

Uncertainty-Driven Exploration

The exploration policy is optimized against an intrinsic reward equal to the disagreement of an ensemble of world model next-state predictors. Maximizing this reward drives the agent toward most informative areas for dynamics learning.

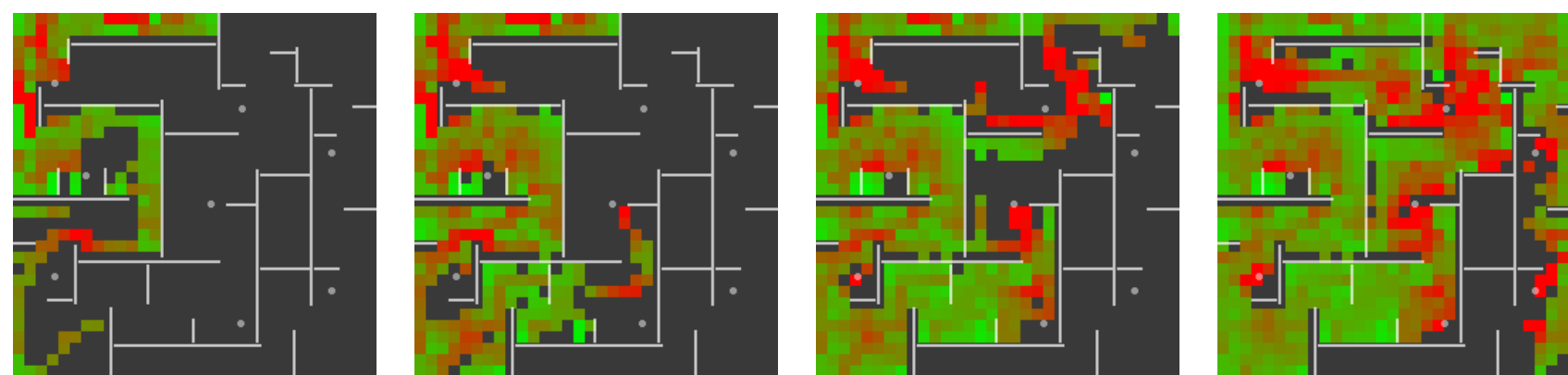


Figure 1: Disagreement over different timesteps during exploration. Red areas have a higher disagreement, meaning the world model is less certain about dynamics there.

Without any progress-based reward the cyberrunner learns how to solve the maze and explore all the positions with different actions and velocities.

Zero-Shot Task Adaptation

A world model learned through exploration can be reused for new, previously unseen tasks by optimizing the policy entirely within imagined rollouts, never interacting with the real environment.

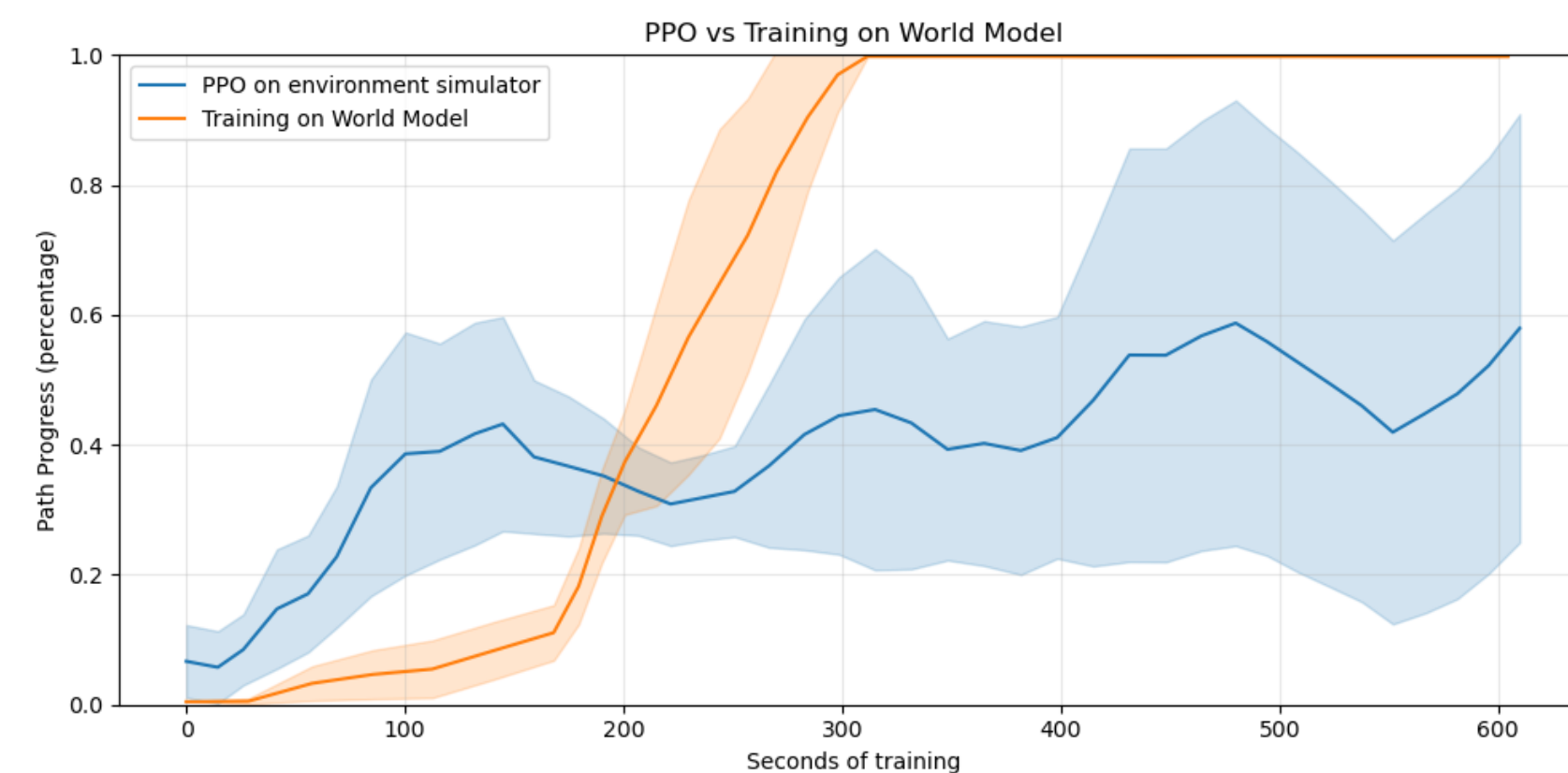


Figure 2: We compare a policy trained in imagination on the learned world model against a PPO baseline trained directly on the simulator.

We trained a world model expressive enough for zero-shot downstream adaptation. This is a major sample-efficiency win for hardware deployment, where real-world interactions are expensive.

CyberRunner is a robotic system that learns to solve a labyrinth by tilting a board to steer a marble through walls and holes. It relies on **Dreamer**, which learns a **world model** and trains the policy in imagined rollouts. We explore three directions: **pure exploration for zero-shot task adaptation**, **safe learning** to reduce falls during training, and **efficient exploration under sparse rewards** driven by epistemic uncertainty.

Safe Learning

The uncertainty-driven explorer keeps pushing the marble into novel states — and into holes that end the episode. **Goal:** keep its coverage while **cutting falls during training**, via a frozen *safe prior* that takes over only near danger.

Method

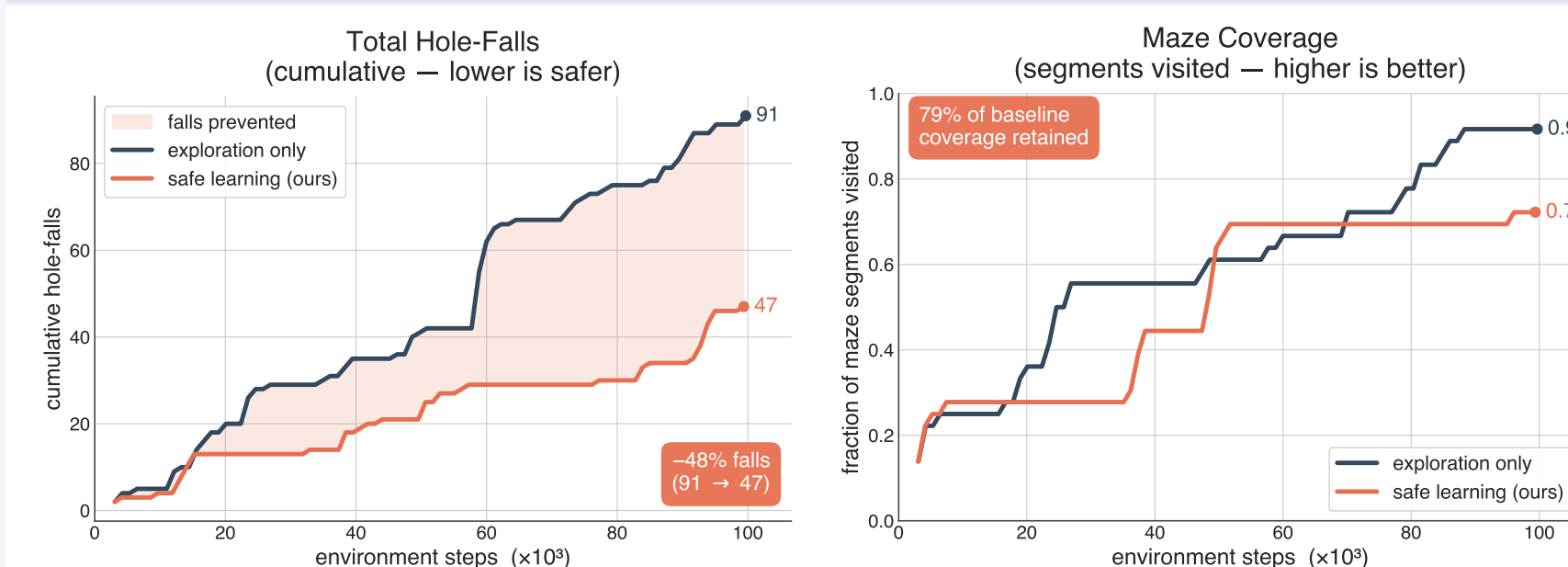
1. Train a safe prior. A policy trained once in simulation to keep the marble alive (reward +1 per surviving step; curriculum + domain randomization), then **frozen**. Its value $V_\pi(s) = \mathbb{E}[\sum_t \gamma^t r_t]$ — the expected survival time — gives a calibrated risk:

$$\text{risk}(s) = \text{clip}\left(1 - \frac{V_\pi(s)}{V_{\text{norm}}}, 0, 1\right).$$

2. Use it while training the world model. The explorer learns the dynamics; the prior keeps it safe two ways:

- **Runtime gate:** when $\Phi_t = c_{<t} + \gamma^t \text{risk}(s_t) \geq d$, the prior takes over for a bounded horizon — no fall on the executed run.
- **Imagination shaping:** penalize the explorer's reward in its rollouts, $\tilde{r}_t = r_t - \lambda \text{risk}(\hat{s}_t)$, so it *learns* to avoid states that need rescue.

Results: a Favourable Trade-off



At 10^5 env-steps, safe learning **cuts falls ~48%** (91→47) and doubles the survival time, at the cost of slightly less coverage (0.92→0.72); exploration *depth* is unchanged. The expected trade-off: *for the same budget, acting safely means exploring a little less.*

Sparse-Reward Exploration

Uncertainty-driven exploration optimises a combined reward

$$r_t = \underbrace{r_t^{\text{task}}}_{\text{extrinsic}} + \lambda \underbrace{r_t^{\text{int}}}_{\text{intrinsic}},$$

with r_t^{int} the **disagreement of an ensemble of world models over the next state** and λ trading task vs. exploration. Previous CyberRunner work used a **dense** r^{task} on path progress; we **sparsify** it, paying out only every N -th waypoint. We then ask: *does optimism ($\lambda > 0$) beat the $\lambda = 0$ baseline, and how does the gap scale with difficulty?* — comparing across *easy*, *medium* and *hard* layouts.

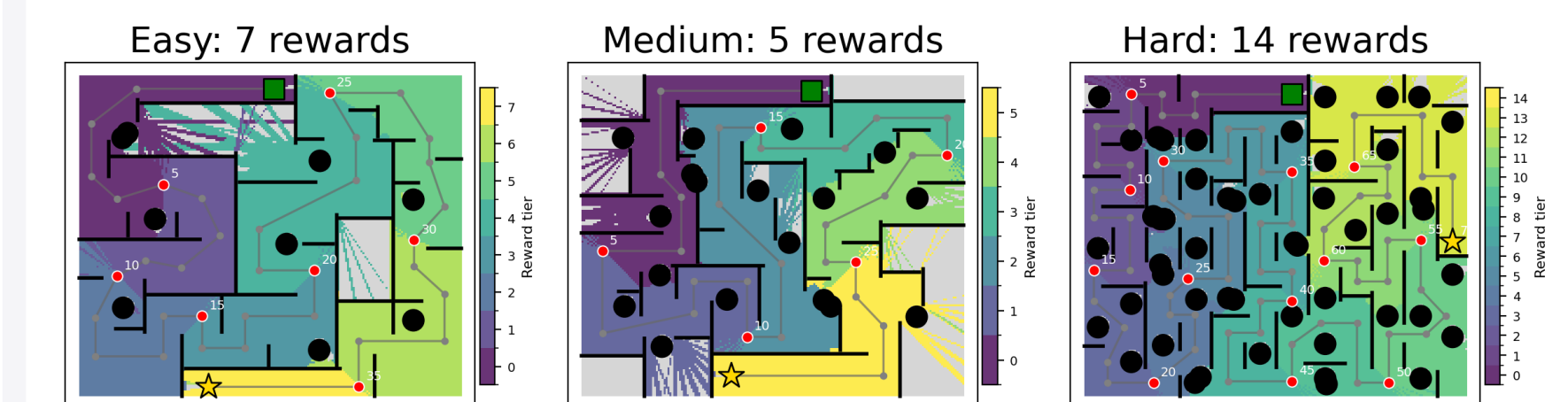


Figure 3: Three layouts of increasing difficulty. Each maze is split into coloured *reward zones* from start to goal; the agent is paid once per zone, on first entry in the episode.

Optimism scales with difficulty

Under the original **dense** reward, optimism brings no measurable benefit: $\lambda=0$ and optimism reach the goal in comparable time. Under **sparse** reward, we observe that as the task gets harder, the optimism edge increases.

