_{(x,y) \in \mathcal{D}_t^{\text{comb}}} [\mathcal{L}_{\text{pred}}(f_\theta(x), y) + \lambda_{\text{KL}} D_{\text{KL}}(q_\zeta(r \mid c(x)) \| p(r))]$
    $\theta \leftarrow \text{AdamStep}(\theta, \nabla_\theta \mathcal{L}_{\text{comb}})$
    Increment counts for all samples used in $\mathcal{D}_t^{\text{comb}}$
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{x^{(q_t)}\}$
    $t \leftarrow t+1$
end while
Return Discovered targets and final metrics.

V Experimental Setup of Domain-Informed Greedy Baseline

Instead of running a simulation like SWAT directly, which is designed for general watershed hydrology modeling but lacks established parameterization for PFAS transport, we implemented a streamlined Python-based workflow more tailored to PFAS-related surface water contamination in collaboration with Water Quality models' experts. This approach allowed us to integrate domain-relevant features without relying on assumptions unsupported by current PFAS science.

First, we assign initial pollutant concentrations in each patch by setting cells flagged as "dischargers" to a high baseline value (e.g., 100), while land cover–based default concentrations provide moderately-

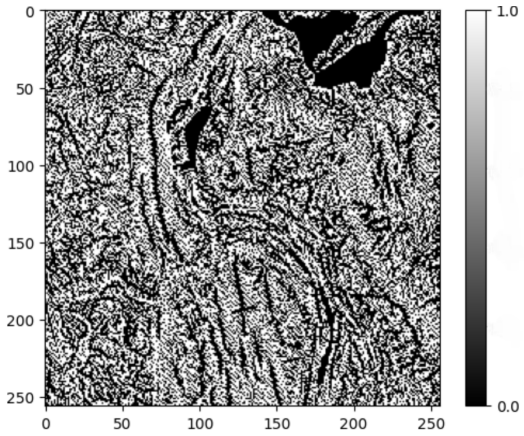


Figure 11: Example binary output from the pollutant transport simulation: 0 represents low contamination, and 1 represents high contamination. The simulation approximates the distribution of PFAS contamination based on hydrological and environmental parameters, providing a practical alternative to full-scale SWAT simulations.

low to low initial contamination levels for other cells. We then construct a Hydrologic Response Unit (HRU) parameter table that assigns infiltration and runoff values based on a cell's land cover, soil type, and slope. Using these parameters, each patch's land cover and soil rasters define infiltration/runoff ratios for every grid cell. We incorporate two key rasters; flow direction and flow accumulation, both exported from System for Automated Geoscientific Analyses (SAGA); a GIS and geospatial analysis tool focused on geospatial data processing, analysis, and visualization, to specify how water and pollutant mass move downstream. In each iteration, a fraction of the pollutant mass in each cell is transferred to its downstream neighbor according to the cell's infiltration/runoff factors, guided by the flow direction raster and scaled by the flow accumulation values. Repeating this process causes pollutant mass to concentrate in lower-lying or higher-flow cells, thereby modeling how contamination evolves over time within the patch.

After the simulation converges, we produce a final pollutant concentration raster, thresholded by the median value of concentrations across all patches for the year to generate a binary contamination map, as illustrated in Fig. 11. The binary outputs, with values of 0 and 1 representing low and high contamination respectively, visualizes the simulation's predictions of contamination distribution. These outputs are compared against test set patches containing actual observed PFAS presence *in surface water* to evaluate how accurately the simulation captures observed contamination patterns. Although this standalone approach remains computationally non-trivial, it is more tractable for batch processing of multiple patches than running a full SWAT project repeatedly. Consequently, it provides a practical baseline for pollutant transport modeling within our broader framework, allowing us to compare simulated outputs to both real-world data and other modeling approaches.

W Impact Statement

This work advances methods for learning and decision-making under extreme data scarcity and resource constraints, motivated by real-world geospatial discovery tasks such as environmental monitoring and public health risk identification. By enabling models to adapt online using limited, costly observations, our approach has the potential to support more efficient allocation of field sampling resources, reduce monitoring costs, and accelerate the identification of regions requiring further investigation.

In applied settings such as pollution or environmental risk mapping, improved sample efficiency may help inform earlier intervention, guide regulatory attention, and support community-level decision-making when comprehensive data collection is infeasible. At the same time, the framework is designed to complement, not replace, expert judgment and established scientific workflows, and its outputs should be interpreted as decision-support signals rather than definitive assessments.

Potential risks include misuse or over-interpretation of model predictions in high-stakes contexts, particularly if deployed without appropriate domain oversight or validation. To mitigate this, our framework emphasizes interpretability through concept-guided relevance and is evaluated using

publicly available data and transparent metrics. We encourage responsible use of such models alongside domain expertise and community engagement, especially in settings with regulatory or societal implications. Importantly, our approach does not rely on large-scale pretraining or extensive retraining, and, in addition to addressing data and resource constraints, may help constrain computational resource requirements in geospatial monitoring pipelines.

Overall, we believe this work contributes positively by providing tools for more responsible and efficient learning in data-limited environments, while acknowledging the importance of careful deployment and human-in-the-loop decision-making.

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: The abstract and introduction accurately reflect the paper's contributions, including the OWL-GPS setting and relevance-guided sampling and adaptation strategies, and these claims are consistently validated through theoretical analysis and experiments demonstrating improved performance under budget-constrained geospatial discovery.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: [Yes]

   Justification: The paper explicitly discusses limitations in the conclusion, including reliance on domain-specific drivers and potential reduced applicability to unstructured settings, thereby clearly acknowledging constraints and scope of the proposed approach.

   Guidelines:

   - The answer [N/A] 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]

Justification: All theoretical results are clearly stated with corresponding assumptions, and complete proofs are provided in the appendix with appropriate cross-referencing from the main text, ensuring clarity and correctness of the theoretical analysis.

Guidelines:

- The answer [N/A] 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: The paper provides detailed descriptions of the experimental setup, including dataset construction, train/test splits, query budgets, model components, and training/inference procedures, with additional implementation details and hyperparameters specified in the appendix to support reproducibility.

Guidelines:

- The answer [N/A] 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: A repository containing code and data is provided in the main paper, and instructions to reproduce the main experimental results is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies key experimental details, including dataset splits, query budgets, model architecture, and training and inference procedures, with additional hyperparameters and implementation specifics provided in the appendix to ensure clarity of the experimental setup.

   Guidelines:

   - The answer [N/A] 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: The paper reports results with error bars (mean ± standard deviation) computed over multiple runs, and the source of variability is clearly defined, providing an appropriate measure of statistical significance.

Guidelines:

- The answer [N/A] 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).
- 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: The paper specifies the compute environment, including GPU type (NVIDIA A100), runtime per inference step, and memory usage, in the appendix.

   Guidelines:

   - The answer [N/A] 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: The work adheres to the NeurIPS Code of Ethics, relying on publicly available environmental datasets, avoiding the use of sensitive personal data, and focusing on applications that support environmental monitoring and public health.

   Guidelines:

   - The answer [N/A] 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: The paper includes a broader impacts discussion in the appendix, outlining both positive societal impacts as well as potential limitations and risks, ensuring a balanced assessment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not involve the release of high-risk models or sensitive datasets, and therefore does not require specific safeguards for responsible deployment.

    Guidelines:

    - The answer [N/A] 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: The paper properly credits all datasets and models used and specifies that they are publicly available and used in accordance with their respective licenses.

Guidelines:

- The answer [N/A] 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.
- 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 paper introduces a newly constructed PFAS dataset, with details on data sources, preprocessing, and splits provided in the appendix, and makes it available through a repository with accompanying documentation.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not involve research with human subjects, and therefore IRB approval is not applicable.

Guidelines:

- The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology does not involve the use of large language models; any use of such tools was limited to grammar or spelling editing assistance and does not impact the scientific contributions.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.