EVIDENTIAL LATENT WORLD MODELS FOR SAFE MODEL-BASED REINFORCEMENT LEARNING

Alisson H. Kolling, Junior D. Jesus, Matheus G. Mateus, Rodrigo S. Guerra & Paulo L. J. Drews-Jr
Computational Science Center
Federal University of Rio Grande
Rio Grande, Brazil
alikolling@gmail.br

Victor A. Kich
Learning Machines Group
University of Kyoto
Kyoto, Japan
victorkich98@gmail.com

Ricardo B. Grando
Robotics and AI Lab
Technological University of Uruguay
Rivera, Uruguay
ricardo.bedin@utec.edu.uy

ABSTRACT

Uncertainty estimation is crucial for deploying reinforcement learning in safety-critical domains such as robotics and autonomous systems. This work introduces Model-based Uncertainty-Aware Reinforcement Learning (MUARL), a constrained model-based reinforcement learning framework that augments TD-MPC2 with evidential deep learning to estimate epistemic and aleatoric uncertainty in a single dynamics-model forward pass. MUARL feeds these estimates into a dual-constraint Model Predictive Path Integral planner that penalizes both predicted safety-cost violations and model uncertainty via adaptive Lagrangian multipliers, so safety is handled directly at planning time. In a dynamic unicycle-car navigation task, evidential uncertainty achieves much sharper out-of-distribution detection than normalizing flows and stochastic ensembles, leading to safer exploration around constraint regions. On Safety Gymnasium navigation benchmarks, MUARL variants obtain higher feasibility and lower cumulative constraint costs than model-free baselines and alternative model-based methods, while keeping task performance competitive. Taken together, these results show that evidential uncertainty can be used in real-time sampling-based planners with only modest extra cost, and that this yields a workable uncertainty-aware constrained MBRL scheme for safety-critical autonomous systems.

1 INTRODUCTION

Deploying reinforcement learning on robots and autonomous systems requires decision-making under uncertainty while respecting operational constraints. In safety-critical domains, the system must avoid constraint violations, such as collisions, hazardous regions, or unsafe operating regimes, throughout training and deployment, not just maximize average reward (Brunke et al., 2022).

Model-based reinforcement learning (MBRL) is appealing here because it uses a learned dynamics model to plan ahead, improving sample efficiency (Plaat et al., 2020). Recent latent-space MBRL methods with sampling-based planning produce good results with simple planning loops (Hansen et al., 2022; 2024). However, this creates a failure mode: if the model is wrong, the planner may confidently choose actions that look good in imagination but fail in reality. This risk is amplified in out-of-distribution states where model error is large and difficult to detect from prediction losses alone.

A natural response is to quantify uncertainty in the world model and use it during planning (Mucsányi et al., 2023). In practice, this is challenging: many uncertainty estimators such as deep ensembles are too expensive when embedded in a planner that evaluates thousands of imagined trajectories per decision step (Amini et al., 2020). For sampling-based MPC, uncertainty must

be both informative and cheap—it should distinguish familiar regions from novel ones without multiplying planning cost by large factors.

This paper proposes MUARL, an uncertainty-aware constrained MBRL framework that augments a latent dynamics world model with single-pass uncertainty estimation and integrates that signal into constrained planning. Our dynamics model uses evidential regression to produce aleatoric and epistemic uncertainty in one forward pass, suitable for high-throughput trajectory evaluation (Amini et al., 2020). We incorporate predicted costs and model uncertainty into an MPPI planner's scoring function via an adaptive Lagrangian formulation, trading off task performance against constraint satisfaction and uncertainty exposure at decision time (Boyd & Vandenberghe, 2004).

We evaluate MUARL in two settings: a controlled navigation task where we explicitly mark out-of-distribution regions, and Safety Gymnasium (Ji et al., 2023a) benchmarks with standard cost and feasibility metrics against model-free and model-based baselines. Our goal is to show that an efficient uncertainty-aware world model can give useful guidance to the planner when decisions are subject to safety constraints.

Contributions. This work makes three contributions. (1) We integrate deep evidential regression into a TD-MPC2-style (Hansen et al., 2024) latent dynamics model to obtain single-pass epistemic and aleatoric uncertainty that is practical for sampling-based MPC. (2) We formulate a dual-constraint MPPI objective that combines learned cost predictions with evidential uncertainty, and we instantiate three uncertainty integration strategies, taking a feasibility-based adaptive variant (MUARL m0f) as our primary algorithm. (3) Through controlled navigation experiments and Safety Gymnasium benchmarks, we show that evidential uncertainty provides a strong OOD/reliability signal and that combining costs with uncertainty at decision time substantially improves constraint feasibility over strong model-free constrained baselines, at the cost of more conservative returns.

2 Related Work

Prior work on safe model-based reinforcement learning can be organized along three axes: firstly, works that provide uncertainty-aware dynamics learning, secondly constrained planning with learned models, and finally model-free constraint satisfaction during policy learning. Most existing methods cover only one or two of these axes, and many rely on multi-model uncertainty estimates that are costly inside sampling-based planners.

PETS (Chua et al., 2018) uses probabilistic ensembles to represent epistemic uncertainty in learned dynamics and achieves sample efficiency, but it does not impose explicit safety constraints during planning. MBPO (Janner et al., 2019) mitigates model bias by limiting rollout horizons and mixing real with synthetic data, improving stability without introducing constraint-aware decision rules. MOReL (Kidambi et al., 2020) adds pessimism by penalizing uncertain transitions, yielding conservative behavior, but it is primarily designed for offline settings and typically inherits the computational cost of ensemble-style uncertainty.

CCE (Wen & Topcu, 2018) demonstrates constrained MPC-style planning by using learned models to predict costs and reject trajectories that violate thresholds, but it treats these predictions as reliable and does not explicitly account for model uncertainty. RCE (Liu et al., 2021) incorporates ensemble uncertainty into cross-entropy planning through robust optimization, improving constraint satisfaction under model error at the cost of maintaining and evaluating multiple models. CAPPETS (Ma et al., 2022) inflates predicted costs using uncertainty-dependent penalties to obtain conservative constraint satisfaction, but similarly depends on ensemble uncertainty and operationalizes uncertainty mainly through cost tightening rather than as a separate planning signal.

Lyapunov-based safe MBRL (Berkenkamp et al., 2017; Chow et al., 2018) can provide formal safety certificates, but often requires careful construction and can be conservative in complex, high-dimensional settings. Model-free constrained RL methods such as CPO (Achiam et al., 2017), SAC-Lag (Ray et al., 2019), SAC-PID (Stooke et al., 2020), and FOCOPS (Zhang et al., 2020) enforce expected cost constraints during learning, but they do not use a world model to evaluate and screen candidate action sequences at decision time and cannot directly reason about epistemic model uncertainty.

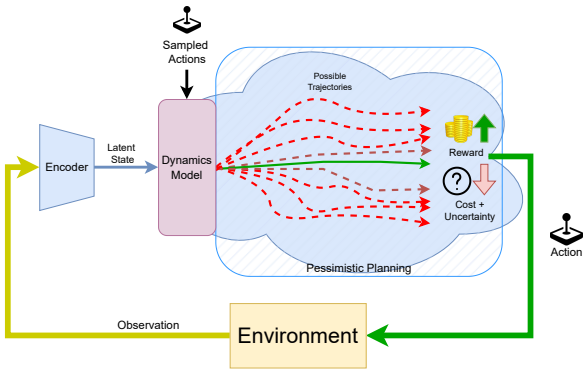


Figure 1: MUARL architecture during inference. The encoder compresses observations into latent states. The MPPI planner samples action sequences, evaluates them through the dynamics model with uncertainty estimation, scores trajectories using Lagrangian penalties on costs and uncertainty, and selects actions for execution.

Evidential learning provides single-pass uncertainty estimates by predicting distributional parameters rather than relying on ensembles or repeated stochastic passes (Sensoy et al., 2018; Amini et al., 2020), and it has been explored in robotics and RL settings (Itkina & Kochenderfer, 2023; Charpentier et al., 2022; Stutts et al., 2024). Recent latent safety filters and safe exploration methods also leverage epistemic uncertainty to avoid unreliable predictions (Seo et al., 2025; As et al., 2025), motivating uncertainty as an online decision-time signal.

Overall, existing work either uses expensive multi-model uncertainty, focuses solely on cost constraints or uncertainty penalties, or lacks decision-time constraint enforcement in latent MPC frameworks such as TD-MPC2. This leaves room for single-pass uncertainty estimators that are cheap enough for sampling-based planners, and for methods that combine cost and uncertainty as complementary signals in the same planning objective.

3 Methodology

We build on latent-space model-based RL with sampling-based MPC, extending a TD-MPC2-style implicit world model with (i) evidential uncertainty estimation in the dynamics model and (ii) planning-time constraint handling that jointly accounts for learned costs and model uncertainty.

3.1 World model with evidential dynamics

Given an observation o_t, an encoder maps it to a latent state $z_t \in \mathbb{R}^{d_z}$, and a learned dynamics model predicts the next latent state conditioned on action a_t:

$$z_t = f_{\text{enc}}(o_t), \qquad \hat{z}_{t+1} = f_{\text{dyn}}(z_t, a_t). \tag{1}$$

In addition, we learn predictors for reward, cost, and termination in latent space:

$$\hat{r}_t = f_r(z_t, a_t), \qquad \hat{c}_t = f_c(z_t, a_t), \qquad \hat{d}_t = f_{\text{term}}(z_t). \tag{2}$$

This follows the implicit-world-model design where planning is performed entirely in the latent space, avoiding reconstruction of high-dimensional observations.

For each latent dimension i, the dynamics head outputs evidential parameters $(\mu_i, \nu_i, \alpha_i, \beta_i)$ that define a Student-t predictive distribution

$$p(z_{t+1,i} \mid z_t, a_t) = \text{St}\Big(z_{t+1,i};\ \mu_i,\ \tfrac{\beta_i(1+\nu_i)}{\nu_i \alpha_i},\ 2\alpha_i\Big),$$

and we use μ_i as the mean next-state prediction. Full NIG priors and derivations are given in Appendix A.

We extract per-dimension aleatoric and epistemic terms as

$$u_i^{\text{al}} = \frac{\beta_i}{\alpha_i - 1}, \qquad u_i^{\text{ep}} = \frac{1}{\nu_i},$$

where u_i^{al} captures irreducible transition noise (distribution width) and u_i^{ep} captures model ignorance via low evidence (small ν_i), which grows in out-of-distribution regions.

MPPI requires a scalar score for each trajectory. We therefore aggregate per-dimension uncertainties into a single transition-level uncertainty. In our experiments we use a conservative aggregation across latent dimensions (treating a transition as uncertain if any latent component is uncertain):

$$u^{\text{al}}(z_t, a_t) = \max_i u_i^{\text{al}}, \qquad u^{\text{ep}}(z_t, a_t) = \max_i u_i^{\text{ep}}, \qquad u(z_t, a_t) = \sqrt{\left(u^{\text{al}}\right)^2 + \left(u^{\text{ep}}\right)^2}. \tag{3}$$

This yields a single uncertainty signal u that is cheap to compute and suitable for trajectory scoring.

We train the dynamics with a standard evidential regression loss (Amini et al., 2020) and supervised losses for reward/cost/termination; implementation details are in Appendix C.

3.2 Dual-constraint MPPI planning

At decision time, MPPI samples N candidate action sequences $A^{(j)} = (a_0^{(j)}, \ldots, a_{H-1}^{(j)})$ and rolls them out through the world model for horizon H:

$$z_{t,0}^{(j)} = z_t, \qquad z_{t,k+1}^{(j)} = f_{\text{dyn}}(z_{t,k}^{(j)}, a_k^{(j)}), \qquad k = 0, \ldots, H-1. \tag{4}$$

Along each rollout we accumulate predicted rewards, costs, and uncertainty terms using f_r, f_c, and the evidential uncertainty computed from f_{dyn}.

For each sampled trajectory j over a horizon H, we roll out the latent dynamics and accumulate discounted reward, cost, and normalized uncertainty:

$$V^{(j)} = \sum_{k=0}^{H-1} \gamma^k \hat{r}_{t,k}^{(j)} + \gamma^H \hat{Q}\left(z_{t,H}^{(j)}, a_{t,H}^{(j)}\right), \tag{5}$$

$$C^{(j)} = \sum_{k=0}^{H-1} \gamma^k \hat{c}_{t,k}^{(j)} + \gamma^H \hat{C}\left(z_{t,H}^{(j)}, a_{t,H}^{(j)}\right), \qquad U^{(j)} = \sum_{k=0}^{H-1} \gamma^k u_{t,k}^{(j)}, \tag{6}$$

where $\hat{r}_{t,k}^{(j)}$ and $\hat{c}_{t,k}^{(j)}$ are the predicted reward and cost at step k, $\hat{Q}$ and $\hat{C}$ are the value and cost bootstraps at the planning horizon, and $u_{t,k}^{(j)}$ is the normalized uncertainty for that transition.

Uncertainty integration strategy. Trajectory scores $S^{(j)}$ are constructed from $(V^{(j)}, C^{(j)}, U^{(j)})$ to penalize or filter trajectories that violate cost or uncertainty limits. In this work we use MUARL m0f as the main instantiation, which combines soft Lagrangian penalties on both cost and uncertainty with a feasibility-based fallback: when fewer than N_e (elite set size) low-uncertainty trajectories are available, the planner switches to cost-minimizing scoring to guide the agent back toward predictable regions. We also study two alternative integration strategies—MUARL m0 (pure soft penalties on cost and uncertainty) and MUARL m1 (hard uncertainty filtering with fallback)—as ablations; detailed formulations are presented in Appendix B.

MPPI weighting and update. MPPI converts trajectory scores into importance weights and updates the action distribution toward high-scoring samples:

$$w^{(j)} \propto \exp\left(\frac{1}{\eta} S^{(j)}\right), \qquad \bar{a}_k \leftarrow \sum_{j=1}^{N} w^{(j)} a_k^{(j)}, \tag{7}$$

with temperature η controlling concentration. The executed action is the first element of the refined mean sequence, in a receding-horizon loop.

Table 1: Zone-based OOD detection and efficiency for three dynamics uncertainty estimators (IID vs. withheld OOD zones).

Method	AUROC	Cohen's d	IID $\bar{u}$	OOD $\bar{u}$	Rel. sep.	Time (ms)
Stochastic Ensemble	0.970	2.06	**0.0034**	**0.0131**	2.8	0.203
Density Regression	0.736	0.29	1.0000	1.0006	0.0	1.451
Evidential	**0.997**	**5.78**	0.0309	0.4678	**14.1**	**0.154**

Why both constraints? Cost predictions encode domain semantics but may be overconfident under distribution shift. Conversely, uncertainty can reliably signal that the model is extrapolating, but it does not distinguish novel-but-safe from novel-and-dangerous states. Using both terms in the planner addresses these failure modes directly at decision time.

4 RESULTS

We evaluate evidential uncertainty as an OOD signal and its impact on safety–performance trade-offs under sampling-based planning. Experiments use a controlled unicycle task and Safety Gymnasium benchmarks. Unless otherwise stated, we refer to MUARL m0f as the default MUARL instantiation, and interpret m0 and m1 as alternative uncertainty-integration strategies examined in ablations. All reported results are averaged over 3 random seeds.

4.1 ABLATIONS ON CONTROLLED NAVIGATION

We consider a continuous 2D navigation task with a nonholonomic unicycle car that must reach a fixed goal region while avoiding a central obstacle and two disjoint constraint zones. The state comprises position (x, y), heading, and speed; the agent controls angular rate and acceleration. Entering a constraint zone incurs a unit cost per time step, and an episode is deemed *feasible* if its cumulative cost stays below a fixed limit. To study epistemic uncertainty, we designate specific regions of the workspace as out-of-distribution (OOD) by withholding their transitions from the replay buffer; in the aleatoric setting, we additionally corrupt observations in those regions with high-variance noise. These design choices allow us to evaluate whether uncertainty estimates distinguish in-distribution from OOD states and how they interact with explicit cost constraints during planning.

4.1.1 UNCERTAINTY ESTIMATION QUALITY AND EFFICIENCY

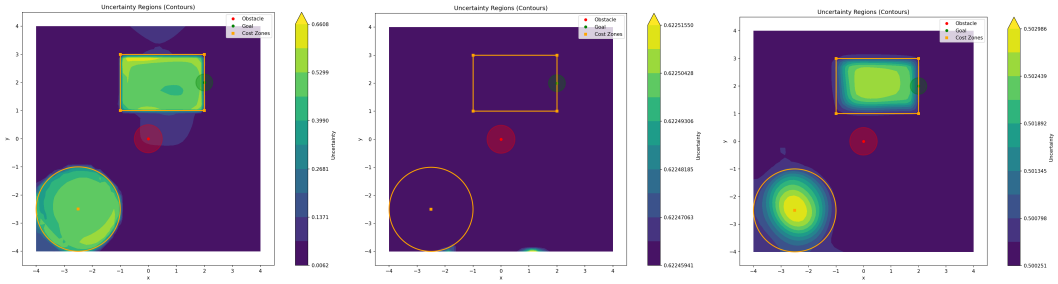
We first compare three uncertainty estimators for the dynamics model: evidential regression (Amini et al., 2020; Meinert et al., 2023), density regression with normalizing flows (Manh Bui & Liu, 2024), and stochastic ensembles (Chua et al., 2018). The environment is partitioned into IID and OOD regions by withholding transitions from designated zones during training and corrupting observations inside those zones, so a good epistemic estimator should assign higher uncertainty in OOD areas.

Table 1 shows that evidential regression yields near-perfect OOD discrimination (AUROC 0.997) and the strongest separation between IID and OOD uncertainties (Cohen's $d = 5.78$, relative separation $14.1\times$), while also achieving the lowest inference latency (0.154 ms per prediction). The stochastic ensemble is competitive in AUROC (0.970) but exhibits substantially weaker separation ($2.8\times$) and higher compute due to multi-model evaluation. Density regression performs poorly for zone discrimination (AUROC 0.736, essentially flat IID/OOD uncertainty), and is the slowest method (1.451 ms), making it unsuitable for MPPI-style planning workloads.

We therefore use evidential uncertainty in all subsequent experiments due to its quality–efficiency trade-off and its single-pass suitability for sampling-based planning.

4.1.2 IS UNCERTAINTY ALONE SUFFICIENT FOR SAFETY?

We next test whether uncertainty can replace explicit cost constraints. We compare an uncertainty-only variant (Unc), a cost-only baseline (Cost), and MUARL variants that combine both mechanisms (m0, m0f, m1).



(a) Evidential (b) Density-based (Flows) (c) Stochastic Ensemble

Figure 2: Uncertainty maps comparing three dynamics uncertainty estimators across the unicycle environment state space. Purple/low values indicate low uncertainty (well-modeled states); yellow-green/high values indicate high uncertainty. **(a)** Evidential regression shows strong IID/OOD separation, with elevated uncertainty at environment boundaries. **(b)** Normalizing flows exhibit poor discrimination (flat uncertainty). **(c)** Stochastic ensembles produce uncertainty near hazards but insufficient separation for reliable OOD detection. Evidential is preferred for MPPI due to quality and single-pass efficiency.

Table 2: Uncertainty-only vs. cost-only vs. combined mechanisms (unicycle ablation). Feasibility is the percentage of episodes below the cost limit used in this study.

Method	Mean Reward	Mean Cost	Feasibility (%)
MUARL m1	47.74 ± 65.03	$\mathbf{0.78 \pm 2.11}$	**86.0**
MUARL m0f	34.75 ± 62.73	2.29 ± 4.20	66.0
MUARL m0	$\mathbf{94.45 \pm 45.72}$	1.30 ± 2.61	75.0
Cost	19.38 ± 49.23	2.46 ± 3.97	67.0
Unc	65.75 ± 61.56	7.64 ± 10.37	49.0

Table 2 shows uncertainty alone is insufficient: Unc reaches moderate reward but incurs substantially higher cost and low feasibility (49%). In contrast, MUARL variants substantially reduce mean cost while maintaining non-trivial reward. MUARL m0 achieves the highest reward among constrained variants in this study (mean cost 1.30), while MUARL m0f trades some return for lower variance and a more conservative cost profile, consistent with its feasibility-based fallback. Additional reconstructed cost and uncertainty maps for all MUARL variants are provided in Appendix D (aleatoric setting) and Appendix E (epistemic setting).

4.1.3 Benefits under aleatoric vs. epistemic uncertainty

In the aleatoric setting we inject high-variance observation noise inside the OOD zones while still withholding their transitions from the replay buffer, whereas in the epistemic setting we only withhold data, so uncertainty arises purely from lack of coverage rather than irreducible noise.

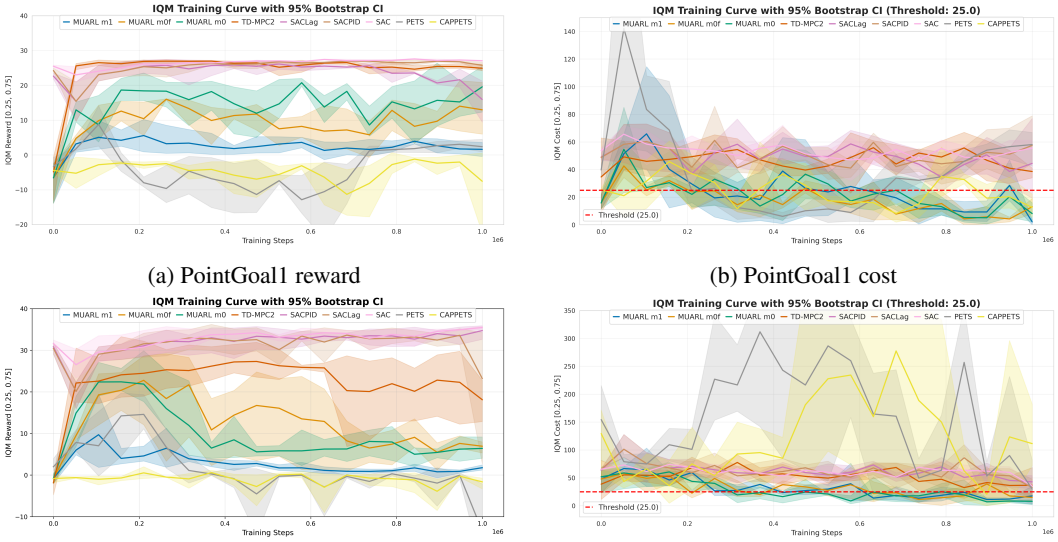
Table 3 shows that uncertainty-aware planning can reduce cost and/or reduce entries into the high-uncertainty region, with clearer gains in the epistemic setting where distribution shift is the dominant failure mode. In aleatoric settings, benefits are smaller and can reverse if uncertainty penalties overreact to benign stochasticity. Full learning curves and additional visualizations are given in Appendix D or the aleatoric case and Appendix E for the epistemic case. Taken together, they suggest that evidential uncertainty is most helpful as a reliability signal under distribution shift, and less useful when high noise is the main difficulty.

4.1.4 Soft penalties vs. hard filtering (uncertainty integration)

We compare two uncertainty integration strategies: m0 (soft uncertainty penalty) and m1 (hard filtering with fallback to cost-based scoring). Among the three MUARL variants, m0f behaves as an intermediate strategy between the high-return, low-conservatism regime of m0 and the strict avoidance of m1, by switching to cost-focused scoring only when the elite set cannot be formed from

Table 3: Uncertainty benefits under aleatoric and epistemic conditions (unicycle ablation). Here, "Inside Unc. Region" measures the average fraction of time steps per episode where the agent's state lies in regions with normalized uncertainty above the learned threshold.

Setting	Method	Avg. Return	Cost	Inside Unc. Region
Aleatoric	Cost	**76.65 ± 66.19**	0.78 ± 2.51	0.52 ± 2.12
Aleatoric	MUARL m0	63.33 ± 64.42	3.13 ± 6.72	1.14 ± 3.79
Aleatoric	MUARL m0f	38.33 ± 68.73	**0.23 ± 1.53**	**0.47 ± 2.97**
Aleatoric	MUARL m1	-5.26 ± 31.45	2.00 ± 7.12	1.22 ± 4.81
Epistemic	Cost	74.91 ± 64.34	1.65 ± 3.33	0.67 ± 1.76
Epistemic	MUARL m0	**93.18 ± 55.43**	1.54 ± 3.25	**0.52 ± 1.57**
Epistemic	MUARL m0f	36.44 ± 60.85	**0.27 ± 1.31**	0.64 ± 1.79
Epistemic	MUARL m1	51.10 ± 61.36	0.61 ± 1.93	0.64 ± 1.75



(a) PointGoal1 reward

(b) PointGoal1 cost

(c) CarGoal1 reward

(d) CarGoal1 cost

Figure 3: Training curves on Safety Gymnasium benchmarks. Top: SafetyPointGoal1; bottom: SafetyCarGoal1. Model-free baselines reach higher returns but incur persistent constraint violations, while MUARL variants learn more conservative policies with substantially lower costs across both environments.

low-uncertainty trajectories. Across ablations, soft penalties preserve higher task return, while hard filtering is more conservative but brittle under stringent thresholds, illustrating a practical performance–safety trade-off.

4.2 Safety Gymnasium benchmarks

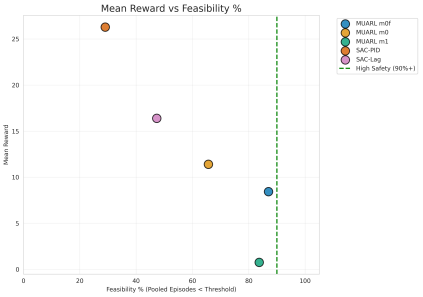
We benchmark MUARL on SafetyPointGoal1 and SafetyCarGoal1 (Ji et al., 2023a), comparing against strong model-free constrained baselines (SAC-PID(Stooke et al., 2020) and SAC-Lagrangian(Ray et al., 2019)) and the model-based RL method PETS (Chua et al., 2018) and its variant that employs constraints and uncertainty CAPPETS(Ma et al., 2022). For more information on the baselines attend to Appendix C.4. Performance is measured using mean return, mean episode cost, and feasibility (fraction of episodes below the cost threshold of 25). We take MUARL m0f as the primary instantiation, and m0 and m1 as ablation variants (see Appendix B for full definitions).

In preliminary experiments on these tasks, PETS and CAPPETS exhibited highly unstable learning dynamics, with rewards and costs oscillating over wide ranges and failing to converge to consistent

Table 4: Safety Gymnasium final evaluation (100 episodes/seed): mean return, mean cost, and feasibility.

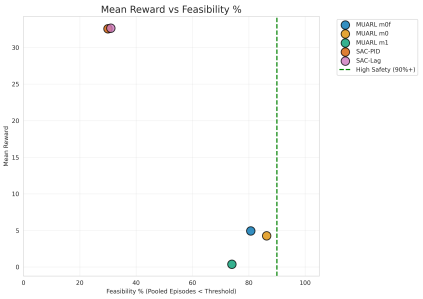
Environment	Method	Mean Reward	Mean Cost	Feasibility (%)
SafetyPointGoal1	MUARL m0	11.41 ± 7.53	25.16 ± 37.64	65.7
SafetyPointGoal1	MUARL m0f	8.44 ± 4.47	**10.36** ± **25.12**	**87.0**
SafetyPointGoal1	MUARL m1	0.77 ± 1.68	12.24 ± 31.93	83.7
SafetyPointGoal1	SAC-PID	**26.28** ± **1.82**	48.84 ± 34.33	29.0
SafetyPointGoal1	SAC-Lag	16.40 ± 9.75	38.46 ± 59.66	47.3
SafetyCarGoal1	MUARL m0	4.26 ± 3.50	**13.18** ± **38.98**	**86.3**
SafetyCarGoal1	MUARL m0f	4.93 ± 3.28	17.52 ± 44.72	80.7
SafetyCarGoal1	MUARL m1	0.38 ± 1.95	24.18 ± 46.66	74.0
SafetyCarGoal1	SAC-PID	32.57 ± 2.47	55.63 ± 44.32	30.0
SafetyCarGoal1	SAC-Lag	**32.64** ± **4.92**	52.89 ± 43.02	31.0

(a) PointGoal1: reward vs. cost



(b) PointGoal1: reward vs. feasibility

(c) CarGoal1: reward vs. cost



(d) CarGoal1: reward vs. feasibility

Figure 4: Final evaluation on SafetyPointGoal1 (top) and SafetyCarGoal1 (bottom). Left: mean reward versus mean episode cost, with the red vertical line marking the cost threshold of 25. Markers denote method means, error bars (if present) show standard deviation across seeds. Right: mean reward versus feasibility, with the green vertical line marking high safety at 90% feasibility. MUARL variants cluster closer to the safe region than model-free baselines in both environments.

policies. As a result, they do not appear in the final quantitative comparison in Table 4, which focuses on methods that reached stable performance.

Table 4 shows MUARL variants achieve substantially higher feasibility than model-free baselines in both environments, while maintaining non-trivial returns. In SafetyPointGoal1, MUARL m0f achieves the strongest safety–performance balance among MUARL variants (cost 10.36, feasibility 87.0%), whereas SAC-PID attains higher return but violates constraints frequently (feasibility 29.0%). In SafetyCarGoal1, MUARL m0 attains feasibility 86.3% with low cost (13.18), while SAC-PID and SAC-Lagrangian again exhibit high costs and low feasibility ($\approx 30\%$).

Across both settings, evidential uncertainty acts as a reliable OOD signal that works at the scale required by sampling-based planning. Using it alongside explicit cost constraints in the MPPI ob-

jective increases feasibility compared to model-free baselines, at the price of more conservative returns. The controlled ablations and Safety Gymnasium runs together suggest that single-pass evidential uncertainty is a practical decision-time signal for constrained MPC.

5 Discussion and Conclusion

Across our experiments, uncertainty was most useful as a *reliability* signal: it highlighted regions where latent rollouts are likely inaccurate due to distribution shift, enabling the planner to avoid trajectories that look good in imagination but are unsupported by experience. In the controlled OOD setting, evidential uncertainty provided strong IID/OOD separation while remaining efficient for MPPI. On Safety Gymnasium, integrating uncertainty improved feasibility over model-free baselines, demonstrating benefits of decision-time trajectory screening in constrained navigation.

However, uncertainty alone is insufficient: the uncertainty-only variant violated constraints frequently, showing that uncertainty does not encode task-specific safety semantics. In high-aleatoric-noise environments, uncertainty penalties can become overly conservative, penalizing safe-but-noisy regions unless thresholds are carefully tuned. Hard filtering is also brittle: strict thresholds can trigger frequent fallbacks to myopic cost-minimizing behavior.

Evidential regression achieved the best quality–efficiency trade-off in our study ($0.154ms$ per prediction vs. $0.203ms$ for ensembles and $1.451ms$ for flows; Table 1), which matters because uncertainty estimation is invoked at every planning step of every candidate trajectory. With N samples and horizon H, multi-pass estimators multiply inner-loop cost proportionally, whereas evidential preserves approximately the same compute as deterministic prediction.

Limitations and future work. Evidential uncertainty behaves more like a distribution-shift detector than a calibrated Bayesian posterior (Juergens et al., 2024), so thresholds may not transfer cleanly across environments or learning stages. The approach relies on an accurate cost predictor; if the cost model is misspecified, the planner can still fail in safety-critical states. Coupling uncertainty penalties with Lagrangian adaptation introduces sensitivity to normalization and hyperparameters, and latent-space planning can mask safety-relevant features through representation errors.

In summary, evidential uncertainty gives a strong OOD signal that is still cheap enough for sampling-based MPC, but it only becomes truly useful for safety once it is paired with explicit cost constraints. Using both signals at decision time improves feasibility over model-free constrained baselines, at the expected cost of more conservative returns. An obvious next step is to study calibration and active data collection under constraints, and to test these ideas on real robots where model mismatch and tight real-time limits are unavoidable.

Acknowledgments

This work was partly supported by CNPq (HydroneLE, 405530/2022-6), FINEP (Subhydrone, 01.23.0481.00), and the NAUTEC - the robotics group at FURG. And the author acknowledges the financial support of the Human Resources Training Program of the National Agency of Petroleum, Natural Gas and Biofuels (PRH-ANP), managed by the São Paulo Research Foundation (FAPESP), Brazil. Process No. 2025/03736-5 - PRH-ANP Program 22/FURG.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. Actsafe: Active exploration with safety constraints for reinforcement learning. In *The Thirteenth*

International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=aKRADWBJ1I.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(Volume 5, 2022):411–444, 2022. ISSN 2573-5144. doi: https://doi.org/10.1146/annurev-control-042920-020211. URL https://www.annualreviews.org/content/journals/10.1146/annurev-control-042920-020211.

Bertrand Charpentier, Ransalu Senanayake, Mykel Kochenderfer, and Stephan Günnemann. Disentangling epistemic and aleatoric uncertainty in reinforcement learning. *arXiv preprint arXiv:2206.01558*, 2022.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/4fe5149039b52765bde64beb9f674940-Paper.pdf.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3de568f8597b94bda53149c7d7f5958c-Paper.pdf.

Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103*, 2019. URL https://arxiv.org/abs/1812.11103.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *ICML*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

Masha Itkina and Mykel Kochenderfer. Interpretable self-aware neural networks for robust trajectory prediction. In Karen Liu, Dana Kulic, and Jeff Ichnowski (eds.), *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 606–617. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/itkina23a.html.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5faf461eff3099671ad63c6f3f094f7f-Paper.pdf.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a. URL https://openreview.net/forum?id=WZmlxIuIGR.

Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304*, 2023b.

Mira Juergens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *International Conference on Machine Learning*, pp. 22624–22642. PMLR, 2024.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

Zuxin Liu, Hongyi Zhou, Baiming Chen, Sicheng Zhong, Martial Hebert, and Ding Zhao. Constrained model-based reinforcement learning with robust cross-entropy method, 2021. URL https://arxiv.org/abs/2010.07968.

Yecheng Jason Ma, Andrew Shen, Osbert Bastani, and Jayaraman Dinesh. Conservative and adaptive penalty for model-based safe reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5404–5412, Jun. 2022. doi: 10.1609/aaai.v36i5.20478. URL https://ojs.aaai.org/index.php/AAAI/article/view/20478.

Ha Manh Bui and Anqi Liu. Density-regression: Efficient and distance-aware deep regressor for uncertainty estimation under distribution shifts. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2998–3006. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/manh-bui24a.html.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26096. URL https://doi.org/10.1609/aaai.v37i8.26096.

Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, and Seong Joon Oh. *Trustworthy Machine Learning*. Number arXiv:2310.08215 in 2310.08215. arXiv, October 2023. doi: 10.48550/arXiv.2310.08215.

Aske Plaat, Walter Kosters, and Mike Preuss. Deep model-based reinforcement learning for high-dimensional problems, a survey. *arXiv preprint arXiv:2008.05598*, 2020. URL https://arxiv.org/abs/2008.05598.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019. URL https://cdn.openai.com/safexp-short.pdf.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *Conference on Robot Learning (CoRL)*, 2025.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. *International Conference on Machine Learning (ICML)*, pp. 9133–9143, 2020. URL http://proceedings.mlr.press/v119/stooke20a/stooke20a.pdf.

Alex Christopher Stutts, Danilo Erricolo, Theja Tulabandhula, and Amit Ranjan Trivedi. Echoes of socratic doubt: Embracing uncertainty in calibrated evidential reinforcement learning. *arXiv preprint arXiv:2402.07107*, 2024.

Min Wen and Ufuk Topcu. Constrained cross-entropy method for safe reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/34ffeb359a192eb8174b6854643cc046-Paper.pdf.

Yiming Zhang, Quan Vuong, and Keith W. Ross. First order constrained optimization in policy space. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

A Evidential Deep Learning Derivation

The evidential dynamics model uses a Normal-Inverse-Gamma (NIG) prior over Gaussian likelihoods. For each latent dimension i, we model:

$$\sigma_i^2 \sim \text{Inv-Gamma}(\alpha_i, \beta_i) \tag{8}$$

$$m_i|\sigma_i^2 \sim \mathcal{N}(\mu_i, \sigma_i^2/\nu_i) \tag{9}$$

$$z_{t+1,i}|m_i, \sigma_i^2 \sim \mathcal{N}(m_i, \sigma_i^2) \tag{10}$$

Marginalizing over m_i and σ_i^2 yields the Student-t predictive distribution shown in the main text. The evidential parameters $(\mu_i, \nu_i, \alpha_i, \beta_i)$ are constrained via activation functions: $\nu_i = \text{Softplus}(\tilde{\nu}_i) + 1$, $\alpha_i = \text{Softplus}(\tilde{\alpha}_i) + 1$, and $\beta_i = \text{Softplus}(\tilde{\beta}_i)$, where $\tilde{\cdot}$ denotes unconstrained network outputs.

The training loss combines a Student-t negative log-likelihood with an evidential regularizer (Amini et al., 2020) that penalizes overconfident predictions on erroneous samples.

B Additional Planning Objectives and Ablations

B.1 Uncertainty integration variants

Method 0: dual-penalty Lagrangian (MUARL m0). MUARL m0 treats both cost and uncertainty as soft constraints via Lagrangian penalties:

$$S^{(j)} = V^{(j)} - \lambda_c \max(0, C^{(j)} - c_{\max}) - \lambda_u \max(0, U^{(j)} - u_{\max}). \tag{11}$$

This allows the planner to accept moderately uncertain trajectories when the predicted return gain is high, while still discouraging violations. (Chua et al., 2018)

Method 0f: feasibility-based adaptive scoring (MUARL m0f). MUARL m0f first checks whether sufficiently many low-uncertainty candidates exist to populate the elite set used by MPPI. Let N_e denote the number of elite trajectories, and define

$$N^u_{\text{feas}} = \sum_{j=1}^{N} \mathbf{1}\{U^{(j)} \le u_{\max}\}. \tag{12}$$

If $N^u_{\text{feas}} \ge N_e$, MUARL m0f uses the standard Lagrangian score (as in m0). If $N^u_{\text{feas}} < N_e$, indicating that most sampled candidates lie in a high-uncertainty regime, MUARL m0f switches to a safety-focused fallback that prioritizes minimizing predicted cost:

$$S^{(j)} = -C^{(j)}. \tag{13}$$

This fallback is intended to guide the agent back toward safer, more predictable regions where low-uncertainty options reappear. (Ray et al., 2019)

Method 1: hard uncertainty filtering (MUARL m1). MUARL m1 enforces uncertainty as a hard constraint by discarding any trajectory whose cumulative uncertainty exceeds a threshold:

$$\mathcal{F}_u = \{j \mid U^{(j)} \leq u_{\max}\}. \tag{14}$$

If $|\mathcal{F}_u| \geq N_e$, only trajectories in $\mathcal{F}_u$ are eligible for elite selection, and they are scored using a cost-constrained Lagrangian:

$$S^{(j)} = V^{(j)} - \lambda_c \max(0, C^{(j)} - c_{\max}), \qquad j \in \mathcal{F}_u. \tag{15}$$

If $|\mathcal{F}_u| < N_e$, the hard constraint is temporarily infeasible; MUARL m1 falls back to cost-based scoring to recover feasibility:

$$S^{(j)} = -\, C^{(j)}. \tag{16}$$

Compared to m0, this strategy provides stricter avoidance of high-uncertainty rollouts, at the potential cost of conservatism when uncertainty is widespread.

Baselines (single-mechanism ablations). To isolate the contribution of each safety signal, we also evaluate:

- **Cost-only**: ignore uncertainty and enforce only the cost constraint,

$$S^{(j)} = V^{(j)} - \lambda_c \max(0, C^{(j)} - c_{\max}). \tag{17}$$

- **Uncertainty-only (Unc)**: ignore operational cost and constrain only uncertainty,

$$S^{(j)} = V^{(j)} - \lambda_u \max(0, U^{(j)} - u_{\max}). \tag{18}$$

C Training and Implementation Details

C.1 MUARL World-Model Architecture

Table 5 summarizes the neural network architecture for the MUARL world-model components (encoder, dynamics, reward/cost models, and termination predictor).

Table 5: Neural network architecture hyperparameters for MUARL world-model components. Values shown are for the standard configuration used in experiments unless otherwise noted.

Component	Parameter	Value	Notes
Encoder			
	Input dimension	d_o	Observation dimension
	FC layer 1	256 hidden units	
	FC layer 2	512 hidden units	
	Output dimension	512	Latent state dimension
	Activation function	Mish	
	Normalization	Layer norm, SimNorm (output)	
	Total parameters	$\approx$ 150K	
Dynamics Model			
	Input dimension	$d_z + d_a = 514$	Concatenated state-action (512 + 2)
	FC layer 1	512 hidden units	
	FC layer 2	512 hidden units	
	Evidential layer output	$4 \times d_z = 2048$ parameters	$(\gamma, \upsilon, \alpha, \beta)$
	Activation (trunk)	Mish	
	Activation (υ)	Softplus	Ensures $\upsilon > 0$
	Activation (α)	Softplus + 1	Ensures $\alpha > 1$
	Activation (β)	Softplus	Ensures $\beta > 0$
	Normalization	Layer norm, spectral norm (evidential)	
	Total parameters	$\approx$ 1.58M	
Reward/Cost Models			
	Input dimension	514	State-action pair
	FC layer 1	512 hidden units	
	FC layer 2	512 hidden units	
	Output dimension	101	Two-hot categorical bins
	Activation function	Mish	
	Normalization	Layer norm	
	Total parameters	$\approx$ 578K	
Termination Predictor			
	Input dimension	512	Latent state only
	FC layer 1	512 hidden units	
	FC layer 2	512 hidden units	
	Output dimension	1 (logit)	
	Output activation	Sigmoid	Converts to probability
	Activation function	Mish	
	Normalization	Layer norm	
	Total parameters	$\approx$ 526K	

C.2 Actor–Critic Architecture

The actor–critic components used for value learning and policy optimization are summarized in Table 6.

Table 6: Neural network architecture hyperparameters for MUARL actor–critic components. Values shown are for the standard configuration used in experiments unless otherwise noted.

Component	Parameter	Value	Notes
Q-Function (per ensemble member)			
	Input dimension	514	State-action pair
	FC layer 1	512 hidden units	Dropout 0.01
	FC layer 2	512 hidden units	
	Output dimension	101	Two-hot categorical bins
	Ensemble size	5	
	Activation function	Mish	
	Normalization	Layer norm	
	Total parameters (single)	$\approx$ 578K	Per ensemble member
	Total parameters (ensemble)	$\approx$ 2.89M	5 members $\times$ 578K
Cost Value Function (per ensemble member)			
	Architecture	Identical to Q-function	
	Ensemble size	5	
	Total parameters (ensemble)	$\approx$ 2.89M	
Policy Network			
	Input dimension	512	Latent state only
	FC layer 1	512 hidden units	
	FC layer 2	512 hidden units	
	Output dimension	4	$2 \times d_a$ (mean and log-std)
	Mean activation	Tanh	Bounds to $[-1, 1]$
	Activation function	Mish	
	Normalization	Layer norm	
	Total parameters	$\approx$ 527K	
Overall Architecture			
	Total learnable parameters	$\approx$ 9,750,000	All components combined

C.3 Training Hyperparameters

Tables 7 and 8 list the optimization settings, loss coefficients, MPPI planning hyperparameters, and general training configuration used in all experiments (unless otherwise specified in ablations).

Table 7: Training hyperparameters for MUARL: optimization settings and loss coefficients. All experiments use these settings unless otherwise specified in ablation studies.

Category	**Parameter**	**Value**
World Model Optimization		
	Optimizer	Adam
	Learning rate	10^{-3}
	Encoder learning rate	10^{-4}
	Beta coefficients	$\beta_1 = 0.9$, $\beta_2 = 0.999$
	Epsilon	10^{-8}
	Gradient clipping	Max norm 10.0
	Batch size	256 transitions
	Sequence length	$H + 1 = 6$
Policy Optimization		
	Optimizer	Adam
	Learning rate	3×10^{-4}
	Beta coefficients	$\beta_1 = 0.9$, $\beta_2 = 0.999$
	Epsilon	10^{-5}
	Gradient clipping	Max norm 10.0
	Batch size	512 states
Loss Coefficients		
	Consistency loss	$\lambda_{\text{cons}} = 20$
	Reward loss	$\lambda_r = 0.5$
	Cost loss	$\lambda_c = 0.5$
	Value loss	$\lambda_Q = 1.0$
	Cost value loss	$\lambda_C = 1.0$
	Termination loss	$\lambda_{\text{term}} = 0.5$
	Evidential regularization	$\lambda_{\text{reg}} = 0.01$
	Exploration bonus	$c_{\text{info}} = 0.01$
	Prior regularization	$\beta_{\text{prior}} = 0.5$
Lagrangian Adaptation		
	Optimizer	Adam
	Learning rate	$\alpha_\lambda = 0.005$
	Beta coefficients	$\beta_1 = 0.9$, $\beta_2 = 0.999$
	Epsilon	10^{-8}

Table 8: Training hyperparameters for MUARL: MPPI planning, uncertainty normalization, and general training settings. All experiments use these settings unless otherwise specified in ablation studies.

Category	Parameter	Value
MPPI Planning		
	Planning horizon	$H = 5$
	Number of samples	$N_s = 512$
	Number of elites	$N_e = 64$
	Policy trajectories	$N_\pi = 64$
	MPPI iterations	$K = 6$
	Temperature	$\beta = 0.5$
	Min action std	$\sigma_{\min} = 0.05$
	Max action std	$\sigma_{\max} = 2.0$
Uncertainty Normalization		
	Reservoir size	20,000 samples
	Lower percentile	5th percentile
	Upper percentile	95th percentile
	Update frequency	Every 100 training steps
	Uncertainty weight	$w_u = 1.0$
General Training		
	Discount factor	$\gamma = 0.99$
	Replay buffer size	1,000,000 transitions
	Random exploration episodes	5-10 episodes
	Gradient updates per step	1
	Target network momentum	$\tau = 0.01$
	Maximum steps	1,000,000
	Random seeds	3

C.4 Baseline Methods

PETS and CAPPETS Instability. PETS (Chua et al., 2018) and CAPPETS (Ma et al., 2022) exhibited highly unstable training dynamics in preliminary experiments on Safety Gymnasium tasks. Figures 3a and 3b show representative training curves for SafetyPointGoal1, where both methods display oscillating rewards and costs with failure to converge. PETS achieved peak rewards around 15-20 before collapsing, while CAPPETS oscillated between 0-25 reward with high variance. All of the hyperparameters used in these experiments were the defaults from OmniSafe (Ji et al., 2023b), but the environments are the original Safety Gymnasium tasks, not the model-based variants.

SAC, SAC-Lagrangian and SAC-PID. SAC (Haarnoja et al., 2019) is a widely adopted off-policy model-free RL algorithm that optimizes a maximum entropy objective to encourage exploration. SAC serves as an unconstrained baseline, optimizing only for task reward without explicit safety mechanisms. SAC-Lagrangian implements constrained policy optimization using gradient-based Lagrangian multiplier updates (Ray et al., 2019). The multiplier is optimized via gradient descent on the dual objective, providing an alternative adaptation mechanism to PID control. SAC-PID variant augments SAC with a PID-controlled Lagrangian multiplier for constraint handling (Stooke et al., 2020). The PID controller dynamically adjusts the penalty weight on cost violations based on recent constraint satisfaction, providing adaptive trade-offs between performance and safety. We use the default hyperparameters from OmniSafe for SAC, SAC-PID and SAC-Lagrangian, without per-task retuning, to match the Safety Gymnasium setup.

D Cost and Uncertainty Fields: Aleatoric Scenario

This section presents the training reward and cost curves together with the reconstructed cost maps and uncertainty fields for all methods under aleatoric uncertainty conditions.

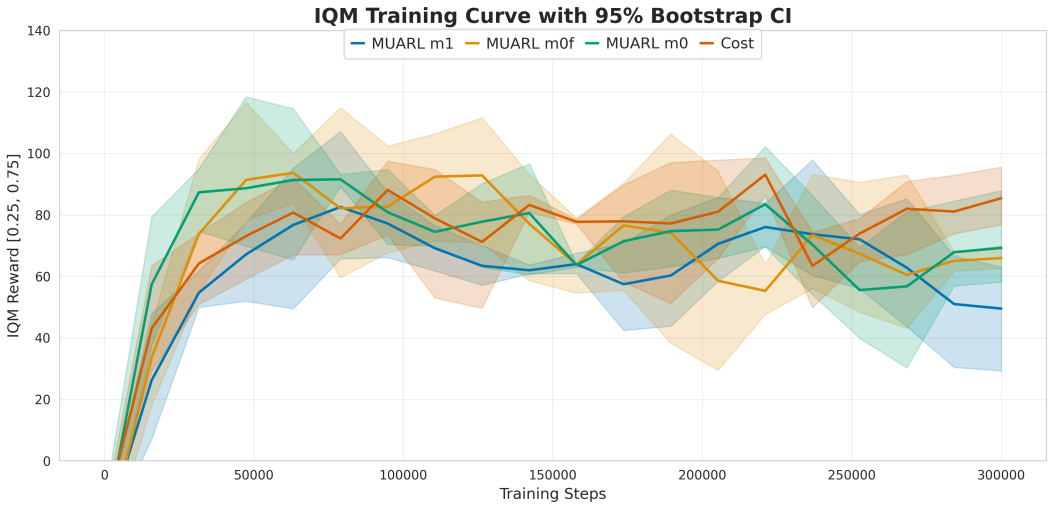


Figure 5: Training performance in the aleatoric uncertainty environment.

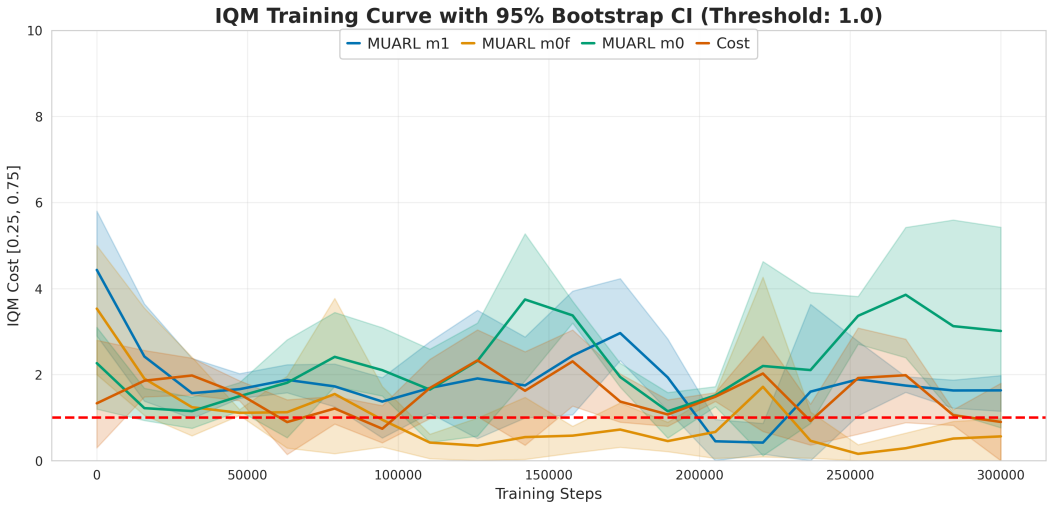


Figure 6: Safety–performance in the aleatoric uncertainty environment.

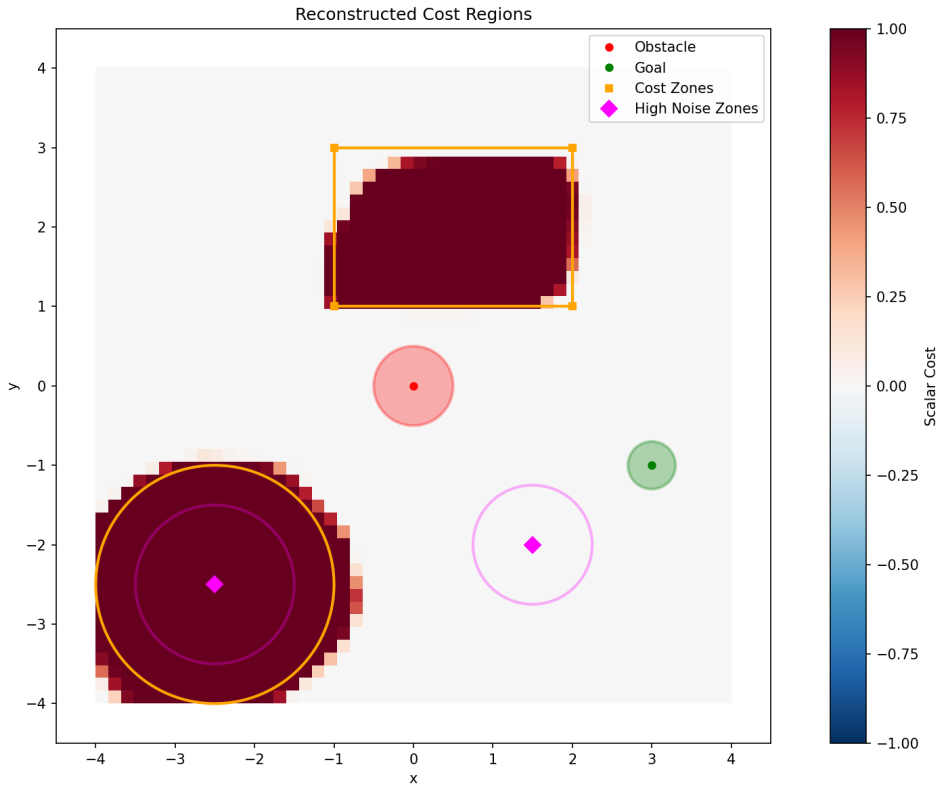
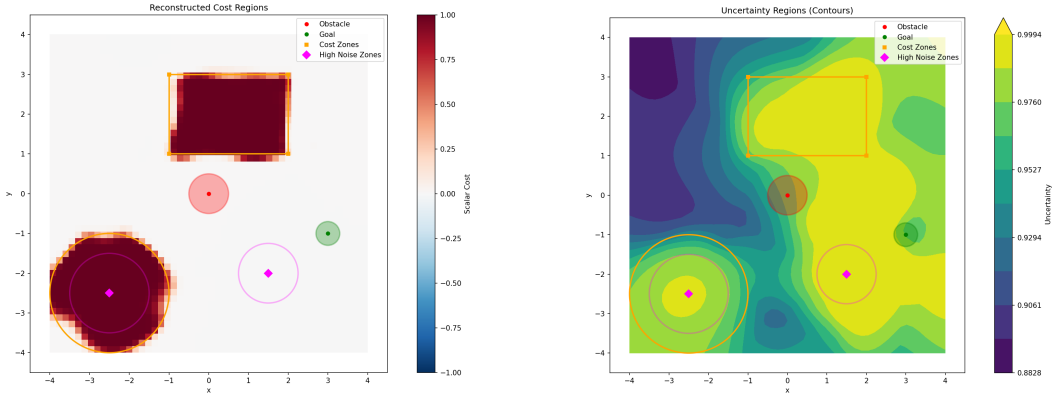


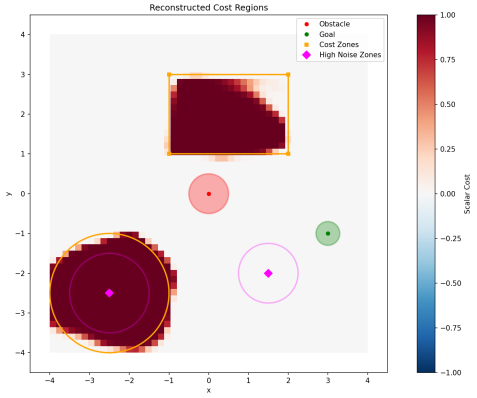
Figure 7: Reconstructed cost map for the cost-only baseline. The learned model predicts high costs (dark red) in obstacle and cost zone regions, with lower costs (lighter colors) in traversable areas. This baseline does not incorporate uncertainty into planning.



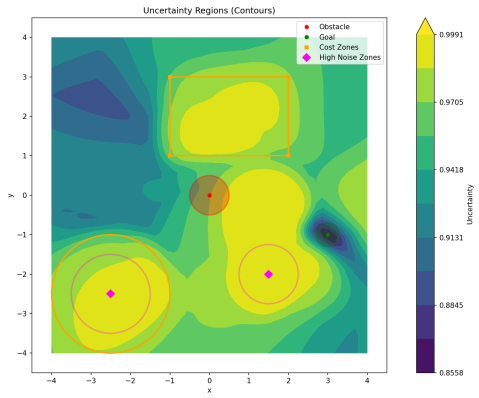
(a) Reconstructed cost map showing predicted operational costs across the state space.

(b) Uncertainty map displaying combined epistemic and aleatoric uncertainty u_{total}.

Figure 8: Method 0 with dual-penalty Lagrangian scoring. The model learns to predict both operational costs and uncertainty across the environment, with elevated uncertainty near obstacles and in regions with limited training data.

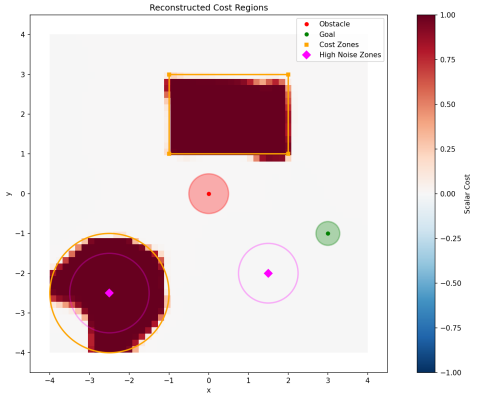


(a) Reconstructed cost map showing predicted operational costs across the state space.

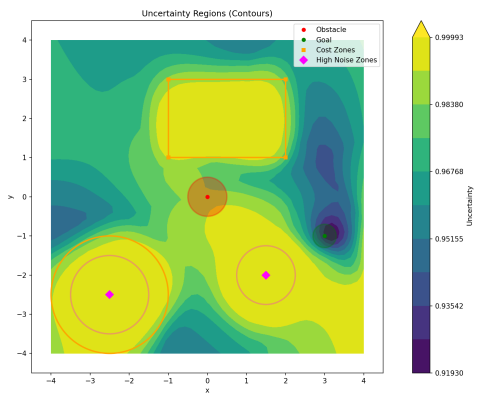


(b) Uncertainty map displaying combined epistemic and aleatoric uncertainty u_{total}.

Figure 9: Method 0 with feasibility-based adaptive scoring. The spatial distribution of cost and uncertainty estimates reflects the strategy's adaptive behavior, switching between reward-focused and safety-focused planning based on uncertainty levels.



(a) Reconstructed cost map showing predicted operational costs across the state space.



(b) Uncertainty map displaying combined epistemic and aleatoric uncertainty u_{total}.

Figure 10: Method 1 with hard uncertainty filtering. The learned model captures distinct uncertainty patterns that guide the hard constraint enforcement during trajectory selection, ensuring strict avoidance of high-uncertainty regions.

E COST AND UNCERTAINTY FIELDS: EPISTEMIC SCENARIO

This section presents the training reward and cost curves together with the reconstructed cost maps and uncertainty fields for all methods under epistemic uncertainty conditions.

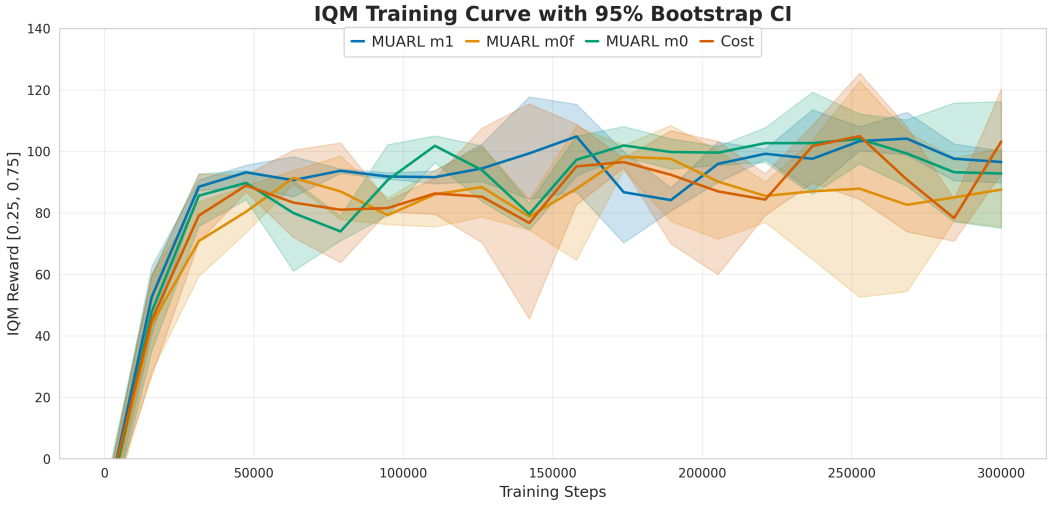


Figure 11: Training performance in the epistemic uncertainty environment.

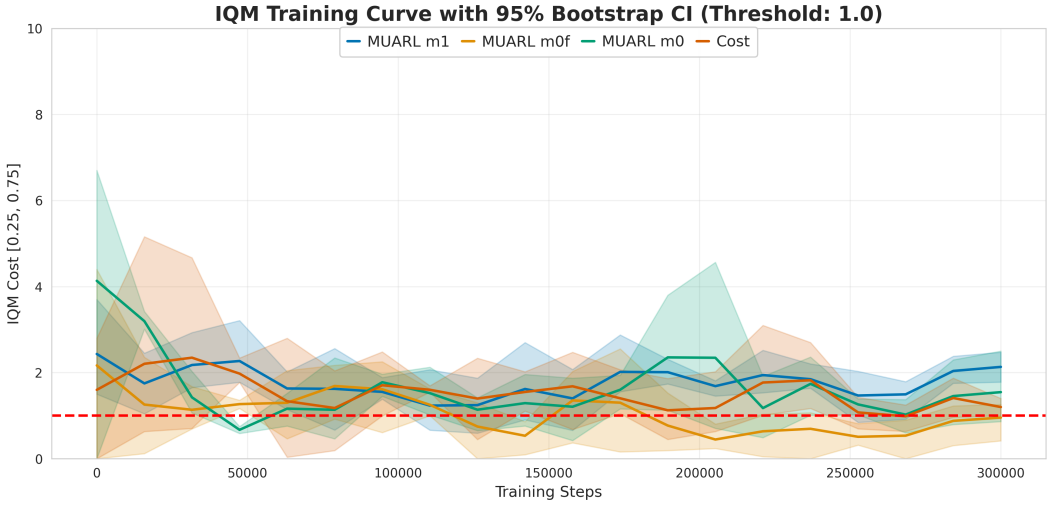


Figure 12: Safety–performance in the epistemic uncertainty environment.

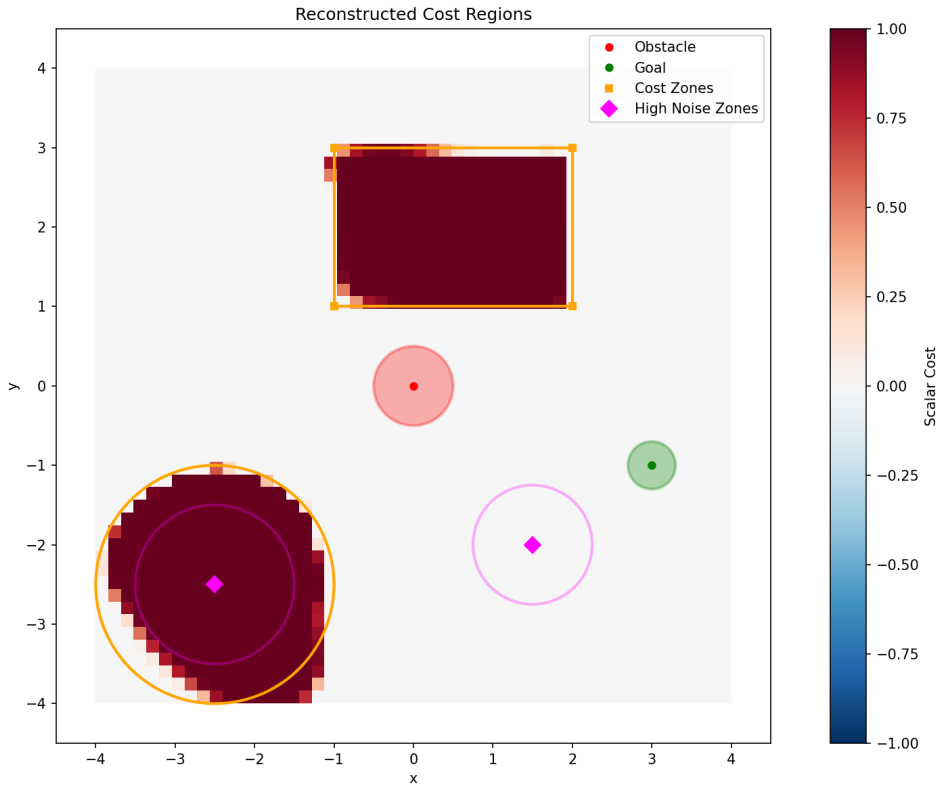
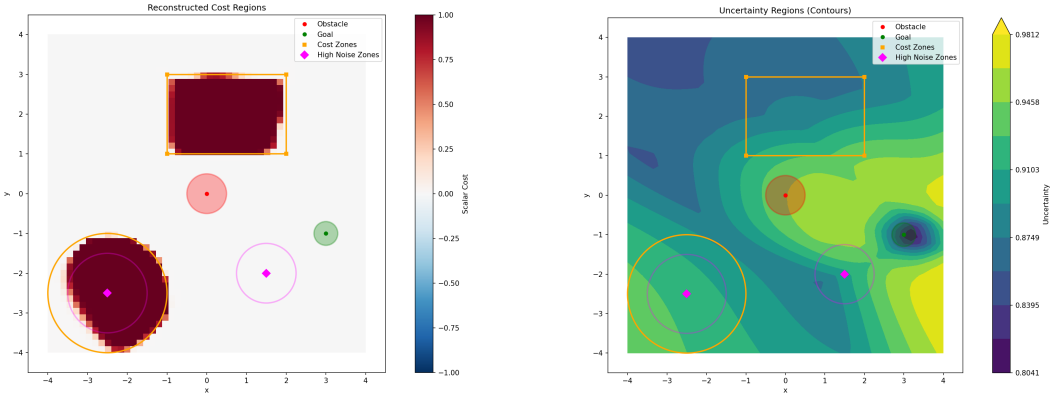


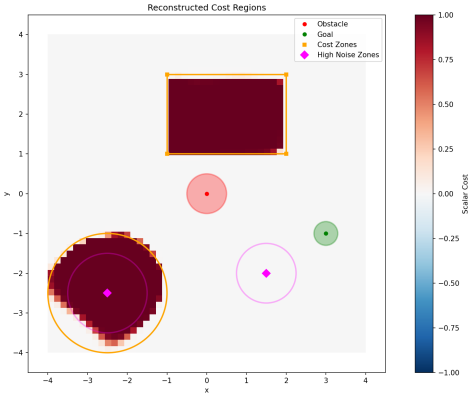
Figure 13: Reconstructed cost map for the cost-only baseline in the epistemic uncertainty scenario. The learned model predicts operational costs without accounting for regions where training data is sparse or absent.



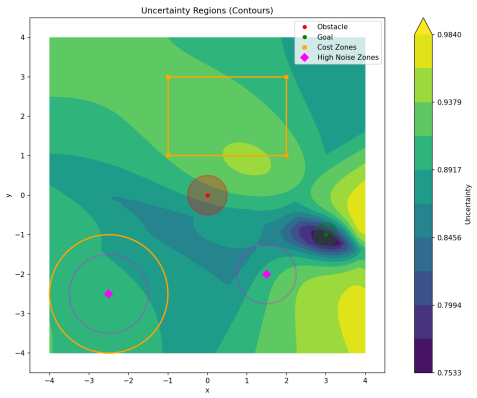
(a) Reconstructed cost map showing predicted operational costs across the state space.

(b) Uncertainty map with epistemic uncertainty dominating in under-explored regions.

Figure 14: Method 0 with dual-penalty Lagrangian scoring under epistemic uncertainty. The model captures elevated uncertainty in regions distant from training data, allowing the dual-penalty framework to balance exploration and exploitation.

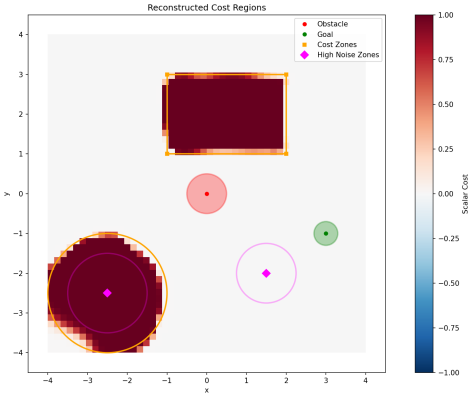


(a) Reconstructed cost map showing predicted operational costs across the state space.

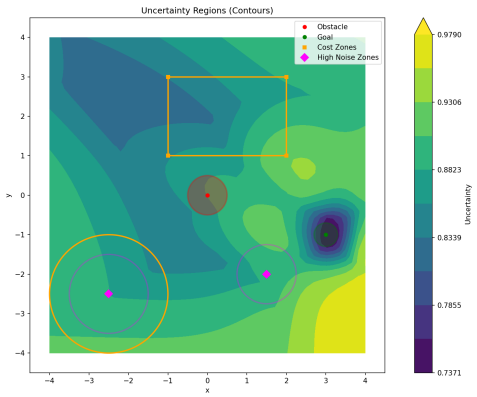


(b) Uncertainty map highlighting epistemic uncertainty in data-sparse regions.

Figure 15: Method 0 with feasibility-based adaptive scoring under epistemic uncertainty. The adaptive mechanism responds to epistemic uncertainty by switching to safety-focused planning when venturing into under-explored state space regions.



(a) Reconstructed cost map showing predicted operational costs across the state space.



(b) Uncertainty map revealing epistemic uncertainty patterns in unexplored regions.

Figure 16: Method 1 with hard uncertainty filtering under epistemic uncertainty. The hard constraint enforcement strictly avoids high-epistemic-uncertainty regions, prioritizing trajectories through well-modeled areas of the state space.