Pragmatic Curiosity: A Unified Framework for Hybrid Learning and Optimization via Active Inference

Anonymous authors
Paper under double-blind review

Abstract

Many engineering and scientific workflows rely on expensive black-box evaluations, requiring sequential decisions that must both improve task performance and reduce uncertainty. Bayesian optimization (BO) and Bayesian experimental design (BED) provide powerful but largely separate treatments of goal-directed optimization and information-seeking experimentation, leaving limited guidance for hybrid settings in which learning and optimization are intrinsically coupled. We propose **Pragmatic Curiosity (PraC)**, a unified framework for hybrid learning and optimization via active inference. PraC evaluates candidate queries by trading information gain about a task-relevant latent symbol against an expected regret-based potential over outcomes. This formulation foregrounds three operational design choices: which latent quantity should be clarified, how task value is encoded as regret, and how strongly information gain should be exchanged against pragmatic regret. We instantiate PraC across three regimes of increasing complexity: decision-oriented monitoring with fixed global symbols and known downstream losses, targeted active search with induced local symbols and evolving coverage goals, and composite Bayesian optimization with hierarchical regret learning under unknown preferences. Across these regimes, PraC reduces downstream decision risk, improves coverage of critical outcome regions, and jointly learns predictive and preference structures without relying on task-specific staging rules.

1 Introduction

Sequential decision-making under uncertainty often requires an agent to decide what to learn and what to do at the same time. Two classical paradigms have addressed these aspects from different directions. Bayesian optimization (BO) is primarily *goal-directed*: it selects queries in order to identify high-performing inputs of an unknown objective function (Močkus, 1975; Jones et al., 1998; Srinivas et al., 2009). Bayesian experimental design (BED) is primarily *information-seeking*: it selects experiments in order to reduce uncertainty about latent quantities of interest (Chaloner & Verdinelli, 1995). Although both can be viewed as instances of adaptive sampling (Di Fiore et al., 2023), they are usually developed under different directives and therefore admit few directly transferable design principles (Hvarfner et al., 2025).

This separation leaves limited guidance for problems in which the relevant uncertainty, the downstream objective, and the sampling policy must be designed jointly. In environmental monitoring (Konakovic Lukovic et al., 2020), a robot may need to learn the hidden state of a chemical plume while deciding where to dispatch a response team. In failure discovery (Ramanagopal et al., 2018; Parashar et al., 2025), an evaluator may need to explore uncertain scenarios while prioritizing outcomes that cover critical failure regions. In preference-guided design (González & Zavala, 2025; Coelho et al., 2025), an optimizer may need to learn not only how actions produce outcomes, but also how those outcomes should be valued. In all these settings, an agent must act before it fully knows, and it only learns by acting. The bottleneck is therefore not learning or optimization alone, but the *meta-decision* that connects them: deciding what knowledge is sufficient for action, and what actions are most informative for future decisions.

Recent work has begun to address this coupling by choosing between specialized tools and accommodating problem-specific adaptation: leveraging information-gain criteria to enhance optimization, or vice versa.

Information-directed sampling (IDS) connects information and regret in online optimization and bandit settings (Russo & Van Roy, 2018). Self-correcting Bayesian optimization introduces a statistical distance-based active learning criterion into the BO loop to improve model learning during optimization (Hvarfner et al., 2023). Expected predictive information gain (EPIG) focuses active learning on predictive information that is relevant to a target distribution (Smith et al., 2023). Preference exploration for BO learns preference models over multi-objective outcomes, but typically relies on stage-wise choices between experimentation and preference learning (Lin et al., 2022). These methods highlight growing synergy between learning and optimization, but they often remain task-specific, and rarely generalize across problem categories.

In this paper, we propose **Pragmatic Curiosity (PraC)**, a unified framework for hybrid learning and optimization via active inference (AIF) (Friston, 2010; Friston et al., 2017). Originally developed in computational neuroscience, AIF prescribes action selection by minimizing *expected free energy* (EFE), a single objective that couples an *epistemic* drive for information gain with a *pragmatic* drive toward preferred outcomes. PraC translates this conceptual principle into a practical acquisition-design template: a candidate query is evaluated by both what it is expected to reveal about a task-relevant latent quantity and how its predicted outcome is expected to affect downstream action. In this sense, PraC formalizes *pragmatic curiosity*: curiosity is not the indiscriminate gathering of information, but the search for knowledge insofar as that knowledge can improve action pragmatically.

PraC not only serves as a unifying umbrella principle that re-interprets many classical acquisition rules in BO and BED, but also foregrounds a *triad* of meta-decisions that govern any hybrid decision-making problems: *representation*, *evaluation*, and *regulation*. In PraC, this triad is conceptualized as *symbols*, *regrets*, and *curiosity*. Symbols specify the latent quantities whose clarification can change what the agent ought to do; regrets provide a modest objective by quantifying task-relevant shortfall; and curiosity sets the exchange rate between epistemic clarification and pragmatic sacrifice. This triad turns PraC from an abstract decision paradigm into a concrete decision rule: it specifies what is worth knowing, what is worth pursuing or avoiding, and how strongly unresolved uncertainty should influence action. It also elevates curiosity from a static exploration heuristic into a regulated gain in the knowledge–action loop, motivating a feedforward–feedback scheduler that adapts curiosity as belief and action co-evolve.

We instantiate PraC across three regimes of increasing complexity. First, in decision-oriented monitoring, PraC uses fixed global symbols and known downstream losses to guide sensing toward environmental distinctions that matter for response decisions. Second, in targeted active search, PraC uses induced local symbols and evolving coverage goals to coordinate exploration in input space with discovery in outcome space. Third, in composite Bayesian optimization with unknown preferences, PraC uses hierarchical regret learning to jointly learn how actions produce outcomes and how those outcomes should be valued. Across these regimes, PraC reduces downstream decision risk, improves coverage of critical outcome regions, and supports joint outcome–regret learning without relying on task-specific staging rules.

Our contributions are as follows.

- **A unified acquisition principle.** We derive PraC as an active-inference acquisition principle that couples task-relevant information gain with goal-directed pragmatic regret.
- **A triad of operational design axes.** We formalize epistemic symbols, regret potentials, and curiosity coefficients as the core meta-decisions underlying hybrid decision-making problems.
- **Empirical validation across three regimes.** We validate PraC in decision-oriented monitoring, targeted active search, and preference-guided composite optimization, spanning fixed global, induced local, and hierarchical symbols under both known and unknown goals.

2 Preliminaries

2.1 Bayesian Optimization

Given an unknown objective function $f : \mathcal{X} \to \mathbb{R}$, BO seeks to identify the input x^* that maximizes the objective f over an admissible set of queries $\mathcal{X}$, *i.e.*, $x^* = \arg\max_{x \in \mathcal{X}} f(x)$. Let the available information regarding

the objective function f be stored in the dataset $\mathcal{D}_t := \{(x_1, y_1), \ldots, (x_t, y_t)\}$, where $y_t \sim \mathcal{N}(f(x_t), \sigma^2(x_t))$ is the noisy observation of the objective function by assuming the noise follows a zero-mean normal distribution with a standard deviation $\sigma(x)$. BO relies on a *surrogate model* $p(f(x)|\mathcal{D}_t)$ that provides a probabilistic representation of the objective f, and uses this information to compute an *acquisition function* $\alpha : \mathcal{X} \to \mathbb{R}$ to drive the selection of the most promising sample to query. Many acquisition functions have been proposed, including *probability of improvement* (Močkus, 1975), *expected improvement* (Jones et al., 1998), *upper confidence bound*, and various *entropy search* methods (Hennig & Christian J. Schuler, 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017; Hvarfner et al., 2022; Neiswanger et al., 2021), as well as practical approaches to optimize them (Wilson et al., 2018).

2.2 Bayesian Experimental Design

Rather than optimizing an objective function $f(x)$, the purpose of BED is to sequentially select a set of experimental designs $x \in \mathcal{X}$ and gather outcomes y, to maximize the amount of information obtained about certain *parameters of interest*, denoted as θ. The parameters θ can correspond to some explicit model parameters, or any implicitly defined quantity of interest (*e.g.*, the optimum of a function, the output of an algorithm, or future downstream predictions). BED finds the next experimental design x_{t+1} by maximizing the *expected information gain* (EIG) (Chaloner & Verdinelli, 1995) that a potential experimental outcome y_{t+1} can provide about θ, measured in terms of expected entropy reduction of the posterior distribution of θ:

$$\mathrm{EIG}(x \mid \mathcal{D}_t) = H[p(\theta|\mathcal{D}_t)] - \mathbb{E}_{p(y|x,\mathcal{D}_t)}[H[p(\theta|\mathcal{D}_t \cup \{(x,y)\})]] = I(\theta; (x,y) \mid \mathcal{D}_t),$$

where $H(\cdot)$ and $I(\cdot)$ denote the entropy and mutual information, respectively.

3 Pragmatic Curiosity: A Framework that Synthesizes Learning and Optimization

Acquisition strategies in BO typically pursue *goal-directed* behavior, whereas those in BED emphasize *information-seeking* behavior. In this section, we synthesize these seemingly competing imperatives through the lens of active inference (AIF), yielding a unified framework for hybrid learning and optimization.

3.1 Active Inference as Expected Free Energy Minimization

Consider a probabilistic surrogate model $q(\cdot)$ that captures the relationship between a decision variable x, an outcome y, and a collection of latent quantities of interest s, and factorizes as

$$q(x, y, s) \coloneqq p(x, y \mid s)\, q(s),$$

where $q(s)$ is the surrogate belief over s. We use $q(\cdot)$ to distinguish the agent's internal surrogate model from the underlying, generally inaccessible, true data-generating model $p(\cdot)$.

In AIF, preferences over possible outcomes are encoded through a probability distribution $p(y)$. Outcomes assigned higher probability are treated as more preferred. The deviation between an observed outcome and those preferred by the agent is measured by its *self-information*, or *surprisal*, $-\log p(y)$. Intuitively, surprisal quantifies how unexpected an outcome y is under the preference distribution $p(y)$: less preferred outcomes incur larger surprisal. The surprisal associated with an outcome y satisfies

$$-\log p(y) = -\log \int \frac{p(y,s)q(s)}{q(s)}\, ds = -\log \mathbb{E}_{q(s)}\left[\frac{p(y,s)}{q(s)}\right] \leq -\mathbb{E}_{q(s)}\left[\log \frac{p(y,s)}{q(s)}\right] = F, \tag{1}$$

where the inequality follows from *Jensen's inequality*.

The right-hand side of equation 1 is called the *variational free energy* (VFE), resembling the *Helmholtz free energy* in physics. It upper bounds surprisal and therefore provides a tractable surrogate objective for inference. In machine learning, the sign of VFE is often reversed, yielding the evidence lower bound (ELBO), whose maximization is a standard variational learning principle (Titsias, 2009).

To obtain a decision rule, we must account not only for realized outcomes, but also for future outcomes that have not yet been observed. AIF does so by considering the expected surprisal of future outcomes under the predictive distribution $q(y \mid x)$:

$$\begin{aligned} -\mathbb{E}_{q(y|x)} \log p(y \mid x) &\leq -\mathbb{E}_{q(y,s|x)} \left[\log \frac{p(y, s \mid x)}{q(s \mid x)} \right] = -\mathbb{E}_{q(y,s|x)} \left[\log \frac{p(s \mid x, y) p(y \mid x)}{q(s \mid x)} \right] \\ &= -\mathbb{E}_{q(y,s|x)} \left[\log p(s \mid x, y) - \log q(s \mid x) \right] - \mathbb{E}_{q(y|x)} \log p(y \mid x) = G, \end{aligned} \tag{2}$$

where the right-hand side of equation 2 is denoted as the *expected free energy* (EFE). AIF prescribes action selection by minimizing this quantity.

3.2 A Principled Decision Paradigm for Hybrid Learning and Optimization

In its original form, the EFE in equation 2 is not yet an actionable acquisition principle, because it depends on the true posterior $p(s \mid x, y)$ and the true outcome model $p(y \mid x)$, both of which are generally unavailable *a priori*. We now turn it into a principled decision paradigm through two approximations that are natural from a decision-theoretic perspective.

The EFE in equation 2 can be decomposed as

$$G = \underbrace{-\mathbb{E}_{q(y,s|x)} \left[\log p(s \mid x, y) - \log q(s \mid x) \right]}_{\text{Term 1}} \underbrace{-\mathbb{E}_{q(y|x)} \log p(y \mid x)}_{\text{Term 2}}.$$

For Term 1, add and subtract $\log q(s \mid x, y)$ inside the expectation:

$$\begin{aligned} \text{Term 1} &= -\mathbb{E}_{q(y,s|x)} \left[\log p(s \mid x, y) - \log q(s \mid x, y) + \log q(s \mid x, y) - \log q(s \mid x) \right] \\ &= -\mathbb{E}_{q(y|x)} \mathbb{E}_{q(s|x,y)} \left[\log p(s \mid x, y) - \log q(s \mid x, y) + \log q(s \mid x, y) - \log q(s \mid x) \right] \\ &= -\mathbb{E}_{q(y|x)} \left[\mathbb{E}_{q(s|x,y)} \left[\log p(s \mid x, y) - \log q(s \mid x, y) \right] + \mathbb{E}_{q(s|x,y)} \left[\log q(s \mid x, y) - \log q(s \mid x) \right] \right] \\ &= -\mathbb{E}_{q(y|x)} \left[\underbrace{-D_{\mathrm{KL}}(q(s \mid x, y) \| p(s \mid x, y))}_{\text{Term 3}} + \underbrace{D_{\mathrm{KL}}(q(s \mid x, y) \| q(s \mid x))}_{\text{Term 4}} \right]. \end{aligned}$$

Since the surrogate belief over s is not updated until an observation is obtained, it does not depend on x alone; hence $q(s \mid x) = q(s)$. Therefore, Term 4 becomes $D_{\mathrm{KL}}(q(s \mid x, y) \| q(s))$, which is precisely the *epistemic value*: the information gained about the latent quantity s from a prospective observation (x, y).

By contrast, Term 3 remains intractable because it depends on the true posterior $p(s \mid x, y)$. The best the agent can access is its current surrogate belief constructed from the historical data $\mathcal{D}_t$, namely $q(\cdot) = p(\cdot \mid \mathcal{D}_t)$. This motivates the *first approximation*: we treat the surrogate belief as the best currently available approximation to the true model, *i.e.*, $q(s) \approx p(s)$. If both the surrogate and true models update according to *Bayes' rule*, $q(s \mid x, y) = \frac{p(x,y|s)q(s)}{\int p(x,y|s)q(s)ds}$, $\quad p(s \mid x, y) = \frac{p(x,y|s)p(s)}{\int p(x,y|s)p(s)ds}$, then Term 3 vanishes under this approximation, and the resulting decision rule is Bayes-optimal with respect to the agent's current belief. This is the only legitimate basis for action: a decision-maker cannot optimize with respect to an inaccessible ground truth, but only with respect to its present posterior approximation. Moreover, under standard consistency conditions, the discrepancy between surrogate and truth shrinks as more data are collected.

The second obstacle lies in Term 2, which depends on the true predictive distribution $p(y \mid x)$. We therefore introduce a *second approximation*: we replace $p(y \mid x)$ by a preference distribution $p_{\mathrm{pref}}(y)$ that does not depend on x. This converts Term 2 into $-\mathbb{E}_{q(y|x)} \log p_{\mathrm{pref}}(y)$, which evaluates preferred outcomes under the predictive model induced by choosing x. Intuitively, the agent now asks: if I take action x, what outcomes am I likely to see, and how desirable are those outcomes?

To define a preference distribution from a task-level notion of desirability, we introduce a Boltzmann operator $\mathcal{B}_\beta$ that maps a nonnegative potential energy function $h : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ into a probability distribution:

$$(\mathcal{B}_\beta[h])(z) \coloneqq \frac{e^{-h(z)/\beta}}{\int_{\mathcal{Z}} e^{-h(z)/\beta} dz}.$$

The parameter β plays the role of temperature: larger β yields a flatter, higher-entropy distribution, whereas smaller β concentrates more sharply around the minimizers of h.

Applying this operator to a possibly time-varying potential energy function $h(y \mid \mathcal{D}_t)$ gives a preference distribution over outcomes, and hence

$$G \approx -I(s;(x,y) \mid \mathcal{D}_t) + 1/\beta \cdot \mathbb{E}_{q(y|x,\mathcal{D}_t)}[h(y \mid \mathcal{D}_t)] + Z,$$

where $q(y \mid x, \mathcal{D}_t)$ is the predictive distribution of a surrogate model constructed from the historical data $\mathcal{D}_t$, and $Z = \log \int_{\mathcal{Y}} e^{-h(y|\mathcal{D}_t)/\beta} dy$ is a normalization constant independent of x.

This leads to our proposed decision paradigm:

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \{\beta_t \underbrace{I(s;(x,y) \mid \mathcal{D}_t)}_{\text{epistemic}} - \underbrace{\mathbb{E}_{q(y|x,\mathcal{D}_t)}[h(y \mid \mathcal{D}_t)]}_{\text{pragmatic}}\}.$$

Here, the conditional mutual information $I(s;(x,y) \mid \mathcal{D}_t)$ captures the *epistemic* drive to reduce uncertainty about latent quantities of interest; the expected potential $\mathbb{E}[h(y \mid \mathcal{D}_t)]$ captures the *pragmatic* drive to avoid undesirable outcomes; and $\beta_t \geq 0$ controls the trade-off between those two.

By construction, this decision paradigm balances *information-seeking* and *goal-directed* behavior within a single objective. Rather than treating exploration and exploitation as competing heuristics, it expresses them as two inseparable aspects of one decision criterion: the agent seeks information insofar as it improves outcome-sensitive action. In other words, it demonstrates a **pragmatic curiosity (PraC)**.

3.3 A Unified View of Acquisition Strategies in BO and BED

PraC serves as a unifying umbrella principle for both learning-oriented and optimization-oriented acquisition strategies. By varying the epistemic target s and the pragmatic potential $h_t(y)$, many classical acquisition rules from BO and BED can be recovered as special cases or close approximations, as summarized in Table 1.

Table 1: A unified view of different acquisition strategies in BO and BED, where x^*, y^* represent the true optimal solution and value, respectively, and $\hat{y}$ is the best value observed in $\mathcal{D}_t$.

Acquisition Strategy	Acquisition Function	Pragmatic Curiosity	
		s	$h_t(y)$
Expected Information Gain (Chaloner & Verdinelli, 1995)	$I(\theta;(x,y) \mid \mathcal{D}_t)$	θ	-
Entropy Search (Hennig & Christian J. Schuler, 2012)	$I(x^*;(x,y) \mid \mathcal{D}_t)$	x^*	-
Max-value Entropy Search (Wang & Jegelka, 2017)	$I(y^*;(x,y) \mid \mathcal{D}_t)$	y^*	-
Joint Entropy Search (Hvarfner et al., 2022)	$I((x^*,y^*);(x,y) \mid \mathcal{D}_t)$	(x^*,y^*)	-
Bayesian Algorithm Execution (Neiswanger et al., 2021)	$I(\mathcal{O}_{\mathcal{A}}(f);(x,y) \mid \mathcal{D}_t)$	$\mathcal{O}_{\mathcal{A}}(f)$	-
GP-Upper Confidence Bound [1] (Srinivas et al., 2009)	$\mu_t(x) + \beta^{1/2}\sigma_t(x)$	$f_{\mathcal{X}}$	$-y$
Probability of Improvement (Močkus, 1975)	$p(y \geq \hat{y})$	-	$-\mathbb{I}(y \geq \hat{y})$
Expected Improvement (Jones et al., 1998)	$\mathbb{E}([y-\hat{y}]_+)$	-	$-[y-\hat{y}]_+$

More importantly, this unifying principle opens the door to acquisition design for a broader class of *hybrid* learning-and-optimization problems that go beyond classical BO and BED. Such problems are typically governed by three coupled meta-decisions: (i) *what the agent seeks to know*, (ii) *what it seeks to achieve*, and (iii) *how it trades off those two*. In the next section, we formalize these three design axes through *symbols*, *regrets*, and *curiosity*.

[1] Strictly speaking, the correspondence for GP-UCB is approximate rather than exact, because GP-UCB uses a first-order standard-deviation term rather than a second-order variance term. Detailed derivations are provided in Appendix B. Still, the connection reveals the information-theoretic structure underlying this heuristic form.

4 The Triad of Meta-Decisions: Symbols, Regrets, and Curiosity

One central contribution of PraC is that it foregrounds three meta-decisions underlying hybrid learning and optimization: what latent quantity should be clarified, encoded by s; how outcomes should be evaluated, encoded by h_t; and how strongly unresolved uncertainty should influence action, encoded by β_t.

These choices are the operational levers through which an abstract decision paradigm becomes a concrete decision rule. This section develops them both conceptually and operationally: *symbols* specify the latent quantities whose clarification can change action; *regrets* provide an outcome-level landscape of task-relevant shortfall; and *curiosity* regulates the exchange rate between clarifying knowledge and pursuing goals.

4.1 Symbols and Models

"All models are wrong, but some are useful." A model becomes useful when it represents the distinctions whose clarification can change what we should do. In PraC, we call these distinctions *symbols*. Different modeling regimes induce different forms of symbols.

Parametric models. In parametric models, symbols are typically *fixed global symbols*: the latent state is represented by a finite-dimensional parameter whose components retain persistent semantic identities across decisions. Examples include unknown physical coefficients, latent modes, transition parameters, or class labels. In this regime, the epistemic objective is to reduce uncertainty about these stable hidden meanings.

Non-parametric models. In non-parametric models, there may be no finite set of globally privileged coordinates. Instead, the relevant symbols are often *induced local symbols*, created by the contemplated observations themselves. This is formalized by the following property.

Lemma 4.1 (**Representation Compression**). *Let $\mathbf{X} \subseteq \mathcal{X}$ be a subset of inputs, and let $f_{\mathbf{X}}$ denote the corresponding function values. Given a historical dataset $\mathcal{D}_t$ and new measurements $\mathbf{Y}$ observed at $\mathbf{X}$, then*

$$I(f_{\mathcal{X}}; (\mathbf{X}, \mathbf{Y}) \mid \mathcal{D}_t) = I(f_{\mathbf{X}}; \mathbf{Y} \mid \mathcal{D}_t),$$

for any (finite or infinite) set $\mathcal{X}$.

Proof. See Appendix A. □

Lemma 4.1 shows that, under a non-parametric model such as a Gaussian process, the information gained from observing $\mathbf{Y}$ at $\mathbf{X}$ is mediated entirely through the function values $f_{\mathbf{X}}$ that directly generate those observations. These local evaluations play a role analogous to parameters in parametric models or latent states in partially observed dynamical systems, but they need not belong to a fixed global representation.

This compression reveals both the power and the cost of non-parametric representation. The power is parsimony: the agent need not reason about every component of an infinite-dimensional latent object when only a local slice can influence the prospective observation. The cost is semantic instability: because the symbol is query-dependent, its meaning is induced by the decision context rather than fixed in advance.

This also suggests a broader role for PraC. PraC asks not merely how expressive a model is, but whether the symbol used by the acquisition is sufficient for action evaluation, minimal enough to avoid epistemic redundancy, and structured enough to support interpretable decision-making. Thus beyond selecting actions, PraC can also guide the construction of representations. When the relevant structure is unclear, induced local symbols can reveal which distinctions are repeatedly useful for decision. Over time, such recurring local distinctions may be consolidated into more stable global symbols. In this sense, PraC does not only act within a representation; it also provides a criterion for which representations deserve to persist.

4.2 Regrets and Goals

"All successes are successful alike, all failures are failed in their own way." In many open-ended decision problems, the goal is not fully known until we arrive there. It seems easier to address what it is not than what

it is: success is often rare and difficult to characterize in full, whereas shortfall, inconsistency, or violation is much more common and easier to recognize and certify. This makes *regret* a more modest object than reward. A reward function typically imposes a complete preference ordering over outcomes. A regret function can impose only a partial ordering: it separates outcomes by their degree of task-relevant shortfall, while leaving outcomes with the same shortfall comparable only up to indifference. Under uncertainty, this prevents the agent from hallucinating fine-grained preferences where the task does not justify them.

Regret therefore can be regarded as the shape cast by a goal onto outcome space: it gives a conservative representation of what matters by quantifying how an outcome falls short of what is desired. We begin by formally defining this generalized notion of regret.

Definition 4.2 (**Regret Function**)**.** Let $\mathcal{Y}$ denote the outcome space, and let ξ denote a desired reference, such as an optimal outcome, a target set, a safety specification, or a preference criterion. A regret function

$$r(\cdot;\xi) : \mathcal{Y} \to \mathbb{R}_{\geq 0}$$

quantifies how an outcome y falls short of the desired reference ξ. It obtains 0 for the desired outcomes, *i.e.*, $\inf_{y\in\mathcal{Y}} r(y;\xi) = 0$, with smaller values indicating outcomes that are closer to what is desired.

This abstraction recovers several important cases:

- When $\xi = y^*$ is the global optimum of an objective function, $r(y;\xi)$ reduces to the conventional instantaneous regret in BO.
- When no outcome-dependent pragmatic preference is imposed, as in pure BED, one may take $r(y;\xi) \equiv 0$, so that decision-making is driven entirely by epistemic value.
- Intermediate choices recover hybrid objectives that jointly encode performance, safety, feasibility, or other task-specific desiderata.

In practice, however, the ideal regret function $r(\cdot;\xi)$ is generally unknown, implicit, or computationally intractable. A decision-maker therefore cannot act directly on r. Instead, it must rely on a belief-conditioned surrogate that reflects its current internal assessment of outcome desirability:

Definition 4.3 (**Potential Energy Function**)**.** The potential energy function is a belief-conditioned surrogate regret

$$h_t : \mathcal{Y} \to \mathbb{R}_{\geq 0}, \qquad h_t(y) := h(y \mid \mathcal{D}_t),$$

that represents the agent's current internal assessment of the undesirability of outcome y based on the information available up to time t. Thus, h_t serves as a time-varying, belief-dependent approximation of the true regret structure, and is updated recursively as new data are acquired.

The distinction between r and h_t is what makes PraC operational. This mirrors the epistemic approximation used earlier in deriving PraC: just as the inaccessible true posterior is replaced by the current surrogate belief, the inaccessible true regret landscape is replaced by a belief-conditioned potential over outcomes. As data accumulate, both the predictive model and the potential energy evolve, so that the agent's internal representation of desire is refined together with its understanding of the world. This distinction also clarifies how h_t should be designed in different regimes.

Known goals. When the goal is known, h_t can be designed to directly encode domain knowledge. Examples include classical regret with respect to an optimum, distance to a target set, penalties for constraint violation, or weighted combinations of task performance and safety requirements. In this regime, the role of h_t is to express a known desiderative structure in a form that can be evaluated under the predictive model.

Unknown goals. When the goal is unknown or only partially specified, the regret itself must be learned. In this regime, the pragmatic term should be endowed with additional structure, so that the agent can infer not only how the world behaves, but also what outcomes should count as desirable or undesirable. A natural construction is hierarchical: outcome desirability is mediated by additional latent variables or higher-level regret models. Given an unknown regret model $g : \mathcal{Y} \to \mathbb{R}_{\geq 0}$, one can define h_t through an inner EFE over that regret model, yielding the nested acquisition

$$\alpha(x \mid \mathcal{D}_t) = \beta_t I(f_{\mathcal{X}}; (x, y) \mid \mathcal{D}_t) + \mathbb{E}_{q(y|x,\mathcal{D}_t)} \left[\, \gamma_t I(g_{\mathcal{Y}}; (y, z) \mid \mathcal{D}_t) - \mathbb{E}_{q(z|y,\mathcal{D}_t)}[z] \, \right], \tag{3}$$

where $\beta_t, \gamma_t \geq 0$, and $z \sim g(y)$ encodes the regret of outcomes.

In this way, the agent does not optimize with respect to a fully specified reward function given in advance; rather, it learns a surrogate of regret jointly with the predictive model. This yields a richer and more flexible account of the interplay between learning and optimization, and enables multiple layers of inference to be composed within a single decision rule. For instance, as we will show later in 5.3, the model $g(y)$ can be learned through preference feedback.

This point highlights a key advantage of the PraC paradigm. Unlike many mainstream approaches that presuppose a fully specified reward function, PraC requires only a current surrogate of what can already be recognized as desirable or regrettable, together with a mechanism for refining that surrogate as evidence accumulates. One need not begin with complete illumination; it is enough to start by seeing a light in the darkness, and to let both understanding of the world and understanding of the goal sharpen together.

4.3 The Degree of Curiosity

If symbols specify what hidden distinctions are worth learning, and regrets specify what desired outcomes are worth pursuing, then the coefficient β_t determines how much immediate pragmatic sacrifice the agent is willing to tolerate in order to refine its epistemic understanding. In this sense, β_t is the exchange rate between clarifying knowledge and pursuing goals, which we denote as the degree of *curiosity*.

Different regimes of β_t produce different modes of behavior. When $\beta_t = 0$, the agent treats its current symbols as fixed and acts myopically with respect to the present surrogate of regret. At the other extreme, a very large β_t prioritizes information acquisition even when that information has weak pragmatic relevance. Neither extreme is satisfactory: seeking knowledge without goals degenerates into aimless information gathering, whereas pursuing goals without clarified meanings risks acting on an incomplete understanding of the world.

This reveals a closed-loop structure between knowledge and action. The agent acts according to its current symbolic understanding and regret surrogate; the resulting observation then reshapes that understanding and changes which future actions become desirable. Thus, curiosity should not be viewed as a static exploration heuristic. Rather, it is a regulated gain in this knowledge–action loop: it determines how strongly unresolved uncertainty should influence action when the agent's current understanding may still be insufficient for reliable decision-making. This motivates designing β_t as an adaptive controller for this knowledge–action loop.

Feedforward scale. The feedforward component estimates the scale of curiosity required for epistemic value to be comparable with pragmatic value. Given a finite candidate set $\mathcal{X}_t^{\text{cand}}$, let $\widehat{I}_t(x) \approx I(s; (x, y) \mid \mathcal{D}_t)$ estimates the epistemic information gain of querying x, and $\widehat{H}_t(x) := \mathbb{E}[h_t(y) \mid x, \mathcal{D}_t]$ estimates its expected pragmatic regret. Let the informative candidate set be $\mathcal{X}_t^{\text{info}} = \{x \in \mathcal{X}_t^{\text{cand}} : \widehat{I}_t(x) > \varepsilon_I\}$, where $\varepsilon_I > 0$. With a quantile level $\rho \in (0, 1)$, the feedforward scale is given as

$$\beta_t^{\text{ff}} = \text{Quantile}^{(\rho)}_{x \in \mathcal{X}_t^{\text{info}}} \frac{\widehat{H}_t(x)}{\widehat{I}_t(x) + \varepsilon_I},$$

which estimates the local exchange rate between information and regret: when useful information is scarce relative to pragmatic regret, the required curiosity scale increases; when information is readily available, a smaller scale is sufficient to make information gain compete with immediate pragmatic regret.

Feedback activation. The feedback component determines whether an epistemic pressure is still needed. Let U_t be a scalar uncertainty measure over the relevant symbol, such as posterior entropy, and let $U^\star$ denote the desired epistemic equilibrium, typically zero or an irreducible uncertainty floor. Define

$$\beta_t^{\text{fb}} = k_\beta \, \text{clip}\left(\frac{U_t - U^\star}{U_0 - U^\star + \varepsilon_U}, 0, 1 \right),$$

where U_0 is the initial uncertainty, k_β is a feedback gain, and $\varepsilon_U > 0$ prevents numerical instability.

The combined schedule for curiosity coefficient is then

$$\beta_t = \text{clip}\left(\beta_t^{\text{fb}} \beta_t^{\text{ff}}, \beta_{\min}, \beta_{\max} \right),$$

with two components playing complementary roles: the feedforward term estimates how large curiosity must be for information gain to compete with pragmatic regret, while the feedback activation suppresses curiosity as the belief approaches the desired epistemic equilibrium. Thus, curiosity is high only when information is both decision-relevant and epistemically needed; it relaxes when the relevant uncertainty has been resolved.

5 Experiments

We instantiate PraC across three regimes of increasing complexity, organized by the structure exposed by the framework. First, we study *fixed global symbols* with *known goals*, where the latent quantity is a finite hypothesis and regret is downstream Bayes risk. Second, we consider *induced local symbols* with *known but evolving goals*, where the latent quantity is induced by a non-parametric surrogate and regret depends on coverage already obtained. Third, we study a *hierarchical* setting with *unknown goals*, where the agent must learn both how actions produce outcomes and how those outcomes should be valued. We analyze dynamic scheduling of β_t in the first regime, where its closed-loop interpretation is most transparent, and use fixed curiosity weights in the latter two regimes to isolate representation and evaluation design.

5.1 Fixed Global Symbols with Known Goals

We begin with the most transparent regime: the latent structure is fixed and the downstream goal is known. Here the symbol is a finite hypothesis θ, and the task objective can be expressed through a downstream loss. This setting allows us to examine whether *curiosity* can be elevated from an exploration heuristic to a monitored and controllable regulation of the decision process.

In **decision-oriented monitoring**, the agent maintains a posterior over hypotheses θ, selects a sensor query x, observes a measurement y, updates its belief, and evaluates the downstream decision loss. Let $a \in \mathcal{A}$ denote a downstream action and $L(a, \theta)$ the loss incurred under hypothesis θ. Given the current surrogate posterior $q(\theta \mid \mathcal{D}_t)$, the Bayes risk is

$$\mathrm{BR}(\mathcal{D}_t) \coloneqq \min_{a \in \mathcal{A}} \mathbb{E}_{q(\theta \mid \mathcal{D}_t)}[L(a, \theta)].$$

For a candidate query x, define the expected Bayes-risk reduction

$$\Delta\mathrm{BR}_t(x) \coloneqq \mathrm{BR}(\mathcal{D}_t) - \mathbb{E}_{q(y \mid x, \mathcal{D}_t)}\left[\mathrm{BR}(\mathcal{D}_t \cup \{(x, y)\})\right].$$

The PraC acquisition is therefore

$$\alpha_t(x) = \beta_t I(\theta; (x, y) \mid \mathcal{D}_t) + \Delta\mathrm{BR}_t(x).$$

Tasks. We evaluate three 2D plume-monitoring tasks. *Source response localization* asks the agent to dispatch a response team near the true source, using normalized squared response distance as loss. *Consequence-weighted dispatch* uses the same response structure, but assigns larger penalties to errors in high-consequence regions. *Active source prioritization* asks the agent to select a limited subset of suspected leaks for repair, with loss given by the weighted risk of true active sources left unrepaired. Details are provided in Appendix C.2.

Baselines and metrics. We compare dynamic PraC, fixed-β PraC with $\beta = 5$, random sampling, pure information gain (EIG), and a decision-greedy policy that maximizes $\Delta\mathrm{BR}_t(x)$. We report posterior Bayes risk and task-specific realized decision metrics: response loss for source localization and consequence-weighted dispatch, and missed-source risk for active source prioritization. Additional metrics, including response success, parameter error, top-k recall, weighted top-k recall, and detailed analysis are reported in Appendix C.2.6.

Results. Fig. 1 shows that dynamic PraC gives the strongest overall performance. In source response localization, it achieves the lowest final Bayes risk while matching the best realized response loss. In consequence-weighted dispatch, decision-greedy obtains the lowest posterior Bayes risk, but dynamic PraC achieves the best realized response loss and lowest parameter error, indicating that reducing posterior risk under the current belief need not coincide with improving the true downstream decision. In active source prioritization, dynamic PraC obtains the lowest final Bayes risk, while both PraC variants substantially

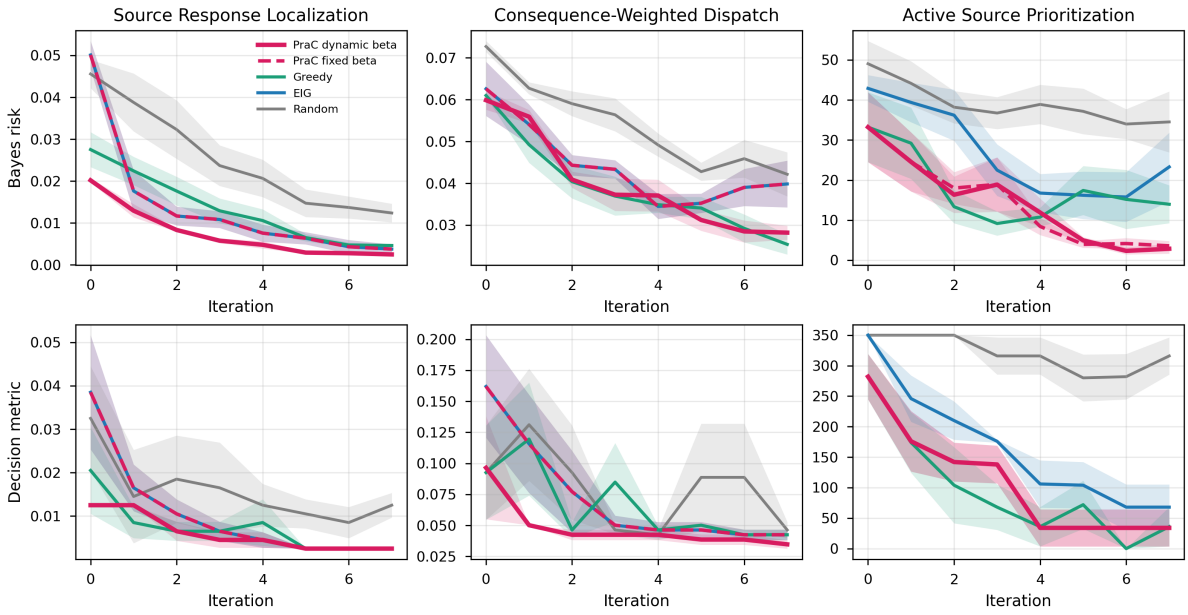

Figure 1: Performance evaluation for decision-oriented plume-monitoring. Top row: posterior Bayes risk. Bottom row: task-specific realized decision metric: response loss for source response localization and consequence-weighted dispatch, and missed-source risk for active source prioritization. Shaded regions represent ± 1 standard deviation over 20 seeds.

improve missed-source risk over EIG and random sampling. Scheduler diagnostics in Appendix C.2.7 further support that dynamic β_t calibrates epistemic pressure according to Bayes-risk reduction and decision-relevant uncertainty, functioning as a regulated exchange rate between clarifying knowledge and pursuing goals.

5.2 Induced Local Symbols with Known Evolving Goals

We next consider a regime where the relevant latent structure is not a fixed finite-dimensional parameter, but a local symbol induced by the contemplated observation. The goal remains known, but evolves with the outcomes already collected: once a target region has been covered, nearby outcomes become less valuable. We use a fixed curiosity weight in this subsection to isolate the effect of *symbolic flexibility*.

A representative problem is **targeted active search** in multi-objective design, where the goal is to cover an important outcome region $\mathcal{S}$ rather than optimize a scalar reward. Following Malkomes et al. (2021), let δ define the resolution at which nearby outcomes are considered redundant, and define

$$\mathbb{C}_\delta(y) := \{y' : d(y, y') < \delta\}, \qquad \mathbb{C}_\delta(Y) := \bigcup_{y \in Y} \mathbb{C}_\delta(y).$$

A natural coverage-shortfall potential is

$$h_t(y) = \mathrm{Vol}(\mathcal{S}) - \mathrm{Vol}(\mathbb{C}_\delta(Y \cup \{y\}) \cap \mathcal{S}),$$

which penalizes outcomes that add little new target coverage. Using the induced local symbol f_x from Lemma 4.1, the PraC acquisition becomes

$$\alpha_t(x) = \beta I(f_x; y \mid \mathcal{D}_t) + \mathbb{E}_{q(y|x,\mathcal{D}_t)} \left[\mathrm{Vol}(\mathbb{C}_\delta(Y \cup \{y\}) \cap \mathcal{S})\right].$$

Tasks. We study failure discovery for a YOLO-based perception module in autonomous-driving scenarios. The input is a 3-dimensional scenario parameterization, and the outcome is a 2-dimensional failure metric describing two collision-relevant failure modes. We consider three nested target sets, $\mathcal{S}_1 \supset \mathcal{S}_2 \supset \mathcal{S}_3$, where smaller sets correspond to rarer and more difficult failure regions. Details are provided in Appendix C.3.

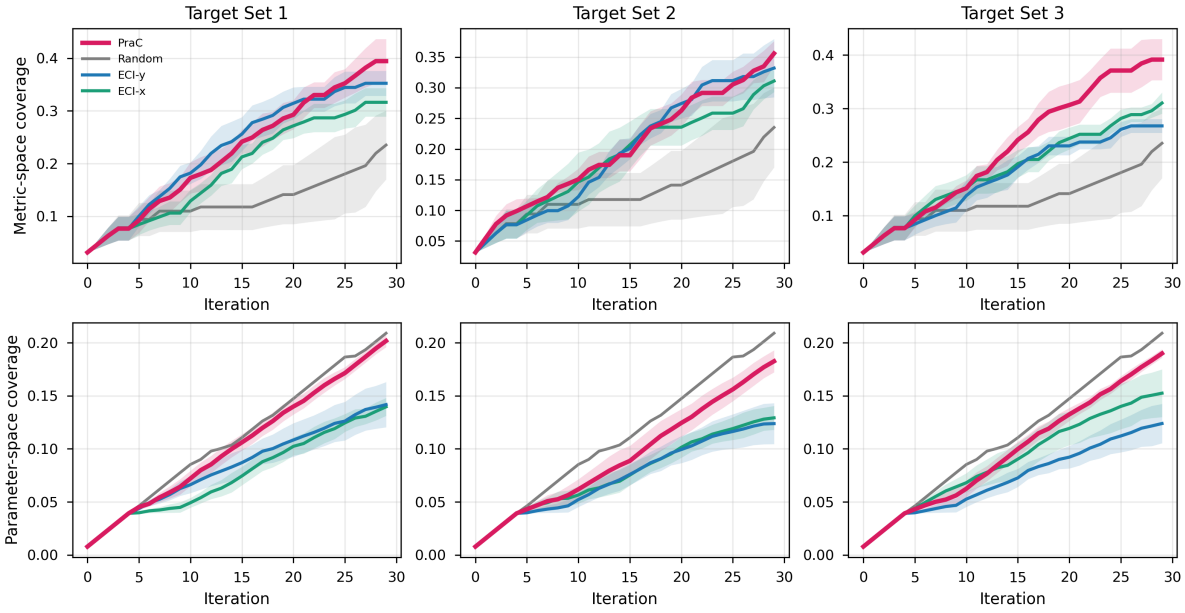

Figure 2: Performance evaluation for targeted active search on failure discovery in autonomous driving scenarios. Top row: coverage in the outcome/metric space. Bottom row: coverage in the input/parameter space. Shaded regions represent ±1 standard deviation over 4 seeds.

Baselines and metrics. We compare PraC against random sampling and two expected coverage improvement (ECI) variants adapted from Malkomes et al. (2021): ECI-y, which greedily prioritizes outcome-space coverage, and ECI-x, which prioritizes input-space coverage. We evaluate both metric-space coverage $\mathbb{C}_\delta(Y)$ and parameter-space coverage $\mathbb{C}_\delta(X)$.

Results. Fig. 2 shows that PraC balances the two coverage pressures: it neither spreads samples broadly in parameter space alone nor myopically chases already discovered target regions in outcome space. The information term learns the local surrogate structure, while the coverage potential directs sampling toward still-uncovered target regions. This is most useful for smaller target sets, where informative failures are sparse. PraC achieves the largest normalized target-region coverage in the most challenging target set, with an improvement of nearly 10 percentage points over the strongest baseline. Those results demonstrate how induced local symbols allow the same acquisition principle to adapt from finite-hypothesis identification to non-parametric target search when the model structure itself is not known *a priori*.

5.3 Hierarchical Symbols with Unknown Goals

Finally, we consider the regime in which the goal itself is not fully specified. The agent must learn both how actions produce outcomes and how those outcomes should be valued. This setting emphasizes the role of *hierarchical regret*: the pragmatic potential is not given directly, but inferred through an additional regret model. We use fixed curiosity weights to isolate the effect of this hierarchical design.

A representative problem is **composite Bayesian optimization** with unknown preferences. An action x produces a multi-objective outcome $y = f(x)$, while an unknown regret model $g(y)$ determines a scalar evaluation over outcomes. Following Lin et al. (2022), we consider learning g from pairwise preference queries. For a pair $\tilde{y} = (y_1, y_2)$, let $l(\tilde{y}) \in \{1, 2\}$ denote the preferred outcome. As in Chu & Ghahramani (2005), we use the probit likelihood

$$\Pr(l(\tilde{y}) = 1 \mid g(y_1), g(y_2)) = \Phi\left(\frac{g(y_2) - g(y_1)}{\sqrt{2}\lambda}\right),$$

where lower $g(y)$ indicates lower regret and hence higher preference.

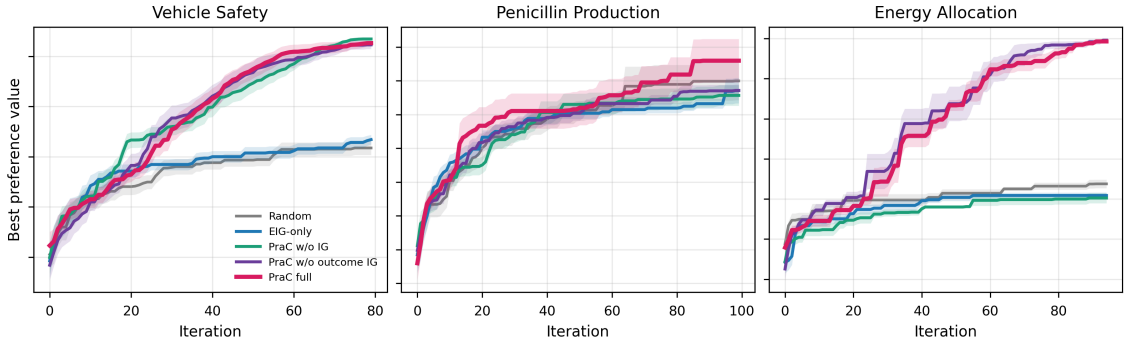

Figure 3: Performance evaluation for composite BO with unknown preferences. Metric is the best preference value attained among all collected outcomes. Shaded regions represent ±1 standard deviation over 20 seeds.

At each iteration, the method selects a candidate pair $\tilde{x} = (x_1, x_2)$, observes outcomes $\tilde{y} = (y_1, y_2)$, and receives a preference label l. The hierarchical PraC acquisition is thus

$$\alpha_t(\tilde{x}) = \beta_t I(f_{\mathcal{X}}; (\tilde{x}, \tilde{y}) \mid \mathcal{D}_t) + \mathbb{E}_{q(\tilde{y}|\tilde{x}, \mathcal{D}_t)} \left[\gamma_t I(g_{\mathcal{Y}}; (\tilde{y}, l) \mid \mathcal{D}_t) - \min_{y \in \{y_1, y_2\}} \mathbb{E}_{q(g(y)|y, \mathcal{D}_t)}[g(y)] \right],$$

where the first term learns the outcome model f, the second learns the regret model g, and the final term exploits the currently predicted lowest-regret outcome.

Tasks. We evaluate three real-world multi-objective optimization problems of increasing complexity: *vehicle safety* with 5-dimensional inputs and 3-dimensional outcomes, *penicillin production* with 7-dimensional inputs and 3-dimensional outcomes, and *energy allocation* in power grids with 40-dimensional inputs and 4-dimensional outcomes. Details are provided in Appendix C.4.

Baselines and metrics. We compare full hierarchical PraC against ablations that remove the outer epistemic term over f, the inner epistemic term over g, or both epistemic terms. These variants isolate the contributions of outcome-model learning, regret-model learning, and their joint interaction. We report the best true preference (negative regret) value attained among all collected outcomes. Additional comparisons with BOPE-style stage-wise variants (Lin et al., 2022) are provided in Appendix C.4.3.

Results. Fig. 3 shows that PraC consistently outperforms the ablated variants in learning and exploiting the unknown regret structure. As tasks become more complex, each component of the hierarchical acquisition becomes more important. In particular, the energy-allocation task shows that ignoring either outcome-model exploration or regret structure can prevent reliable discovery of high-preference regions, whereas the full PraC acquisition continues to improve. Additional comparisons with BOPE-style stage-wise variants in Appendix C.4.3 shows that stage-wise methods are sensitive to their switching schedules, while PraC more reliably discovers high-preference regions. These results support the central claim of hierarchical PraC: the agent can jointly learn what the world does and what matters, without requiring a manually staged separation between learning and optimization.

6 Conclusion

This paper introduced **Pragmatic Curiosity (PraC)**, a unified framework for hybrid learning and optimization. The framework exposes three operational design choices: *symbols*, the latent quantities whose clarification can change action; *regrets*, the potential landscapes that encode task value or shortfall; and *curiosity*, the exchange rate between information and pragmatic value. It offers a mechanism for refining knowledge and action together when grounded in task-relevant symbols and regulated by regret-based potentials. Across decision-oriented monitoring, targeted active search, and composite Bayesian optimization with unknown preferences, PraC reduced downstream decision risk, improved coverage of critical outcome regions, and jointly learned predictive and regret structures without relying on task-specific staging rules.

References

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Advances in neural information processing systems*, 33:21524–21538, 10 2020.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Wei Chu and Zoubin Ghahramani. Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pp. 137–144, 2005.

RP Cardoso Coelho, A Francisca Carvalho Alves, TM Nogueira Pires, and FM Andrade Pires. A composite bayesian optimisation framework for material and structural design. *Computer Methods in Applied Mechanics and Engineering*, 434:117516, 2025.

Francesco Di Fiore, Michela Nardelli, and Laura Mainini. Active learning and bayesian optimization: A unified perspective to learn with a goal. *arXiv preprint arXiv:2303.01560*, 2023.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16, 2017.

Tommaso Dreossi, Daniel J Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A Seshia. Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *International Conference on Computer Aided Verification*, pp. 432–442, 2019.

Daniel J Fremont, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L Sangiovanni-Vincentelli, and Sanjit A Seshia. Scenic: a language for scenario specification and scene generation. In *Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation*, pp. 63–78, 2019.

Karl Friston. The free-energy principle: A unified brain theory?, 2 2010. ISSN 1471003X.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active Inference: A Process Theory. *Neural Computation*, 29(1):1–49, 1 2017. ISSN 0899-7667. doi: 10.1162/NECO{_}a{_}00912.

Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. *Advances in neural information processing systems*, 31, 9 2018.

Leonardo D González and Victor M Zavala. Implementation of a bayesian optimization framework for interconnected systems. *Industrial & Engineering Chemistry Research*, 64(4):2168–2182, 2025.

Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

José Miguel Hernández-Lobato, W. Matthew Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, 2014.

Carl Hvarfner, Frank Hutter, and Luigi Nardi. Joint Entropy Search for Maximally-Informed Bayesian Optimization. 6 2022. URL `http://arxiv.org/abs/2206.04771`.

Carl Hvarfner, Erik Hellsten, Frank Hutter, and Luigi Nardi. Self-Correcting Bayesian Optimization through Bayesian Active Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2 2023. URL `http://arxiv.org/abs/2304.11005`.

Carl Hvarfner, David Eriksson, Eytan Bakshy, and Max Balandat. Informed initialization for bayesian optimization and active learning. *arXiv preprint arXiv:2510.23681*, 2025.

Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A Review of Yolo algorithm developments. *Procedia computer science*, 199:1066–1073, 2022.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998. ISSN 09255001. doi: 10.1023/A:1008306431147.

Mina Konakovic Lukovic, Yunsheng Tian, and Wojciech Matusik. Diversity-guided multi-objective bayesian optimization with batch evaluations. In *Advances in Neural Information Processing Systems*, pp. 17708–17720, 2020.

Qiaohao Liang and Lipeng Lai. Scalable bayesian optimization accelerates process optimization of penicillin production. In *NeurIPS 2021 AI for Science Workshop*, 2021.

Zhiyuan Jerry Lin, Raul Astudillo, Peter I. Frazier, and Eytan Bakshy. Preference Exploration for Efficient Bayesian Optimization with Multiple Outcomes. In *International Conference on Artificial Intelligence and Statistics*, pp. 4235–4258, 3 2022. URL `http://arxiv.org/abs/2203.11382`.

Gustavo Malkomes, Bolong Cheng, Eric Hans Lee, and Michael Mccourt. Beyond the Pareto Efficient Frontier: Constraint Active Search for Multiobjective Experimental Design. Technical report, 2021.

J. Močkus. On Bayesian Methods for Seeking the Extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975. doi: 10.1007/978-3-662-38527-2{_}55.

Willie Neiswanger, Ke Alexander Wang, and Stefano Ermon. Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information. In *International Conference on Machine Learning*, pp. 8005–8015, 2021.

Anjali Parashar, Joseph Zhang, Yingke Li, and Chuchu Fan. Cost-aware discovery of contextual failures using bayesian active learning. In *Conference on Robot Learning*, pp. 2239–2267. PMLR, 2025.

Manikandasriram Srinivasan Ramanagopal, Cyrus Anderson, Ram Vasudevan, and Matthew Johnson-Roberson. Failing to Learn: Autonomously Identifying Perception Failures for Self-Driving Cars. *IEEE Robotics and Automation Letters*, 3(4):3860–3867, 10 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2857402.

Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 4 2018.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. *Operations Research*, 66(1):230–252, 2 2018. ISSN 0030-364X. doi: 10.1287/opre.2017.1663. URL `https://pubsonline.informs.org/doi/10.1287/opre.2017.1663`.

Mark J. Schervish. *Theory of Statistics*. Springer New York, New York, NY, 1995. ISBN 978-1-4612-8708-7. doi: 10.1007/978-1-4612-4250-5.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-Oriented Bayesian Active Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. 12 2009. doi: 10.1109/TIT.2011.2182033. URL `http://arxiv.org/abs/0912.3995http://dx.doi.org/10.1109/TIT.2011.2182033`.

Ryoji Tanabe and Hisao Ishibuchi. An Easy-to-use Real-world Multi-objective Optimization Problem Suite. 9 2020. doi: 10.1016/j.asoc.2020.106078.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, pp. 567–574, 2009.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A Generative Image-to-text Transformer for Vision and Language. 5 2022.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pp. 3627–3635, 2017.

James T. Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for Bayesian optimization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 12 2018. URL `http://arxiv.org/abs/1805.10196`.

A Proof of Lemma 4.1 (Representation Compression)

To prove Lemma 4.1, we first introduce a lemma:

Lemma A.1. *For any (finite or infinite) set $\mathcal{X}$, after a few new measurements $(\mathbf{X}, \mathbf{Y}), \mathbf{X} \subseteq \mathcal{X}$ the KL divergence between $p(f_{\mathcal{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))$ and $p(f_{\mathcal{X}}|\mathcal{D})$ is*

$$D_{KL}[p(f_{\mathcal{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))\|p(f_{\mathcal{X}}|\mathcal{D})] = \mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} \log \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}],$$

where $L(\mathbf{Y}|f_{\mathbf{X}})$ is the likelihood of observations, $L(\mathbf{Y}) = \int L(\mathbf{Y}|f_{\mathbf{X}})p(f_{\mathbf{X}}|\mathcal{D})df_{\mathbf{X}}$ the marginal likelihood.

Proof. We first consider the case where $\mathcal{X}$ is *finite* to provide an intuitive insight. The KL divergence between $p(f_{\mathcal{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))$ and $p(f_{\mathcal{X}}|\mathcal{D})$ can be given as

$$\begin{aligned}
&D_{\mathrm{KL}}[p(f_{\mathcal{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))\|p(f_{\mathcal{X}}|\mathcal{D})] \\
=&D_{\mathrm{KL}}[p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))\|p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D})] \\
=&\int p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y})) \log \frac{p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))}{p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D})} df_{\mathcal{X}\setminus\mathbf{X}} df_{\mathbf{X}} \\
=&\int p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y})) \log \frac{p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D})L(\mathbf{Y}|f_{\mathbf{X}})}{p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D})L(\mathbf{Y})} df_{\mathcal{X}\setminus\mathbf{X}} df_{\mathbf{X}} \\
=&\int p(f_{\mathcal{X}\setminus\mathbf{X}}, f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y})) \log \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} df_{\mathcal{X}\setminus\mathbf{X}} df_{\mathbf{X}} \\
=&\int p(f_{\mathbf{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y})) \log \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} df_{\mathbf{X}} \\
=&\int \frac{p(f_{\mathbf{X}}|\mathcal{D})L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} \log \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} df_{\mathbf{X}} \\
=&\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})} \log \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}],
\end{aligned} \tag{4}$$

where $L(\mathbf{Y}|f_{\mathbf{X}})$ is the likelihood of observations, $L(\mathbf{Y}) = \int L(\mathbf{Y}|f_{\mathbf{X}})p(f_{\mathbf{X}}|\mathcal{D})df_{\mathbf{X}}$ the marginal likelihood.

Now we move to the more general cases where $\mathcal{X}$ is *infinite*. In such cases, there is no useful infinite-dimensional Lebesgue measure with respect to an "infinite-dimensional vector" $f_{\mathcal{X}}$. Thus, we need to resort to a more general definition for KL divergence based on the Radon-Nikodym derivative:

Definition A.2. If P and Q are probability measures over a set $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the KL divergence from P to Q is defined as

$$D_{\mathrm{KL}}[P\|Q] = \int_{\mathcal{X}} \log(\frac{dP}{dQ})dP,$$

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of P with respect to Q, and provided the expression on the right-hand side exists.

According to the measure-theoretic definition of Bayes' theorem for a dominated model (Schervish, 1995), the Radon-Nikodym derivative of the posterior $P(\cdot) := p(\cdot|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))$ with respect to the prior $\hat{P}(\cdot) := p(\cdot|\mathcal{D})$ is given as

$$\frac{dP}{d\hat{P}}(f_{\mathcal{X}}) = \frac{L(\mathbf{Y}|f_{\mathcal{X}})}{L(\mathbf{Y})}.$$

Since the dataset $\mathbf{Y}$ is finite, so similar to previous, we restrict the likelihood to only depend on the finite dataset:

$$\frac{dP}{d\hat{P}}(f_{\mathcal{X}}) = \frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}.$$

Now the KL divergence between $\hat{P}$ and P is quantified as

$$\begin{aligned}
&D_{\mathrm{KL}}[P(f_{\mathcal{X}})\|\hat{P}(f_{\mathcal{X}})] \\
=&\int_{f_{\mathcal{X}}} \log(\frac{dP}{d\hat{P}}(f_{\mathcal{X}}))dP(f_{\mathcal{X}}) \\
=&\int_{f_{\mathcal{X}}} \log(\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})})dP(f_{\mathcal{X}}) \\
=&\int_{f_{\mathcal{X}}} \log(\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})})\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}d\hat{P}(f_{\mathcal{X}}) \\
=&\int_{f_{\mathbf{X}}} \log(\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})})\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}d\hat{P}(f_{\mathbf{X}}) \\
=&\mathbb{E}_{\hat{P}(f_{\mathbf{X}})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}],
\end{aligned} \tag{5}$$

which has the exact same form as equation 4.

Therefore, we can conclude that regardless of the set $\mathcal{X}$ being finite or infinite, the KL divergence between the prior and posterior only depends on the evaluations of the observed data. That is to say, whilst we are in fact quantifying the KL divergence between the full distributions, we only need to keep track of the distribution over finite function values $f_{\mathbf{X}}$. □

Now we are ready to prove Lemma 4.1.

Proof. According to Lemma A.1,

$$\begin{aligned}
I(f_{\mathcal{X}}; (\mathbf{X}, \mathbf{Y})|\mathcal{D}) =&\mathbb{E}_{p(\mathbf{Y}|\mathbf{X},\mathcal{D})}[D_{\mathrm{KL}}[p(f_{\mathcal{X}}|\mathcal{D} \cup (\mathbf{X}, \mathbf{Y}))\|p(f_{\mathcal{X}}|\mathcal{D})]] \\
=&\mathbb{E}_{p(\mathbf{Y}|\mathbf{X},\mathcal{D})}\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}] \\
=&\mathbb{E}_{L(\mathbf{Y})}\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}] \\
=&\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}\mathbb{E}_{L(\mathbf{Y})}[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}] \\
=&\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}\int[\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}]L(\mathbf{Y})d\mathbf{Y} \\
=&\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}\int L(\mathbf{Y}|f_{\mathbf{X}})\log\frac{L(\mathbf{Y}|f_{\mathbf{X}})}{L(\mathbf{Y})}d\mathbf{Y} \\
=&\mathbb{E}_{p(f_{\mathbf{X}}|\mathcal{D})}[D_{\mathrm{KL}}[L(\mathbf{Y}|f_{\mathbf{X}})\|L(\mathbf{Y})]] \\
=&I(f_{\mathbf{X}}; \mathbf{Y}|\mathcal{D}).
\end{aligned} \tag{6}$$

□

B Re-interpretation of GP-UCB

The reinterpretations of most of the acquisition strategies in Table 1 are straightforward according to their definitions. However, placing a rather intuitive GP-UCB strategy within this framework seems implicit. To reveal their connection, we resort to Lemma 4.1. Then if we assume constant Gaussian noises $\mathcal{N} \sim (0, \sigma^2)$ for the observations, we have

$$\begin{aligned} I(f_x; y|\mathcal{D}_t) &= h(y \mid \mathcal{D}_t) - H(y|f_x, \mathcal{D}_t) \\ &= \frac{1}{2}\log(1 + \sigma^{-2}\sigma_t^2(x)), \end{aligned}$$

where $\sigma_t^2(x)$ is the variance evaluated on the GP model $p(f_x|\mathcal{D}_t)$.

When further assuming that the GP kernel $\kappa_t(x, x') \leq 1, \forall x, x' \in \mathcal{X}$, then $0 \leq \sigma_t^2(x) \leq \kappa_t(x, x') \leq 1$, which gives

$$\log(1 + \sigma^{-2}\sigma_t^2(x)) \geq \log(1 + \sigma^{-2})\sigma_t^2(x).$$

If we choose $\beta = \frac{1}{2}\log(1 + \sigma^{-2})$, then the epistemic term in PraC, *i.e.*, $I(f_x; y|\mathcal{D}_t)$, provides an upper bound of the square of the exploration term $\beta^{1/2}\sigma_t(x)$ in GP-UCB.

This reveals the close relationship between GP-UCB and PraC, showing that the uncertainty bonus in GP-UCB can be viewed as a first-order surrogate for an information-theoretic epistemic term.

C Experimental Details

This appendix provides a comprehensive overview of the simulation environment, model parameters, and hyper-parameter choices used to generate the results in this paper. The experiments were designed to be reproducible given the configurations outlined below.

C.1 Simulation Environment

All simulations for the perception failure evaluation in CARLA (Section C.3) were conducted on a Linux workstation with Ubuntu 22.04 LTS equipped with an Intel 13th Gen Core i7-13700KF CPU (16 cores, 24 threads, up to 5.4 GHz) and an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). The system ran CARLA simulations using CUDA 12.2 and NVIDIA driver version 535.230.02.

All other experiments were run on a MacBook Pro equipped with an Apple M2 Pro processor (10-core CPU, 16-core GPU) and a 3024 × 1964 Retina display. The GPU supports Metal 3, and the system was used as-is without external accelerators.

All experiments were conducted using Python 3.9. The core scientific computing libraries utilized were:

- BoTorch (Balandat et al., 2020)
- GPyTorch (Gardner et al., 2018)

In all experiments, we utilized the built-in Monte Carlo sampler in Botorch for the optimization of acquisition functions. The Monte Carlo samples are drawn from the posteriors for each model to approximate the expectations of acquisition functions, and default optimization method in Botorch are used to optimize the acquisition functions.

C.2 Decision-Oriented Plume Monitoring

This appendix provides implementation details and additional analysis for the decision-oriented plume-monitoring experiments in Section 5.1. All three tasks share the same sequential decision structure. The agent maintains a posterior distribution over latent environmental hypotheses θ, selects one sensor query x at each iteration, observes a plume count y, updates the posterior, and evaluates the downstream decision induced by the updated belief.

C.2.1 Plume Field Model

We consider the monitoring of a chemical plume field within a 100×100 domain, where multiple plume sources generate plume particles that can be measured by sensors. The field function is represented by the rate of hits, defined as the average number of particles per unit time measured by the sensor at a certain location.

The rate of hits for a chemical plume source is given as:

$$R_\theta(x) = \frac{R_s}{\log \frac{\gamma}{a}} \exp(-\frac{\langle \theta - x, V \rangle}{2D}) K_0(\frac{||\theta - x||_2}{\gamma}),$$

where θ is the location of the plume source, R_s is the rate at which the plume source releases the plume particles in the environment, $\gamma = \sqrt{D\tau/(1 + \frac{||V||^2\tau}{4D})}$ is the average distance traveled by a plume particle in its lifetime, a is the size of the sensor detecting plume particles, V is the average wind velocity, D is the diffusivity of the plume particles, and K_0 is the Bessel function of zeroth order.

The measurement y, *i.e.*, the number of particles measured, is modeled as a Poisson random variable with $R_\theta(x)\Delta t$ as the rate parameter, which leads to a likelihood model as

$$L_\theta(y|x) = \frac{\exp(-R_\theta(x)\Delta t)(R_\theta(x)\Delta t)^y}{y!},$$

where Δt is the time taken to obtain a measurement.

C.2.2 Task 1: Source Response Localization

The first task asks the agent to localize a plume source well enough to dispatch a response team near it. The true source is located at $(35, 65)$. The latent hypotheses θ consist of source-location hypotheses on a grid over the 100×100 domain with spacing 5, together with source-strength related multipliers

$$[0.4, 0.7, 1.0, 1.6].$$

The response actions form a coarser grid with spacing 20. The prior is intentionally multimodal, with modes near $(75, 25)$ and $(35, 65)$, so the agent must resolve a decision-relevant ambiguity rather than simply exploit a unimodal belief.

For a response action a and hypothesis θ, let ℓ_θ denote the source location specified by θ. The downstream loss is the normalized squared response distance,

$$L_{\text{loc}}(a, \theta) = \frac{\|a - \ell_\theta\|_2^2}{100^2 + 100^2}.$$

Given posterior $q_t(\theta) = q(\theta \mid \mathcal{D}_t)$, the Bayes risk is

$$BR(q_t) = \min_a \sum_\theta q_t(\theta) L_{\text{loc}}(a, \theta),$$

and the Bayes action is

$$a_t^\star = \arg\min_a \sum_\theta q_t(\theta) L_{\text{loc}}(a, \theta).$$

For dynamic curiosity schedule, the default dynamic settings are

$$\rho = 0.75, \quad k_\beta = 1, \quad \varepsilon_I = 10^{-8}, \quad \varepsilon_U = 10^{-8}, \quad \beta_{\min} = 10^{-3}, \quad \beta_{\max} = 10, \quad U^\star = 0.$$

We report:

- **Bayes risk**: $BR(q_t)$.

- **Response loss**: realized loss of the Bayes action under the true source,

$$L_{\mathrm{loc}}(a_t^\star, \theta_{\mathrm{true}}).$$

- **Response success**: whether the selected response action lies within one response-grid spacing of the true source,

$$\mathbf{1}\{\|a_t^\star - \ell_{\theta_{\mathrm{true}}}\|_2 \le 20\}.$$

- **Parameter error**: Euclidean error between the posterior MAP source location and the true source location.

C.2.3 Task 2: Consequence-Weighted Dispatch

The second task uses the same response-decision structure, but errors in high-consequence regions are more costly. The true source is located at $(68, 72)$. The prior has modes near $(34, 66)$, $(68, 72)$, and $(76, 24)$, and the high-consequence region is centered near the true source.

Let $C(\ell_\theta)$ denote the consequence weight associated with the source location under hypothesis θ. In the implementation, this weight is defined by a baseline term plus Gaussian consequence regions,

$$C(\ell_\theta) = C_0 + \sum_j w_j \exp\left(-\frac{\|\ell_\theta - c_j\|_2^2}{2\sigma_j^2}\right),$$

with an additional source-strength multiplier. The consequence-weighted downstream loss is

$$L_{\mathrm{cw}}(a, \theta) = C(\ell_\theta)\frac{\|a - \ell_\theta\|_2^2}{100^2 + 100^2}.$$

Given posterior $q_t(\theta)$, the Bayes risk is

$$BR(q_t) = \min_a \sum_\theta q_t(\theta) L_{\mathrm{cw}}(a, \theta).$$

For dynamic curiosity schedule, the default dynamic settings are

$$\rho = 0.75, \quad k_\beta = 1, \quad \varepsilon_I = 10^{-8}, \quad \varepsilon_U = 10^{-8}, \quad \beta_{\min} = 10^{-3}, \quad \beta_{\max} = 10, \quad U^\star = 0.$$

We report:

- **Bayes risk**: posterior expected consequence-weighted loss of the Bayes action.
- **Response loss**: realized consequence-weighted response loss under the true source.
- **Response success**: the same geometric success criterion as in source response localization.
- **Parameter error**: Euclidean MAP source-location error.

C.2.4 Task 3: Active Source Prioritization

The third task models limited repair or inspection resources. There are six possible sources, and the true active sources are indices $[4, 5]$. The repair budget is $k = 2$. The source consequence weights are

$$[70, 65, 5, 5, 180, 170],$$

and the prior activity probabilities are

$$[0.82, 0.80, 0.40, 0.40, 0.15, 0.15].$$

This creates a decision tension: the prior favors high-probability decoys, while the true active sources have high consequence weights but low prior activity.

The posterior is over nonempty active-source subsets θ. Let

$$z_i(\theta) \in \{0, 1\}$$

indicate whether source i is active under hypothesis θ. The posterior marginal activity probability of source i is

$$q_{t,i} = \sum_{\theta} q_t(\theta) z_i(\theta).$$

With zero false-positive cost, the expected missed-source risk before selecting repairs is

$$\sum_i w_i q_{t,i}.$$

If the selected repair set is S, its posterior expected benefit is

$$\sum_{i \in S} w_i q_{t,i}.$$

The Bayes action selects the top k sources by $w_i q_{t,i}$, and the Bayes risk is

$$BR(q_t) = \sum_i w_i q_{t,i} - \max_{|S|=k} \sum_{i \in S} w_i q_{t,i}.$$

For dynamic curiosity schedule, the default dynamic settings are

$$\rho = 0.9, \quad k_\beta = 2, \quad \varepsilon_I = 10^{-8}, \quad \varepsilon_U = 10^{-8}, \quad \beta_{\min} = 1, \quad \beta_{\max} = 10, \quad U^\star = 0.$$

We report:

- **Bayes risk**: posterior expected weighted risk left unrepaired after selecting k sources.
- **Missed-source risk**: realized total weight of true active sources not selected,

$$\sum_{i \in A_{\text{true}} \setminus S_t} w_i.$$

- **Top-k recall**:

$$\frac{|S_t \cap A_{\text{true}}|}{\min(k, |A_{\text{true}}|)}.$$

- **Weighted top-k recall**:

$$\frac{\sum_{i \in S_t \cap A_{\text{true}}} w_i}{\sum_{i \in A_{\text{true}}} w_i}.$$

- **Parameter error**: symmetric-difference size between the MAP active-source set and the true active-source set.

C.2.5 Baselines

We compare the following methods.

Random. Random selects a sensor query uniformly or according to the predefined random sampling protocol over the candidate set.

Pure information gain. EIG selects

$$x_t = \arg\max_{x \in \mathcal{X}_t^{\text{cand}}} \widehat{I}_t(x),$$

and therefore learns the latent environment without regard to downstream decision consequence.

Decision-greedy. Greedy selects

$$x_t = \arg\max_{x \in \mathcal{X}_t^{\text{cand}}} \Delta BR_t(x),$$

and therefore maximizes immediate expected downstream Bayes-risk reduction without explicitly valuing epistemic clarification.

Fixed-β PraC. Fixed PraC uses

$$\alpha_t(x) = \Delta BR_t(x) + 5\,\widehat{I}_t(x).$$

Dynamic-β PraC. Dynamic PraC uses the same acquisition but computes β_t adaptively from the feedforward–feedback schedule described in Section 4.3.

C.2.6 Analysis of Decision-Oriented Monitoring Results

The decision-oriented plume-monitoring experiments are designed to test whether PraC learns environmental distinctions that matter for downstream action, rather than merely reducing parameter uncertainty. The detailed results are reported in Table 2.

Across the three tasks, the results show three complementary behaviors: dynamic PraC improves downstream Bayes-risk reduction in the response-localization and source-prioritization tasks; decision-greedy can be competitive when posterior risk is already well aligned with the downstream loss; and pure information gain can fail when uncertainty reduction is not sufficiently targeted toward the decision-relevant ambiguity.

Source response localization. In source response localization, dynamic PraC achieves the lowest final Bayes risk, $0.0025 \pm 1.59 \times 10^{-12}$, and also obtains zero parameter error. All non-random adaptive methods achieve the best realized response loss and response success, indicating that the response decision itself is relatively easy once the posterior identifies a response-equivalent source region. However, the Bayes-risk and parameter-error results reveal a sharper distinction among methods. Dynamic PraC resolves the latent ambiguity most completely, whereas fixed PraC and EIG retain residual MAP source-location error, and Greedy retains larger parameter error. This suggests that, in this task, downstream decision improvement and epistemic clarification are aligned, but dynamic curiosity still improves the posterior quality beyond what is required for merely selecting a successful response action.

Consequence-weighted dispatch. The consequence-weighted dispatch task is more diagnostic because the posterior Bayes risk and the realized response loss no longer rank methods identically. Greedy achieves the lowest final Bayes risk, 0.0254 ± 0.00538, because it directly optimizes the one-step expected reduction in posterior risk. However, dynamic PraC achieves the best realized response loss, 0.0347 ± 0.00771, and the lowest parameter error, 7.62 ± 5.56. This gap is important: posterior Bayes risk measures expected loss under the agent's current belief, whereas realized response loss evaluates the downstream decision under the true latent hypothesis. A purely greedy policy can therefore reduce posterior risk quickly while still failing to collect information that would improve the realized decision under the true source. Dynamic PraC sometimes sacrifices immediate posterior-risk reduction to clarify the environmental hypothesis, which improves the eventual response decision in the realized environment.

Active source prioritization. Active source prioritization creates a different form of decision ambiguity. The prior assigns high activity probabilities to some lower-consequence decoy sources, while the true active sources have lower prior probability but much larger consequence weights. Dynamic PraC achieves the lowest final Bayes risk, 2.86 ± 2.72, improving over fixed PraC, Greedy, EIG, and Random. Dynamic and fixed PraC tie on missed-source risk, top-k recall, and weighted top-k recall, showing that both PraC variants

reliably identify the high-consequence sources needed for the repair decision. By contrast, EIG and Random perform substantially worse on missed-source risk and weighted recall. This supports the central point of the decision-oriented formulation: information is useful only when it helps resolve the ambiguity that affects the downstream repair set. Pure information gain may reduce uncertainty over the active-source subset, but it does not necessarily prioritize distinctions that change the top-k consequence-weighted decision.

Comparison with pure information gain. EIG is a strong baseline when learning the latent hypothesis is well aligned with the downstream decision. This explains why it performs competitively in source response localization and matches fixed PraC on several metrics. However, EIG lacks an explicit mechanism for distinguishing decision-relevant uncertainty from decision-irrelevant uncertainty. In consequence-weighted dispatch and active source prioritization, this limitation becomes more visible: EIG can collect information about the environment without adequately targeting the distinctions that reduce consequence-weighted response loss or missed-source risk. PraC addresses this by coupling information gain with expected Bayes-risk reduction.

Comparison with decision-greedy. The decision-greedy baseline isolates the pragmatic term by choosing queries that maximize $\Delta BR_t(x)$ without any explicit epistemic value. Its strong Bayes-risk performance in consequence-weighted dispatch shows that myopic posterior-risk reduction can be effective when the current belief already points toward useful response decisions. However, Greedy is more vulnerable when the current posterior is brittle or biased by the prior. In active source prioritization, Greedy is substantially worse than dynamic PraC in final Bayes risk, even though it remains competitive on some top-k metrics. This indicates that myopic decision improvement may not be sufficient when the downstream action depends on resolving low-prior but high-consequence hypotheses.

Why Bayes risk and realized loss can disagree. The distinction between Bayes risk and realized downstream loss is essential for interpreting these experiments. Bayes risk is the expected loss under the agent's posterior belief, while realized loss is evaluated under the true latent hypothesis. A method can reduce Bayes risk by becoming confident under its current posterior, even if the posterior remains biased in a way that hurts realized performance. This is why Greedy can obtain the lowest final Bayes risk in consequence-weighted dispatch while dynamic PraC obtains the best realized response loss and parameter error. PraC's epistemic term helps correct the posterior before committing too strongly to the current decision surrogate.

Overall, the three tasks show that PraC is most beneficial when downstream action depends on resolving specific decision-relevant ambiguity. When the pragmatic objective and epistemic objective are aligned, PraC remains competitive with EIG and Greedy. When they diverge, dynamic PraC provides a mechanism for deciding when information is worth acquiring because it improves downstream action. This supports the role of PraC as a hybrid learning-and-optimization framework: the agent does not learn the environment for its own sake, nor does it merely optimize the current posterior decision; it learns the environmental distinctions that make better downstream action possible.

C.2.7 Diagnostics of Dynamic Curiosity

In addition to the main performance curves, we report scheduler diagnostics to verify that dynamic curiosity behaves as intended. The diagnostic plots include:

- β_t, the scheduled curiosity coefficient;
- β_t^{ff}, the feedforward exchange-rate scale;
- β_t^{fb}, the feedback activation factor;
- $\beta_t \widehat{I}_t(x_t)$, the selected effective epistemic pressure.

These diagnostics distinguish three regimes. When decision-relevant uncertainty is high, the feedback activation keeps curiosity active. When useful information is costly relative to downstream Bayes-risk

Table 2: Full final performance metrics for decision-oriented plume monitoring over 20 seeds. Lower is better for Bayes risk, response loss, parameter error, and missed-source risk. Higher is better for response success, top-k recall, and weighted top-k recall. Bold indicates the best or tied-best mean within each task and metric. Dyn. PraC denotes dynamic-β_t PraC; Fixed PraC denotes fixed-β PraC with $\beta = 5$.

Source Response Localization					
Metric	Dyn. PraC	Fixed PraC	Greedy	EIG	Random
Bayes risk ↓	$\mathbf{0.0025 \pm 1.59 \times 10^{-18}}$	0.00377 ± 0.00254	0.00463 ± 0.000765	0.00377 ± 0.00254	0.0124 ± 0.00495
Response loss ↓	$\mathbf{0.0025 \pm 0}$	$\mathbf{0.0025 \pm 0}$	$\mathbf{0.0025 \pm 0}$	$\mathbf{0.0025 \pm 0}$	0.0125 ± 0.00632
Response success ↑	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	0.80 ± 0.40
Parameter error ↓	$\mathbf{0.00 \pm 0}$	2.00 ± 4.00	5.24 ± 5.25	2.00 ± 4.00	8.61 ± 5.09

Consequence-Weighted Dispatch					
Metric	Dyn. PraC	Fixed PraC	Greedy	EIG	Random
Bayes risk ↓	0.0282 ± 0.00374	0.0398 ± 0.0125	$\mathbf{0.0254 \pm 0.00538}$	0.0398 ± 0.0125	0.0422 ± 0.0118
Response loss ↓	$\mathbf{0.0347 \pm 0.00771}$	0.0424 ± 0.00945	0.0424 ± 0.00945	0.0424 ± 0.00945	0.0463 ± 0.0144
Response success ↑	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$	$\mathbf{1.00 \pm 0}$
Parameter error ↓	$\mathbf{7.62 \pm 5.56}$	8.86 ± 6.60	8.04 ± 3.06	8.86 ± 6.60	17.8 ± 9.72

Active Source Prioritization					
Metric	Dyn. PraC	Fixed PraC	Greedy	EIG	Random
Bayes risk ↓	$\mathbf{2.86 \pm 2.72}$	3.59 ± 2.70	14.0 ± 10.6	23.3 ± 19.1	34.5 ± 17.0
Missed-source risk ↓	$\mathbf{34.0 \pm 68.0}$	$\mathbf{34.0 \pm 68.0}$	36.0 ± 72.0	68.0 ± 83.3	316 ± 68.0
Top-k recall ↑	$\mathbf{0.90 \pm 0.20}$	$\mathbf{0.90 \pm 0.20}$	$\mathbf{0.90 \pm 0.20}$	0.80 ± 0.245	0.10 ± 0.20
Weighted top-k recall ↑	$\mathbf{0.903 \pm 0.194}$	$\mathbf{0.903 \pm 0.194}$	0.897 ± 0.206	0.806 ± 0.238	0.0971 ± 0.194

reduction, the feedforward scale increases. When the decision-symbol posterior concentrates, the feedback activation suppresses curiosity even if raw information-gain ratios become numerically large. This prevents late-stage ratio explosions and makes the dynamic schedule interpretable.

Role of dynamic curiosity. The dynamic schedule gives β_t an operational meaning as an exchange rate between epistemic value and downstream Bayes-risk reduction. The feedforward scale estimates how large curiosity must be for information gain to compete with pragmatic improvement, while the feedback activation suppresses curiosity as decision-relevant uncertainty decreases. This prevents β_t from acting as a fixed exploration knob. Instead, the relevant quantity is the effective epistemic pressure $\beta_t \widehat{I}_t(x)$. The empirical results are consistent with this interpretation: dynamic PraC improves posterior quality in source response localization, improves realized downstream loss in consequence-weighted dispatch, and achieves the lowest Bayes risk in active source prioritization.

C.3 Targeted Active Search

We consider the failure discovery for YOLO-based object detection (Jiang et al., 2022; Redmon & Farhadi, 2018) in the CARLA simulator (Dosovitskiy et al., 2017).

C.3.1 Perception in self driving simulation CARLA

This requires the generation of various scenarios in the environment using CARLA simulator. The environment is a composition of a static context and scenario ϕ. We use the probabilistic programming framework Scenic (Fremont et al., 2019) for sampling scenarios with varying contextual information for a fixed scenario variable ϕ. For the simulations presented in this paper, we used a publicly available, pre-existing environment (Dreossi et al., 2019), which consists of the ego vehicle maneuvering on the road with two non-ego agents– two non-ego cars and a pedestrian crossing the road. We use YOLO object detection model to detect all non-ego agents in a scene. The generated scenario is seeded for reproducibility, so that for a given scenario ϕ, the environment can be treated as a deterministic quantity. Each scenario is defined using $\phi = [b_e, b_l, s]$, where $b_e, b_l \in [5, 15]$ represent the braking threshold of the ego car and lead car (non-ego car in-front of the

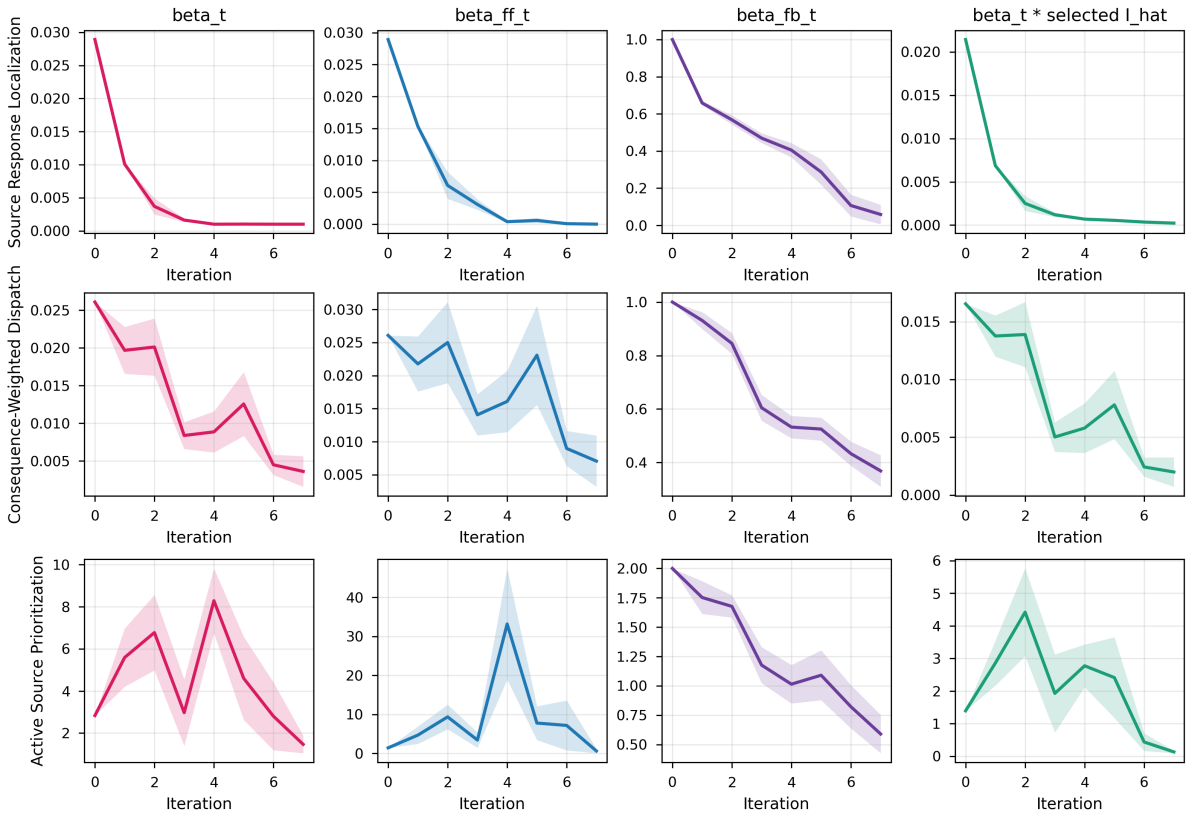

Figure 4: Diagnostics of dynamic curiosity scheduler. Shaded regions show mean ±1 standard error over 20 seeds.

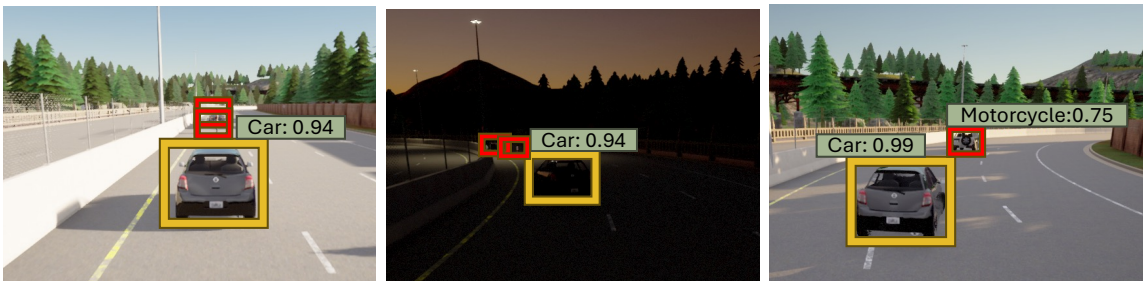

Figure 5: Examples of missed object detection by YOLO due to two reasons considered in perception failure case study in Section C.3. Fig (left to right): example of Failure-1 (distance), Failure-2 (poor light) and Failure-1 and Failure-2 both in one scene (distance and poor light), respectively. Bounding boxes for detected objects (misdetections) shown in yellow (red) with detection confidence numbers. Each scene has two cars and a pedestrian.

ego car) (m) and $s \in [0, \pi/2]$ denotes the sun altitude angle (rad). Each of these quantities is normalized to be within $[0, 1]$, and the normalized scenario is chosen as the decision variable x for active inference.

We are interested in environment variables (scenarios) that lead to two specific types of failures– failure to detect non-ego agents due to large distance from ego vehicle (Failure 1), and failure to detect non-ego agents due to poor scene lighting (Failure 2). Fig. 5 shows examples of the discussed object detection failures we aim to discover.

C.3.2 LLM-based evaluation for CARLA

We use LLM to perform evaluations for whether a generated scene corresponds to a failure due to specific type.

Each evaluation for a specific scenario corresponds to $T = 60$ steps of simulations. Images recorded from the camera view are used for object detection and classification at every 10 steps using YOLO-v3 (Redmon & Farhadi, 2018), and the classified image are used as inputs to GiT (Wang et al., 2022) to predict a likely failure type based on fine-tuning data.

Results obtained from YOLO, along with the reports from GiT are used as inputs to GPT3 model for failure evaluation, which is queried 6 times per evaluation, and assigns binary scores pertaining to each failure mode for each scene (camera image). The average value reported across 6 scenes is used to construct $c_i : \mathcal{Z} \to \mathbb{R}$ for $i = 1, 2$ as: $c_i(z) = \frac{1}{T}\sum_{t=1}^{T} b_t^i$. Here $b_t^i \in \{0, 1\}$ is a scene-specific binary evaluation provided by the LLM based on report generated by GiT to assess if an object detection failure is observed in a given scene and corresponds to Failure-i.

We use GPT-3.5 Turbo model for LLM-based binary evaluations, with each evaluation we query the LLM 6 times and combine the binary evaluations for all 6 runs. Note that usage of LLM is not a core part of our methodology and is used only as a subjective evaluator.

We show the prompt used for failure evaluation of each scene using the LLM in the box below. Information shown in red and blue is obtained from GiT captioning system, and CARLA respectively. The output of the LLM is used to obtain a binary number for each scene which is composed to give a scenario specific cost function c_1, c_2.

C.3.3 Task-Specific Configurations

- Goal: We consider two cost functions c_1, c_2 associated with each type of failure. The goal is to sample from the set $\Omega = \{z | c_1(z) \geq C_1, c_2(z) \geq C_2\}$, and we consider three target sets defined by
  $C_1 = C_2 = 0.1$,
  $C_1 = C_2 = 0.5$, and
  $C_1 = C_2 = 0.8$.

- Degree of curiosity: $\beta = 20.0$.

Prompt Used for CARLA evaluation

You will be provides the analysis of YOLO Object detection on an image that was taken from the camera feed of CARLA simulator. The simulator is simulating a pedestrian crossing the road before a car in front of the ego car. There are two cars and one pedestrian in each image.
The information provided:

1. *Objects detected*: List of objects detected by YOLO in the image. This list should have atleast one object from the *Objects to detect* list
2. *Objects to detect* list:
   - One object with one of the following labels: 'car','truck',
   - One object with one of the following labels: 'car', 'truck','bus','motorcycle','bicycle'
   - One object with one of the following labels: 'person'
3. Reason: The reason is a brief explanation of the failure to detect all objects, if that happens, and is generated by a pre-trained GiT model in the form of captions for the image.

We are looking to discover images where YOLO fails to detect an object due to **bad light** and/or **large distance.** If the list of *objects detected* has an object missing from the *objects to detect* list, look at the reason. The reason can have other components as well, but it can 'only' be considered as **bad light** if at least one of the objects was failed to detect strictly due to **bad light**. Similarly, the reason can have other components as well, but it can 'only' be considered as **large distance** if the reason contains the phrase 'far away'. Follow the response instructions while responding.
Response Instructions: Respond should be an integer 0, 1, 2, 3 or 4:

- 0 indicating that at least one object was missing from the 'objects to detect' list, but the reason provided does not correspond to bad light or large distance.
- 1 indicating that an object was not detected and the reason provided corresponds to bad light only.
- 2 indicating that an object was not detected, and the reason corresponds to large distance only.
- 3 indicating that an object was not detected, and the reason corresponds to both large distance and bad light.
- 4 indicating all objects are detected. Do not provide explanation.

Response format: Response: [integer], where integer = 0,1,2,3,4.
The list of objects detected and reason for incomplete detection for the image are as follows:

- Objects detected: {objects}
- Reason: {reason}

C.4 Composite Bayesian Optimization

C.4.1 Preference Evaluation

We simulate human-in-the-loop or policy-driven decision-making via pairwise preference queries. That is, for selected pairs of outcomes (y_1, y_2), a preference function indicates which design is preferred. These preferences are generated based on a latent utility function, not revealed to the optimizer.

An initial set of 1 pairwise preferences is randomly sampled to initialize the model. Each step of the optimization selects new pairs to query, guided by the used acquisition strategy.

C.4.2 Task-Specific Configurations

Vehicle Safety.

- Goal: Optimize vehicle crash-worthiness.
- Testbed: See Tanabe & Ishibuchi (2020) for details.
- Ground Truth: $g(y) = (y - y^*)^2$, where $y^* = [1864.7202, 11.8199, 0.2904]$.
- Degree of curiosity: $\beta = \gamma = 1.0$.

Penicillin.

- Goal: Maximize the penicillin yield while minimizing time to ferment and the CO2 byproduct.
- Testbed: See Liang & Lai (2021) for details.
- Ground Truth: $g(y) = (y - y^*)^2$, where $y^* = [25.935, 57.612, 935.5]$.
- Degree of curiosity: $\beta = \gamma = 1.0$.

Energy Resource Allocation.

- Goal: Identify deployment strategies for Distributed Energy Resources (DERs) in Optimal Power Flow (OPF) that align with implicit ethical preferences across multiple performance dimensions detailed in the following table.
- Testbed: IEEE 30-bus network in pandapower library.
- Ground Truth: $g(y) = a^\intercal y$, where $a = [-1, 1, -2, -1]$
- Degree of curiosity: $\beta = \gamma = 1.0$.

Performance Metrics	Definition
Voltage Fairness	Measures the variance in bus voltages across the network; lower variance implies more equitable voltage delivery.
Total Cost	Combines capital expenditures for DER installation and operational costs related to reactive power support.
Priority Area Coverage	Quantifies the share of power delivered to high-priority buses, such as rural or underserved regions.
Resilience	Assesses the percentage of time that all bus voltages remain within safe operating limits under perturbations (*e.g.*, load uncertainty or line outages).

C.4.3 Comparison with BOPE

To highlight the benefits of jointly learning and optimizing, rather than separating these into stages, we extend the baseline comparison with **BOPE** from Lin et al. (2022).

The original BOPE framework is intentionally flexible and leaves many problem-specific design choices open, especially regarding how and when to switch between preference exploration and experimentation. In our comparison, we consider four representative stage-wise variants:

- **BOPE-I**: A two-phase strategy that starts with qEUBO (preference-focused exploration) and switches to qNEI in the second half (objective-driven refinement), illustrating the effect of premature exploitation.

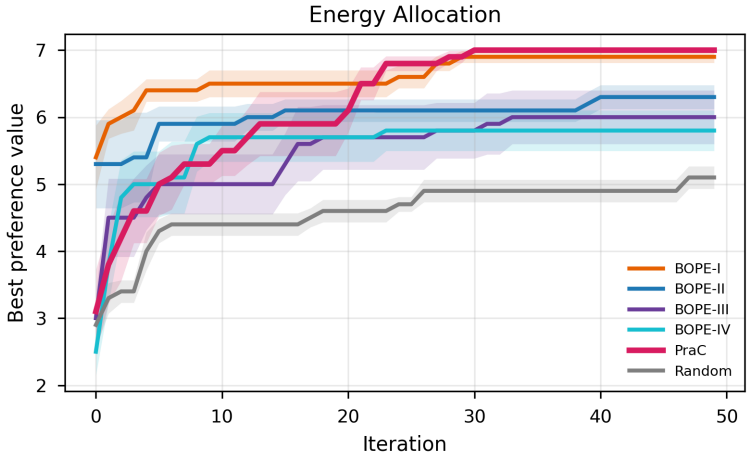

Figure 6: Comparison with BOPE-style stage-wise variants for energy resource allocation. Metric is the best preference value attained. Shaded regions represent ±1 standard deviation over 20 seeds.

- **BOPE-II**: A two-phase strategy that starts with qNEI (exploring the objective space) and switches to qEUBO in the second half (exploiting the learned preference model), using a frozen outcome snapshot for qEUBO.
- **BOPE-III**: A qEUBO-only variant where experiments are selected by qEUBO with a newly sampled objective realization at each iteration, encouraging stronger exploration through objective variation.
- **BOPE-IV**: A BOPE variant that uses standard preference exploration (qEUBO) and selects experiments exclusively with qNEI, fully refitting both outcome and preference GPs after each update.

Figure 6 reports their preference scores over random sampling (**RS**). It is evident that our method (**PraC**) consistently discovers higher-preference regions after a brief initial exploration phase, while BOPE variants are highly sensitive to their stage-wise design choices. **BOPE-I**, being preference-driven in the first phase, initially attains higher preference values but fails to balance exploration and exploitation, leading to poor convergence in the second half; its performance is also sensitive to the precise switching point. **BOPE-II** explores first and then optimizes, achieving better final performance than BOPE-I, but the strict separation between exploration and exploitation still yields suboptimal outcomes. **BOPE-III** mixes both aspects but remains more exploitation-centric, performing better than BOPE-I/II yet still below PraC. **BOPE-IV** and PraC share the idea of refitting both models at each step, but BOPE-IV converges to a lower-preference solution. In contrast, our acquisition strategy jointly leverages information from both the outcome and preference models at every iteration, leading to higher sample efficiency and more reliable discovery of high-preference regions.