

COvolve: Adversarial Co-Evolution of Large-Language-Model-Generated Policies and Environments via Two-Player Zero-Sum Game

Alkis Sygkounas
Machine Perception and Interaction
Lab (MPI), Örebro University
Sweden
alkis.sygkounas@oru.se

Rishi Hazra
Machine Perception and Interaction
Lab (MPI), Örebro University
Sweden
rishi.hazra@oru.se

Andreas Persson
Machine Perception and Interaction
Lab (MPI), Örebro University
Sweden
andreas.persson@oru.se

Pedro Zuidberg Dos Martires
Machine Perception and Interaction
Lab (MPI), Örebro University
Sweden
pedro.zuidberg-dos-martires@oru.se

Amy Loutfi
Machine Perception and Interaction
Lab (MPI), Örebro University
Sweden
amy.loutfi@oru.se

Abstract

A central challenge in building continually improving agents is that training environments are typically static or manually constructed. This restricts continual learning and generalization beyond the training distribution. We address this with COvolve, a co-evolutionary framework that leverages large language models (LLMs) to generate both environments and agent policies, expressed as executable Python code. We model the interaction between environment and policy designers as a two-player zero-sum game, ensuring adversarial co-evolution in which environments expose policy weaknesses and policies adapt in response. This process induces an automated curriculum in which environments and policies co-evolve toward increasing complexity. To guarantee robustness and prevent forgetting as the curriculum progresses, we compute the mixed-strategy Nash equilibrium (MSNE) of the zero-sum game, thereby yielding a meta-policy. This MSNE meta-policy ensures that the agent does not forget to solve previously seen environments while learning to solve previously unseen ones. Experiments in urban driving, symbolic maze-solving, and geometric navigation showcase that COvolve produces progressively more complex environments. Our results demonstrate the potential of LLM-driven co-evolution to achieve open-ended learning without predefined task distributions or manual intervention.

Keywords

Co-evolution, Unsupervised Environment Design, Mixed-strategy Nash Equilibrium, Large Language Models

ACM Reference Format:

Alkis Sygkounas, Rishi Hazra, Andreas Persson, Pedro Zuidberg Dos Martires, and Amy Loutfi. 2026. COvolve: Adversarial Co-Evolution of Large-Language-Model-Generated Policies and Environments via Two-Player Zero-Sum Game. In *Genetic and Evolutionary Computation Conference*



This work is licensed under a Creative Commons Attribution 4.0 International License. *GECCO '26, San Jose, Costa Rica*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2487-9/2026/07
<https://doi.org/10.1145/3795095.3805144>

(*GECCO '26*), July 13–17, 2026, San Jose, Costa Rica. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3795095.3805144>

1 Introduction

Developing agents that continually acquire new skills in dynamic and unpredictable settings remains a core challenge in AI. Most current training pipelines still depend on large amounts of human-curated data, which is costly and often produces agents that generalize poorly beyond their training distribution [47]. While reinforcement learning (RL) offers an appealing alternative by allowing agents to learn through extensive interaction in a simulator [41], it inherits a fundamental limitation: the environments used for training are either fixed and/or manually designed. Constructing an environment distribution that captures the diversity and variability of real-world conditions is inherently difficult [5], and RL agents often fail to generalize beyond the narrow distribution they encounter during training [15, 22, 23]. Achieving robustness and transferability, therefore, requires exposing agents to a diverse and continually evolving curriculum of environments that adapt to their capabilities and expand the range of behaviors they must master [6, 23].

Unsupervised environment design (UED) [9, 21] addresses these limitations in training environments by automatically generating a curriculum of environments that adapts in difficulty based on the agent's performance. By dynamically tailoring environments to expose and challenge the agent's weaknesses, UED encourages continual learning. However, UED typically generates environments via randomization or simple heuristics, which limits the diversity and relevance of the resulting tasks. We overcome this by introducing COvolve¹, a co-evolutionary framework that frames UED as a two-player zero-sum game. COvolve leverages LLM-based code generation with rich priors to imaginatively design both environments and policies. In COvolve, an LLM-powered environment designer and policy designer compete adversarially to co-evolve more challenging levels and more capable policies, respectively, as conceptually illustrated in Figure 1.

¹Code and videos are available at: <https://github.com/sygkounas/covolve>

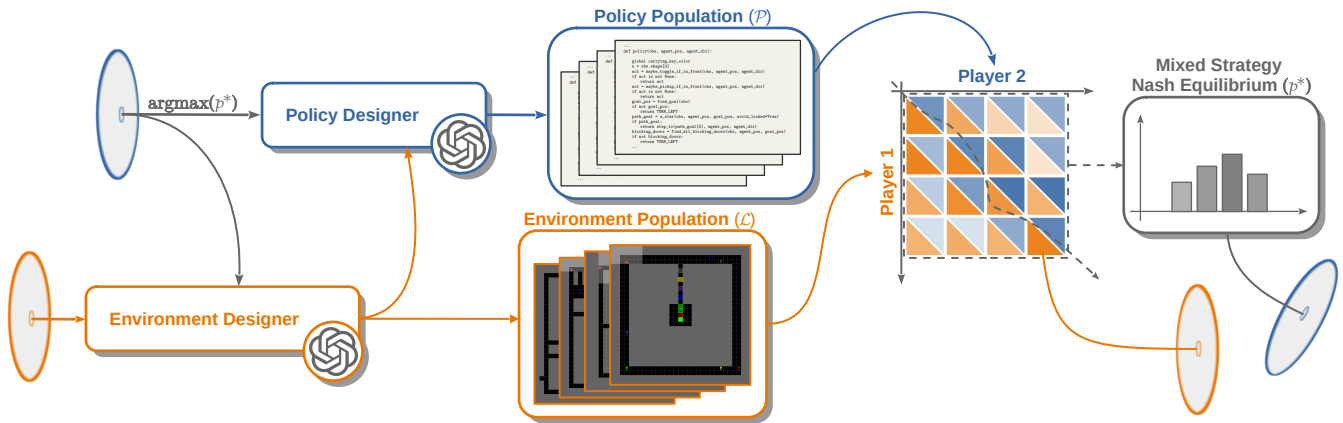


Figure 1: A conceptual overview of the proposed COvolve, comprised of an Environment Designer and a Policy Designer that co-evolve by playing a two-player zero-sum game. The Environment Designer generates increasingly challenging environments (as code), while the Policy Designer creates policies (as code) to solve them. A mixed-strategy Nash equilibrium enables robust, open-ended learning through continual adaptation.

Previous research has explored LLM-driven environment generation [12] and policy design [25] using code-based outputs, where both environments and policies are represented directly as executable programs. Such programmatic representations provide advantages over neural encodings, including improved generalizability to unseen scenarios [19, 45], greater modularity and reuse of behavioral components [11, 48], and stronger verifiability and interpretability [2, 46]. They also allow Python to express arbitrarily complex environment and policy logic, while LLMs contribute priors that enable the automatic synthesis of diverse tasks without hand-crafted templates [12]. However, existing approaches typically address either environment generation without promoting robust agent learning [12], or policy design without continual adaptation to new challenges [26]. In contrast, COvolve integrates both aspects into a closed-loop LLM-driven co-evolution process that simultaneously advances environment complexity and policy capability. Concretely, we make the following contributions:

(1) *Game-theoretic Framework for Robust Policy Design.* We frame the co-evolution as a two-player zero-sum game between a policy player and an environment player, where the payoff is the policy’s success rate in each environment. At each iteration, COvolve maintains populations of policies and environments, evaluates all pairs to form an empirical payoff matrix, and computes the mixed-strategy Nash equilibrium of this matrix game. The resulting meta-policy distribution solves the max–min objective within the empirical meta-game [32], improving worst-case performance against the current environment set and guiding the environment player to generate levels that exploit weaknesses of the equilibrium distribution [24]. In contrast, prior approaches [50] train independent policies per environment, thereby compromising population-level robustness and causing catastrophic forgetting.

(2) *Empirical Evidence of Emergent Curriculum and Generalization.* We empirically demonstrate that COvolve produces increasingly challenging environments across diverse domains (urban driving,

maze-solving, and 2D navigation), with generated levels exhibiting escalating complexity and diversity over time. Crucially, our evaluation shows that computing the MSNE is essential to prevent catastrophic forgetting, unlike approaches such as Eurekaverse [26], which retain only the latest best policy and finetune on new environments, leading to forgetting.

2 Related Works

Domain randomization (DR) exposes agents to a broad distribution of environments [20, 44] but lacks adaptivity and often produces trivial or unsolvable tasks [9]. Unsupervised environment design (UED) addresses this by automatically generating curricula tailored to agent performance. For instance, minimax adversarial training selects environments that minimize the agent’s reward [31, 35, 39]. However, it can produce overly difficult tasks unless constrained [9]. Regret-based methods like PAIRED [9] address this by defining regret relative to an approximated optimal policy to ensure solvability. While our work uses a minimax adversary, future directions could also incorporate regret-based strategies to avoid generating unsolvable levels. Crucially, our LLM-driven co-evolution introduces data-driven priors that enable the design of more challenging and relevant environments than classical, heuristic-based UED.

Recent work uses LLMs to generate and automate environment design [12, 49], world model generation [8, 43], and reward specification in RL [17, 29]. However, most frameworks either decouple environment and agent learning or focus only on environment generation, limiting agent robustness. Our framework enables fully closed-loop co-evolution, automatically generating a curriculum that adapts to both the agent and the environment. We use game-theoretic principles to maintain a diverse policy population via a mixed-strategy Nash equilibrium, yielding a meta-policy that is robust across the evolving set of generated environments and provides a principled population-level objective for continual adaptation.

In parallel with environmental design, LLMs have been used to synthesize modular, generalizable, and interpretable code-based

policies. Approaches like Code-as-Policies[25], RL-GPT[28], and ProgPrompt[42] leverage LLMs to generate executable plans or combine code with RL controllers, but are typically limited to narrow task distributions. In contrast, our approach constructs robust, continually adaptive policies that learn within an open-ended, co-evolving curriculum.

Our work is also related to the self-play paradigm, in which models play dual roles to create a self-improvement loop. Here, LLMs create copies of themselves with different roles to improve without relying on human data. This has been used in domains like coding (Coder-Tester Agents) [27, 51] and reasoning (Challenger-Solver Agents) [3, 18]. The improvement step is directly applied to the LLMs, which can be inefficient for domains where solutions can be represented by compact policies rather than large, monolithic models. In contrast, COvolve harnesses LLMs to drive the design of specialized agents that are modular, interpretable, and easier to deploy. A more concurrent work is Bachrach et al. [1] where LLMs produce strategies as code for playing against a Nash equilibrium mixture over the current population of strategies.

3 Preliminaries

3.1 Unsupervised Environment Design (UED)

Formally, a UED is defined over an *underspecified partially observable Markov decision process* (UPOMDP) [9], given by the 8-tuple $\mathcal{M} = (\Theta, \mathcal{S}, \mathcal{A}, \mathcal{O}, T, O, R, \gamma)$, with the last seven elements having the same meaning as in a standard POMDP: \mathcal{S} , \mathcal{A} , and \mathcal{O} are the sets of states, actions, and observations, respectively, T and O denote the transition and observation functions, R is the reward function, and $\gamma \in [0, 1]$ the discount factor.

The first element Θ of a UPOMDP \mathcal{M} denotes the space of *underspecified* environment parameters (e.g., number and position of obstacles, size of the grid). Picking a specific $\theta \in \Theta$ materializes a concrete POMDP. A UPOMDP can hence be viewed as a set of POMDPs. A concrete set of parameters $\theta \in \Theta$ is also referred to as a *level*. The choice of θ may influence the reward function $R : \mathcal{S} \times \mathcal{A} \times \Theta \rightarrow \mathbb{R}$, the transition function $T : \mathcal{S} \times \mathcal{A} \times \Theta \rightarrow \Delta(\mathcal{S})$, and the observation function $O : \mathcal{S} \times \Theta \rightarrow \mathcal{O}$, where $\Delta(\mathcal{S})$ is the set of all probability distributions over \mathcal{S} .

Given a level $\theta \in \Theta$, the expected discounted return (i.e. utility) of a policy π and a level θ is denoted as $U_\theta(\pi) = \mathbb{E}_{\tau \sim (\pi, \theta)} [G_\tau]$, with τ denoting trajectories sampled under the policy π at level θ . G_τ is the sum of discounted rewards for each trajectory: $\sum_{t=0}^T \gamma^t r_t$, with r_t being the reward collected at time step t . The optimal policy for level θ is then given by $\pi_\theta^* = \arg \max_\pi U_\theta(\pi)$.

The goal of UED is to train a policy that performs well across a broad distribution of environments. To this end, UED is typically framed as a two-player game, with an adversary Λ from which we can sample levels given a policy: $\theta \sim \Lambda(\pi)$. The adversary’s goal is to identify levels that challenge the policy π by optimizing the utility function $U_\theta(\pi)$ that exposes its weaknesses.

A simple example is that of a *minimax adversary* [31, 35, 39], which is a point distribution $\Lambda(\pi) = \arg \min_{\theta \in \Theta} U_\theta(\pi)$ and proposes new levels to minimize the policy performance. In response, a *maximin* policy is one that tries to perform well under the most adversarial level $\pi^* = \arg \max_{\pi \in \Pi} \min_{\theta \in \Theta} U_\theta(\pi)$. However, solving

this exactly is computationally intractable. In the following section, we introduce an efficient approximation method.

3.2 Policy Space Response Oracles (PSRO)

PSRO [24] is a general framework for multi-agent learning that addresses the fundamental challenge of non-stationarity in multi-agent environments. In such settings, the optimal policy for one agent depends on the policies of other agents, creating a moving target that makes traditional single-agent reinforcement learning approaches ineffective. Rather than attempting to learn a single “best” policy, PSRO builds and maintains a diverse population of policies over time. This approach provides robustness against various opponent strategies and reduces exploitability in competitive scenarios. In a 2-player setting, the PSRO framework operates through the following four-step iterative process:

- (1) Each player $i \in \{1, 2\}$ maintains a growing set of policies $\mathcal{P}^i = \{\pi_1^i, \pi_2^i, \dots, \pi_t^i\}$, creating a library of strategies for that individual player.
- (2) PSRO constructs a payoff matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{P}^1| \times |\mathcal{P}^2|}$ by evaluating all pairwise policy combinations, where each entry represents the expected payoffs for players 1 when player 1 uses policy π_i^1 and player 2 uses policy π_j^2 .
- (3) The framework computes a meta-policy for player 1 that determines how to mix the existing policies in the population.
- (4) For player 2, a new *best response* policy π_{t+1}^2 is trained to maximize performance against the player 1 meta-policy, and is subsequently added to the policy population \mathcal{P}^2 of player 2. This iterative process continues until convergence, resulting in a diverse and robust policy population.

4 Methodology

We adapt PSRO to UED by formulating environment and policy generation as a co-evolutionary process between an *Environment Designer* and a *Policy Designer*. The two designers iteratively generate, evaluate, and retain populations of environments and policies. At each iteration, new candidates are produced via structural program mutations [14, 52] of previously generated environments and policies, followed by fitness-based selection. This interaction is governed by a minimax objective, yielding a two-player zero-sum game summarized in Algorithm 1, with additional algorithmic details provided in Appendix A. Since both environments and policies are represented as executable Python code, Figure 2 illustrates their co-evolution across successive iterations.

4.1 Policy Designer

For the current level, the Policy Designer Ψ synthesizes a policy via an iterative best-response procedure. Given a level θ , the policy designer generates K candidate policy mutations by applying LLM-guided program transformations to the current best policy, conditioning on the level-specific observation and action spaces \mathcal{O}^θ and \mathcal{A}^θ . Each candidate policy $\tilde{\pi}_k$ is evaluated on θ , and the highest-performing candidate according to the utility U_θ is retained. The selected policy π constitutes an approximate best response for level θ and is appended to the growing policy sequence \mathcal{P} .

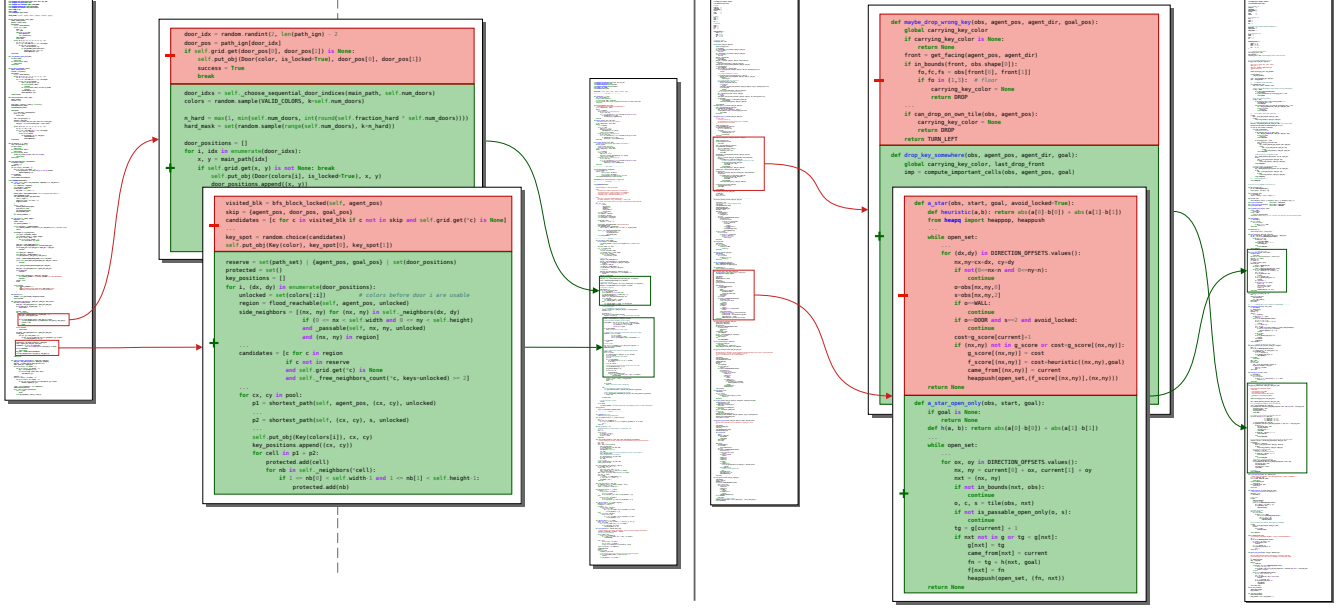


Figure 2: Example of co-evolution in COvolve. Left: successive environment implementations, progressing from ad hoc generation to a structured, parameterized design with explicit solvability checks and controllable chokepoints. Right: successive policy implementations, progressing from basic navigation to improved handling of keys and doors, together with refinements to the A* based planner for more reliable action selection. Highlighted code blocks indicate the changes introduced.

Algorithm 1 COvolve

- 1: **Require:** Initial environment θ_0
- 2: **Hyperparameters:** Total iterations T , candidate generated per level K
- 3: Initialize environment levels $\mathcal{L} \leftarrow \{\theta_0\}$, policy sequence $\mathcal{P} \leftarrow ()$, Payoff Matrix $M = []$ ▷ Initialize co-evolving populations
- 4: **for** $t = 0$ to T **do**
- 5: # 1. Policy Design (structural mutation + selection)
- 6: Generate K candidates: $\{\tilde{\pi}_1, \dots, \tilde{\pi}_K\} = \Psi(\pi_{t-1}, O^{\theta_{t-1}}, \mathcal{A}^{\theta_{t-1}})$
- 7: $\pi_t = \arg \max_j U_{\theta_{t-1}}(\tilde{\pi}_j)$
- 8: Append $\mathcal{P} \leftarrow \pi_t$
- 9: # 2. Update Payoff Matrix M (fitness evaluation)
- 10: **for** $i, j = 0$ to t **do**
- 11: $m_{ij} \leftarrow U_{\theta_j}(\pi_i)$
- 12: **end for** ▷ Cross-population fitness computation
- 13: # 3. Recompute MSNE (population-level update)
- 14: $p^* \leftarrow \text{SolveNash}(M)$ ▷ Eq. 1 ▷ Mixture over evolved policies
- 15: # 4. Best Response Environment Design (structural mutation + selection)
- 16: Generate K candidates: $\{\tilde{\theta}_1, \dots, \tilde{\theta}_K\} = \Lambda(\theta_{t-1}, p^*)$
- 17: $\theta_{t+1} = \arg \min_j \{\mathbb{E}_{\pi_i \sim p^*} [U_{\theta_j}(\pi_i)]\}$
- 18: Add $\mathcal{L} \leftarrow \theta_{t+1}$
- 19: **end for**
- 20: **Return:** MSNE Policy distribution p^* , environment levels \mathcal{L}

While each policy is tailored to an individual level, our broader goal is to obtain an approximation of the optimal policy π^* that performs well across levels. Since the interaction between the policy designer and the environment designer forms a two-player zero-sum game, a mixed-strategy Nash Equilibrium (MSNE) provides

a principled solution, ensuring that the obtained policy is robust under adversarial conditions. Building on PSRO, we achieve this by maintaining a growing sequence of policies $\mathcal{P} = \{\pi_1, \dots, \pi_t\}$ and previously generated levels $\mathcal{L} = \{\theta_1, \dots, \theta_t\}$, where each policy π_i is optimized for its corresponding level θ_i using the process described above. Let the payoff matrix be $M \in \mathbb{R}^{r \times t}$, where each entry $m_{ij} = U_{\theta_j}(\pi_i)$ denotes the expected return of policy π_i on level θ_j . The new MSNE is then computed via a minimax optimization where the policy agent maximizes its worst-case expected payoff [34]:

$$p^* = \arg \max_{p \in \Delta^r} \min_{j \in \{1, \dots, r\}} \sum_{i=1}^r p_i m_{ij} \quad \text{where,} \tag{1}$$

$$\Delta^r = \left\{ p \in \mathbb{R}^r \mid \sum_{i=1}^r p_i = 1, p_i \geq 0 \forall i \right\}.$$

Here, p^* is the mixture weights over policies that maximizes the worst-case expected return across all levels, defining the MSNE policy distribution p^* , where each policy π_i is sampled from π_{MSNE} with probability p_i^* . At the beginning of each episode, the agent samples a policy $\pi_i \sim p^*$ and takes actions according to $a \sim \pi$ in the given level.

4.2 Environment Designer

The Environment Designer generates a new level as a *best response* to the current MSNE policy. Its goal is to minimize the expected return of the mixture policy, thereby revealing its weaknesses. This adversarial loop encourages curriculum-like progression, automatically increasing the environment’s difficulty in response to the agent’s improvement.

Using a best-response update, the environment designer uses an LLM-based adversary Λ to generate K candidate mutations of the current environment $\{\hat{\theta}_1, \dots, \hat{\theta}_K\} = \Lambda(\theta, p^*)$. The candidate that minimizes performance under π_{MSNE} is selected.

The selected environment is added to the level set \mathcal{L} , after which a new policy is synthesized in response and appended to the policy sequence \mathcal{P} . This procedure repeats across iterations, expanding both the environment and policy sets.

5 Experiments

We evaluate COvolve across three complementary domains that capture distinct challenges in agent learning. MiniGrid [4] requires symbolic planning in procedurally generated mazes with sequential dependencies such as keys and doors. PyGame [36] emphasizes continuous 2D navigation with geometric reasoning and collision constraints, with difficulty scaling through denser obstacles and narrower traversable passages. CARLA [10] provides a high-fidelity urban driving setting with partial observability, dynamic vehicles, pedestrians, and traffic lights. Together, these domains provide a testbed for evaluating the emergence and robustness of curriculum under co-evolution.

5.1 Environments and Tasks

All main experiments use GPT-5.2 [33] as the generative model for both environment and policy synthesis. Prompts are provided in Appendix C. All generated code is dynamically validated and executed using `exec()`. For each policy–environment pair, we evaluate the payoff $U_\theta(\pi) \in [0, 1]$, averaged over 100 episodes, to populate the empirical payoff matrix \mathbf{M} . Complete environment specifications (action and observation spaces, task scaling, termination, and feasibility checks) are provided in Appendix B.

MiniGrid Symbolic Maze Solving. We use the MiniGrid–Empty environment as a base for generating symbolic maze-solving tasks. It is a fully-observable environment with $n \times n \times 3$ grids, augmented with the agent’s absolute position and orientation, resulting in an $n \times n \times 3 + 2$ observation vector, where n is the grid width and height. Each cell encodes object type, color identifier, and state (e.g., open/closed doors). The agent acts in a 5-action discrete space (*turn, move, pick up, drop, toggle*). Difficulty is scaled by enlarging grids, adding walls, and introducing locked doors with keys that must be retrieved and used in sequence, enforcing multi-step planning even in small mazes. We use handcrafted heuristics to validate if the generated environments are feasible (c.f., § 6 Limitations). Episodes terminate when the agent reaches the goal tile or when the step horizon is reached. A selection of evolved environments is shown in Figure 3, with further implementation details in Appendix B.1.

PyGame Geometric 2D Navigation. To test LLM’s ability to deal with *continuous action spaces*, we use a custom 2D navigation environment in which a circular agent must reach a rectangular goal zone while avoiding fixed rectangular obstacles. States are fully observable, consisting of the agent’s position and a list of all objects (obstacles and goals) with their positions and sizes (growing in size for each new level). The agent acts in a continuous space through 2D velocity commands. Difficulty increases when obstacles are added, the agent–goal distance is increased, and narrow

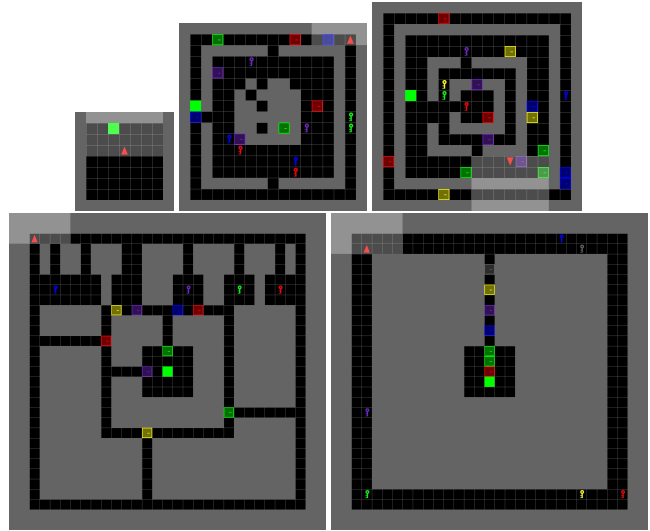


Figure 3: A selection of evolved MiniGrid environments produced by COvolve. Complexity increases from empty grids to larger mazes with dense walls and locked doors requiring corresponding keys. The agent must reach the green goal tile, often by planning multi-step sequences of key retrieval and door unlocking.

passages are created that may block traversal. This requires agents to identify traversable corridors relative to their size and to plan long detours when direct routes are infeasible. Episodes terminate when the agent overlaps the goal zone or when the step horizon is reached. A selection of evolved environments is shown in Figure 4, with further details in Appendix B.2.

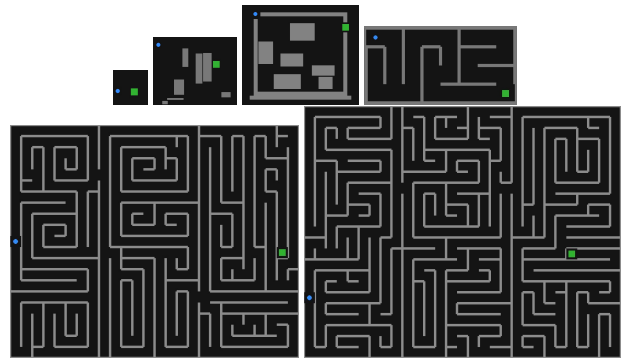


Figure 4: A selection of evolved PyGame environments produced by COvolve. Tasks progress from open arenas to cluttered maps with dense obstacles and narrow corridors. The agent must reach the goal zone while navigating collision-free paths through increasingly constrained layouts.

CARLA Urban Driving. We evaluate urban driving in CARLA Town01, a high-fidelity simulator with vehicles, pedestrians, and traffic lights. The vehicle follows a prescribed route using continuous steering, throttle, and brake controls. Observations are

egocentric and partial, consisting of the vehicle’s kinematics, the nearest traffic light, and compact features of nearby vehicles and pedestrians. Task difficulty increases with varying traffic density and pedestrian activity, introducing adversarial behaviors such as abrupt braking or traffic-light violations. Episodes terminate upon route completion (success) or any infraction (collision, red-light violation, or timeout). This setting tests policy robustness under partial observability and multi-agent interactions with stochastic and sometimes adversarial actors. A selection of evolved environments is shown in Figure 5, with further details in Appendix B.3.

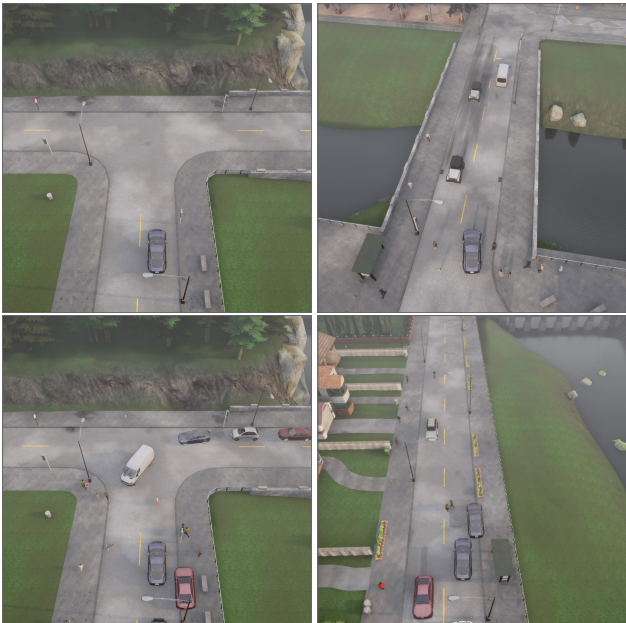


Figure 5: Selected CARLA environments produced by COvolve. Tasks progress from urban driving on empty roads to crowded streets with increasingly aggressive actor behaviors. The task for the agent is to drive along the street while following traffic rules (such as stopping at red lights) and, at the same time, adapt to increasingly unpredictable behaviors from fellow drivers and pedestrians.

5.2 Results

We evaluate whether adversarial co-evolution with equilibrium reasoning yields policies that both adapt to increasingly difficult environments and retain performance on previously generated ones. At each iteration, we compare three strategies: (i) *UED-Greedy*, which retains only the latest best policy; (ii) *UED-Uniform*, which samples uniformly from all policies generated up to the current iteration; and (iii) *COvolve*, which computes a mixed-strategy Nash equilibrium (MSNE) over the policy population. At iteration k , all strategies are evaluated on the full environment archive $\{\theta_0, \dots, \theta_k\}$.

UED-Greedy retains only the latest policy. *UED-Uniform* evaluates a uniform mixture over all policies generated up to the current iteration, controlling for mixture size without optimizing mixture

weights. *COvolve* instead computes a mixed-strategy Nash equilibrium over the policy population, selecting mixture weights that maximize the minimum return across the environment archive.

Because co-evolution produces distinct environment archives across runs, results from different random seeds are not directly comparable: averaging would mix performance over non-identical evaluation sets. We therefore report results from a single representative run and provide results for a second seed in Appendix D.3.

The results, presented in Figure 6, report three views: (i) *UED-Greedy*, where the latest policy π_k is evaluated on all environments generated up to iteration k ; (ii) *COvolve*, where policies are sampled from the MSNE and evaluated on the same environments; and (iii) a direct comparison between *UED-Greedy* and *COvolve* at iteration k , with performance averaged across all environments.

UED-Greedy policies are optimized for the most recently generated environment and exhibit reduced performance on earlier environments (Figure 6, left). In contrast, the MSNE mixture maintains performance across the full environment set as it grows over iterations (Figure 6, center). The aggregated comparison shows that, when the equilibrium is non-trivial, MSNE selection yields higher average performance than the latest-only *UED-Greedy* strategy when evaluated across all environments (Figure 6, right).

5.3 Generalization

We evaluate generalization beyond co-evolution on unseen standardized benchmark environments that preserve the same underlying task structure as the evolved environments. For MiniGrid, we consider *MiniGrid-MultiRoom-N6-v0* (six rooms), *MiniGrid-LockedRoom-v0*, and *MiniGrid-DoorKey-16x16-v0*. For CARLA, we evaluate on *Town02*, which is not encountered during co-evolution.

At iteration k , we compare three strategies under identical rollout settings: *UED-Greedy* (latest-only policy), *UED-Uniform*, and the *COvolve* MSNE policy distribution. For *UED-Uniform* and *COvolve*, a policy is sampled at the beginning of each episode according to the corresponding mixture distribution. Detailed environment specifications and their differences from the evolved tasks are provided in Appendix D.1. For each evolutionary seed, we run 100 evaluation episodes and compute the mean return. We then report the mean and standard deviation across two seeds. We do not report PyGame generalization results, as no standardized evaluation benchmarks exist for this domain.

Environment	UED-Greedy	UED-Uniform	COvolve
MiniGrid-MultiRoom-N6-v0	1.00 ± 0.00	0.86 ± 0.06	1.00 ± 0.00
DoorKey-16x16-v0 (MiniGrid)	1.00 ± 0.00	0.62 ± 0.24	1.00 ± 0.00
LockedRoom-v0 (MiniGrid)	1.00 ± 0.00	0.66 ± 0.16	1.00 ± 0.00
Town02 (CARLA)	0.62 ± 0.09	0.13 ± 0.06	0.71 ± 0.05

Table 1: Generalization to unseen environments. Results are reported as mean ± standard deviation across evolutionary seeds, with 100 evaluation episodes per seed.

Importantly, these standardized environments are substantially simpler than the environments generated during co-evolution. In contrast, the evolved environments typically contain multiple sequential key–door dependencies, narrow chokepoints, and adversarial obstacle placements; the benchmark tasks considered here

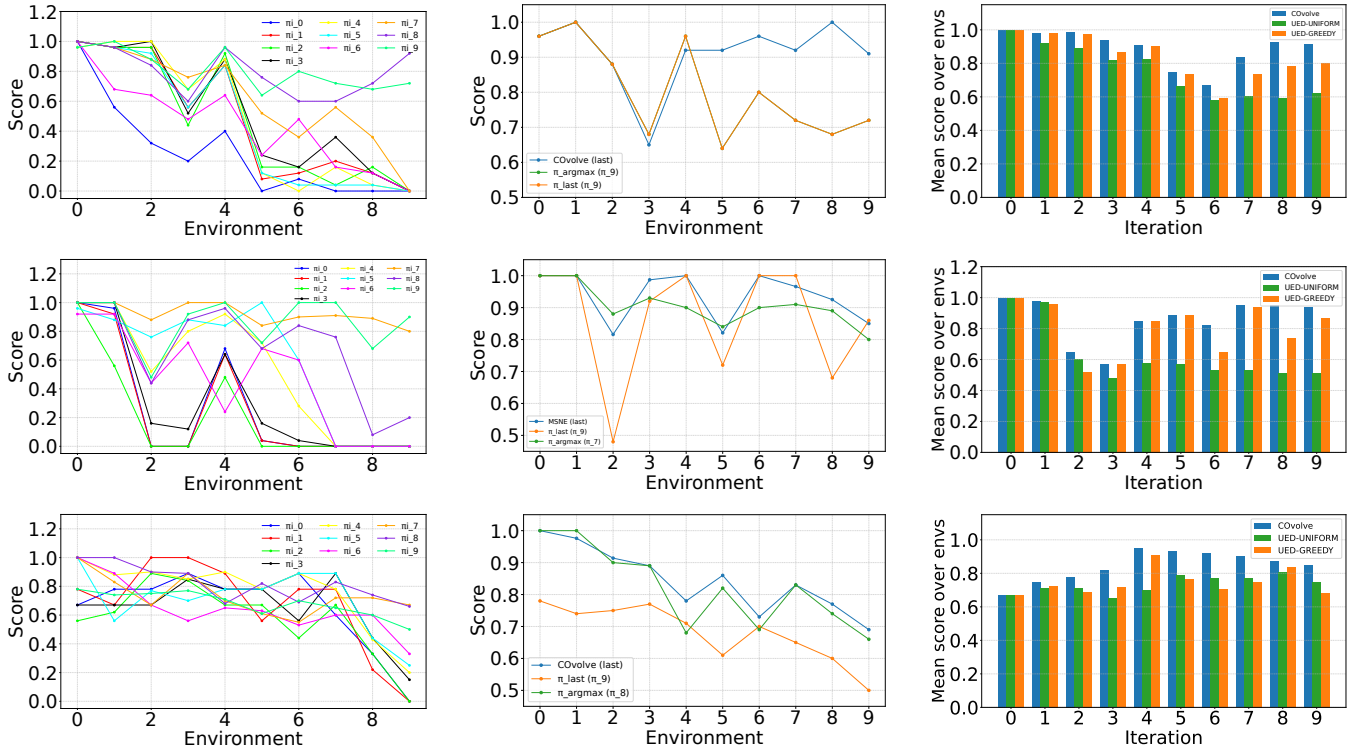


Figure 6: Performance during environment-policy co-evolution. Left: Success rates of all discovered policies evaluated on all environments generated during evolution (policy-environment payoff matrix). Center: Comparison between the mixed-strategy Nash equilibrium (MSNE) policy mixture, the best single policy π_{argmax} , and the latest policy π_k , evaluated on the environment archive $\{\theta_0, \dots, \theta_k\}$. Here, π_{argmax} denotes the policy that maximizes mean performance over the entire archive. For MiniGrid, $\pi_{\text{argmax}} = \pi_k$. Right: Mean success over $\{\theta_0, \dots, \theta_k\}$ for three strategies: UED-Greedy (latest policy only), UED-Uniform (uniform mixture over all policies), and COvolve (MSNE mixture). As evolution progresses and the latest policy forgets earlier environments, the MSNE mixture assigns probabilities to earlier policies to preserve worst-case performance over the archive.

involve at most a single locked door and a single key, with significantly less constrained geometry. As a result, strong performance on these unseen benchmarks, as indicated by Table 1, does not indicate overfitting of the specific environments, but rather that the learned policies have generalized. This highlights the role of the environment designer in constructing tasks harder than canonical benchmarks, thereby inducing a challenging curriculum.

5.4 Ablation Studies

Is curriculum necessary? We perform a zero-shot ablation in which the LLM generates policies directly for the hardest environment in each domain, without exposure to intermediate environments. Starting from an initial policy, the LLM applies up to k mutation steps, and we retain the best-performing policy. As shown in Figure 7, zero-shot generation consistently fails, demonstrating that progressive curriculum construction is necessary for effective policy synthesis.

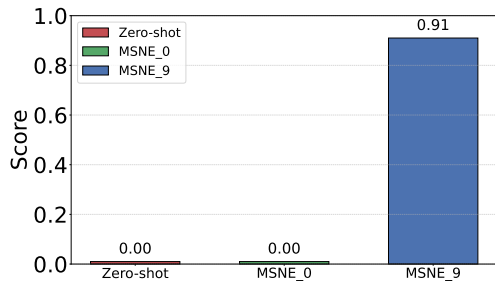
How do neural network-based RL approaches compare to LLM-generated code-as-policies in solving fully observable environments? To test whether standard RL can solve our fully observable evaluation tasks, we trained representative Stable-Baselines3 agents [37]

– PPO [40] and SAC (continuous control) [16] for PyGame, and PPO and QR-DQN (discrete) [7] for MiniGrid. Despite full observability, performance degrades sharply with task complexity: PPO and QR-DQN achieve near-zero rewards and success rates in the harder environments (SAC shows only limited improvements). See Appendix D.2 for details regarding training settings, curves, and success rates.

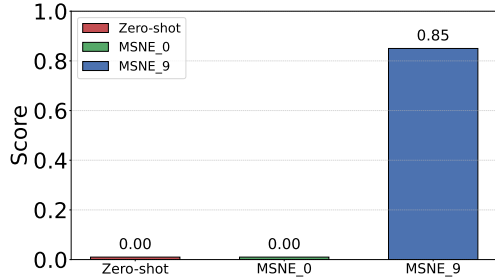
How do weaker language models perform compared to the latest, state-of-the-art models? We repeat the co-evolution procedure using GPT-4.1 as a weaker generative model, keeping prompts and evaluation settings the same. Figure 8 reports results for a single run, comparing MSNE and UED-Greedy across domains. Although GPT-4.1 produces simpler environments and weaker policies, MSNE consistently outperforms UED-Greedy and remains robust over the environment archive.

6 Conclusions

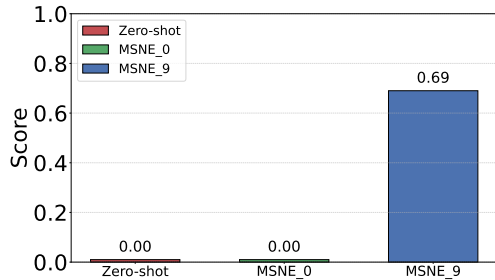
We introduced COvolve, a framework in which large language models generate environments and policies in a closed loop. The interaction between environment design and policy design is formulated as a two-player zero-sum game, and learning is performed



(a) MiniGrid



(b) PyGame

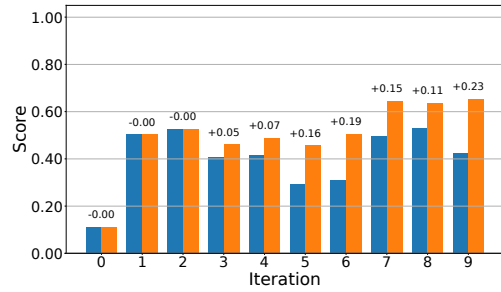


(c) CARLA

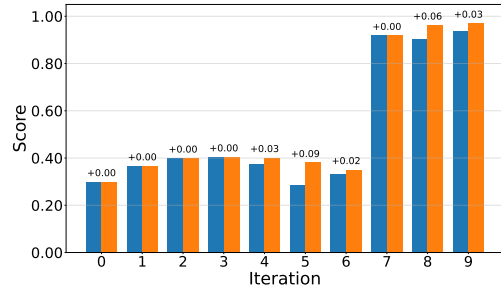
Figure 7: Result on curriculum learning. Direct training on the hardest environment (“Zero-shot”) fails, while co-evolutionary MSNE with curriculum (*right bars*) yields non-trivial performance.

over growing populations of environments and policies. Solving for a mixed-strategy Nash equilibrium yields a meta-policy that optimizes worst-case performance across the empirical set of generated environments and provides a population-level objective for continual adaptation. Experiments in urban driving, symbolic maze solving, and geometric navigation show that C0volve produces an emergent curriculum with environments that exhibit increasing structural complexity. These results demonstrate that a game-theoretic formulation of policy and environment generation enables robust co-evolution and automated curriculum construction without predefined task distributions.

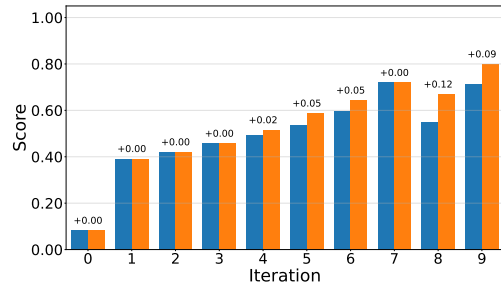
Limitations and Future Work. Unconstrained LLM-generated environments can be infeasible. To prevent this, the environment designer is restricted to predefined helper functions that ensure feasibility in each domain, and environments that violate these constraints are never sampled or evaluated. Details are provided



(a) MiniGrid



(b) PyGame



(c) CARLA

Figure 8: Ablation with a weaker generative model (GPT-4.1). Comparison of UED-Greedy and C0volve (MSNE) across domains. Although a weaker generation reduces overall performance, MSNE consistently mitigates forgetting and maintains robustness over the environment archive.

in Appendix B. Future work could pursue more principled control over environment difficulty, for example, by incorporating minimax regret [9] and providing formal guarantees on curriculum progression rather than relying on domain-specific heuristics. Another direction is to strengthen diversity checks during environment generation.

Acknowledgments

This work is supported by Knut and Alice Wallenberg Foundation via the Wallenberg AI Autonomous Sensors Systems and the Wallenberg Scholars Grant.

References

- [1] Yoram Bachrach, Edan Toledo, Karen Hambardzumyan, Despoina Magka, Martin Josifoski, Minqi Jiang, Jakob Foerster, Roberta Raileanu, Tatiana Shavrina, Nicola Cancedda, Avraham Ruderman, Katie Millican, Andrei Lupu, and Rishi Hazra. 2025. Combining Code Generating Large Language Models and Self-Play to Iteratively Refine Strategies in Games. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence Organization*, 10999–11003. doi:10.24963/ijcai.2025/1249 Demo Track.
- [2] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems* 31 (2018).
- [3] Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Self-Questioning Language Models. arXiv:2508.03682 [cs.LG] <https://arxiv.org/abs/2508.03682>
- [4] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigridd & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems* 36 (2023), 73383–73394.
- [5] Jeff Clune. 2020. AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv:1905.10985 [cs.AI] <https://arxiv.org/abs/1905.10985>
- [6] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. 2020. Leveraging procedural generation to benchmark reinforcement learning (*ICML'20*). *JMLR.org*, Article 191, 9 pages.
- [7] Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. 2017. Distributional Reinforcement Learning with Quantile Regression. arXiv:1710.10044 [cs.AI] <https://arxiv.org/abs/1710.10044>
- [8] Nicola Dainese, Matteo Merler, Minttu Alakuijala, and Pekka Marttinen. 2024. Generating Code World Models with Large Language Models Guided by Monte Carlo Tree Search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=9SpWvX9ykp>
- [9] Michael Dennis, Natasha Jaques, Eugene Viniitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. 2020. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in neural information processing systems*.
- [10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16.
- [11] Kevin Ellis, Lionel Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lore Anaya Pozo, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. 2023. DreamCoder: growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *Philosophical Transactions of the Royal Society A* 381, 2251 (2023), 20220050.
- [12] Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. 2025. OMNI-EPIC: Open-endedness via Models of human Notions of Interestingness with Environments Programmed in Code. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=Y1XkzMjPd>
- [13] John Forrest and Ted Ralphs. 2005. CBC: COIN-OR Branch and Cut Solver. <https://github.com/coin-or/Cbc>. Version accessed: 2024.
- [14] Léo François Dal Piccol Sotto, Paul Kaufmann, Timothy Atkinson, Roman Kalkreuth, and Márcio Porto Basgalupp. 2021. Graph representations in genetic programming. *Genetic Programming and Evolvable Machines* 22, 4 (2021), 607–636.
- [15] Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P Adams, and Sergey Levine. 2021. Why generalization in rl is difficult: Epistemic pomdps and implicit partial observability. *Advances in neural information processing systems* 34 (2021), 25502–25515.
- [16] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290 [cs.LG] <https://arxiv.org/abs/1801.01290>
- [17] Rishi Hazra, Alkis Sygkounas, Andreas Persson, Amy Loutfi, and Pedro Zuidberg Dos Martires. 2025. REvolve: Reward Evolution with Large Language Models using Human Feedback. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=cJPuL8mOw>
- [18] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-Zero: Self-Evolving Reasoning LLM from Zero Data. arXiv:2508.05004 [cs.LG] <https://arxiv.org/abs/2508.05004>
- [19] Jeevana Priya Inala, Osbert Bastani, Zenna Tavares, and Armando Solar-Lezama. 2020. Synthesizing programmatic policies that inductively generalize. In *8th International Conference on Learning Representations*.
- [20] Nick Jakobi. 1997. Evolutionary robotics and the radical envelope-of-noise hypothesis. *Adaptive behavior* 6, 2 (1997), 325–368.
- [21] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. 2021. Prioritized level replay. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [22] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2023. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning. *J. Artif. Int. Res.* 76 (May 2023), 64 pages. doi:10.1613/jair.1.14174
- [23] Ezgi Korkmaz. 2024. A Survey Analyzing Generalization in Deep Reinforcement Learning. arXiv:2401.02349 [cs.LG] <https://arxiv.org/abs/2401.02349>
- [24] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3323fe11e9595c09af38fe67567a9394-Paper.pdf
- [25] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9493–9500.
- [26] William Liang, Sam Wang, Hung-Ju Wang, Osbert Bastani, Dinesh Jayaraman, and Yecheng Jason Ma. 2024. Environment Curriculum Generation via Large Language Models. In *8th Annual Conference on Robot Learning*. <https://openreview.net/forum?id=F0rWEID2gb>
- [27] Zi Lin, Sheng Shen, Jingbo Shang, Jason Weston, and Yixin Nie. 2025. Learning to Solve and Verify: A Self-Play Framework for Code and Test Generation. arXiv:2502.14948 [cs.SE] <https://arxiv.org/abs/2502.14948>
- [28] Shaoteng Liu, Haoqi Yuan, Minda Hu, Yanwei Li, Yukang Chen, Shu Liu, Zongqing Lu, and Jiaya Jia. 2024. RL-GPT: Integrating Reinforcement Learning and Code-as-policy. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=LEzx6QKRH>
- [29] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Eureka: Human-Level Reward Design via Coding Large Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=IEduRUO5F>
- [30] Stuart Mitchell, Michael O’Sullivan, and Iain Dunning. 2011. Pulp: a linear programming toolkit for python. *The University of Auckland, Auckland, New Zealand* 65 (2011), 25.
- [31] Jun Morimoto and Kenji Doya. 2005. Robust reinforcement learning. *Neural computation* 17, 2 (2005), 335–359.
- [32] John Nash. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 2 (1951), 286–295. <http://www.jstor.org/stable/1969529>
- [33] OpenAI. 2025. Introducing GPT-5.2. <https://openai.com/index/introducing-gpt-5-2/>
- [34] Martin J Osborne and Ariel Rubinstein. 1994. *A course in game theory*. MIT press.
- [35] Lrel Pinto, James Davidson, and Abhinav Gupta. 2017. Supervision via competition: Robot adversaries for learning tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1601–1608.
- [36] PyGame Community. 2000–2024. PyGame: Python Game Development. <https://www.pygame.org/>. Accessed: 2025-05-09.
- [37] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [38] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8.
- [39] Mikayel Samvelyan, Akbir Khan, Michael D Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Nicolaus Foerster, Roberta Raileanu, and Tim Rocktäschel. 2023. MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=sKWIRDzPfd>
- [40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] <https://arxiv.org/abs/1707.06347>
- [41] David Silver and Richard S. Sutton. 2025. Welcome to the Era of Experience. In *Designing an Intelligence*. MIT Press.
- [42] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In *Workshop on Language and Robotics at CoRL 2022*. https://openreview.net/forum?id=3K4-U_5cRw
- [43] Hao Tang, Darren Key, and Kevin Ellis. 2024. WorldCoder, a Model-Based LLM Agent: Building World Models by Writing Code and Interacting with the Environment. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 70148–70212. https://proceedings.neurips.cc/paper_files/paper/2024/file/820c61a0cd419163ccb2c33b268816e-Paper-Conference.pdf
- [44] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on*

- Intelligent Robots and Systems (IROS)*. 23–30.
- [45] Dweep Trivedi, Jesse Zhang, Shao-Hua Sun, and Joseph J Lim. 2021. Learning to synthesize programs as interpretable and generalizable policies. *Advances in neural information processing systems* 34 (2021), 25146–25163.
 - [46] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2018. Programmatically interpretable reinforcement learning. In *International conference on machine learning*. PMLR, 5045–5054.
 - [47] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325 [cs.LG] <https://arxiv.org/abs/2211.04325>
 - [48] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=ehfRiF0R3a>
 - [49] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2024. GenSim: Generating Robotic Simulation Tasks via Large Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=OI3RoHoWAN>
 - [50] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. 2019. Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. arXiv:1901.01753 [cs.NE] <https://arxiv.org/abs/1901.01753>
 - [51] Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. 2025. Co-Evolving LLM Coder and Unit Tester via Reinforcement Learning. arXiv:2506.03136 [cs.CL] <https://arxiv.org/abs/2506.03136>
 - [52] Yao Wen Yang, Jian Feng Xu, and Chee Kiong Soh. 2006. An evolutionary programming algorithm for continuous global optimization. *European journal of operational research* 168, 2 (2006), 354–369.

Appendix

The appendix is organized as follows. Section A details the co-evolution algorithm and Nash distribution computation. Section B provides environment implementation specifics for the used simulators (i.e., MiniGrid, PyGame, and CARLA). Section C lists the exact prompts used for environment and policy generation. Section D reports additional experimental results, including generalization and reinforcement learning baselines. Section E presents the best-performing policies per domain. Finally, Section F illustrates examples of evolved environments and their mutation progress.

A Algorithmic Details

Co-Evolution Loop. At each generation t , the system performs a mutation-based search to synthesize a new environment θ_t and a corresponding policy π_t . Environment mutation generates K candidate environments by perturbing the previous one, θ_{t-1} . Each candidate is evaluated under the Nash distribution w^{t-1} over policies from earlier iterations, and the candidate with the lowest expected return is selected. Policy mutation is initialized from the highest-weighted policy π^{best} under w^{t-1} . The policy LLM generates K mutated versions of this base policy, which are evaluated solely on θ_t , and the best-performing policy is selected. The pair (θ_t, π_t) is added to the archive, and the payoff matrix $M \in [0, 1]^{t \times t}$ is updated accordingly.

Nash Distribution Computation. To determine the current policy mixture, we solve a two-player zero-sum game defined by the empirical payoff matrix M . We compute the Nash equilibrium policy distribution by solving the dual linear program of a two-player zero-sum game over the empirical payoff matrix M . The optimization is formulated using PuLP [30] and solved using the CBC backend solver [13]. The solution yields a probability distribution over policies that minimizes the worst-case environment return:

$$w^t = \arg \max_{w \in \Delta^t} \min_i \sum_j w_j M_{ij}.$$

B Environment Details

At generation 0, all environments are provided with heuristics to ensure solvability for the specific task. For the environment to be generated, at least one solution is required.

B.1 MiniGrid Implementation Details

The MiniGrid environment represents a 2D grid-world where each cell encodes the presence of objects, walls, keys, doors, and other entities. The environment supports flexible configurations of size, object placement, and symbolic dependencies, making it suitable for general planning tasks.

Action Space. The agent interacts with the environment using a discrete action space of six primitive actions:

- TURN_LEFT (0): Rotate the agent 90° counterclockwise.
- TURN_RIGHT (1): Rotate the agent 90° clockwise.
- MOVE_FORWARD (2): Advance one tile forward, if the path is free.
- PICK_UP (3): Pick up an object in front, used for collecting keys.

- DROP (4): Drop the currently carried object onto the tile in front.
- TOGGLE (5): Interact with doors in front of the agent:
 - Open a closed door (STATE = 1).
 - Unlock a locked door (STATE = 2) if carrying the correct key.

Tile Encoding. Each grid tile is encoded as a 3-tuple of integers:

$$(\text{OBJECT_IDX}, \text{COLOR_IDX}, \text{STATE})$$

This structured representation is provided in a fully observable grid array. The indexing is spatial, with (x, y) referring to grid row and column, respectively.

Table 2: MiniGrid OBJECT_IDX Mappings

0	Unseen
1	Empty
2	Wall
3	Floor
4	Door
5	Key
6	Ball
7	Box
8	Goal
9	Lava
10	Agent

Table 3: Door State Field

0	Open (passable)
1	Closed (toggle to open)
2	Locked (requires key to unlock and toggle)

Environment Logic. Doors and keys are linked by color indices, with up to six distinct colors available. Locked doors block the agent’s path until the corresponding key is acquired. The environment enforces procedural placement constraints, ensuring at least one feasible path exists through BFS-based solvability checks. Walls and other obstacles further complicate navigation. The agent maintains a single-key capacity, necessitating key management and path re-planning in multi-door configurations.

Observations. At each timestep, the agent receives a fully observable grid state represented as a flattened tensor of shape $(\text{grid_size} \times \text{grid_size} \times 3)$, normalized to $[0, 1]$. Each tile encodes the object type, color index, and dynamic state (e.g., door status) as defined by the environment’s tile encoding scheme. In addition, the policy receives the agent’s absolute position (`agent_pos`) and current orientation (`agent_dir`), enabling precise spatial reasoning and orientation-dependent actions. This structured input enables policies to perform symbolic reasoning without perceptual ambiguity, allowing them to focus solely on decision-making and planning.

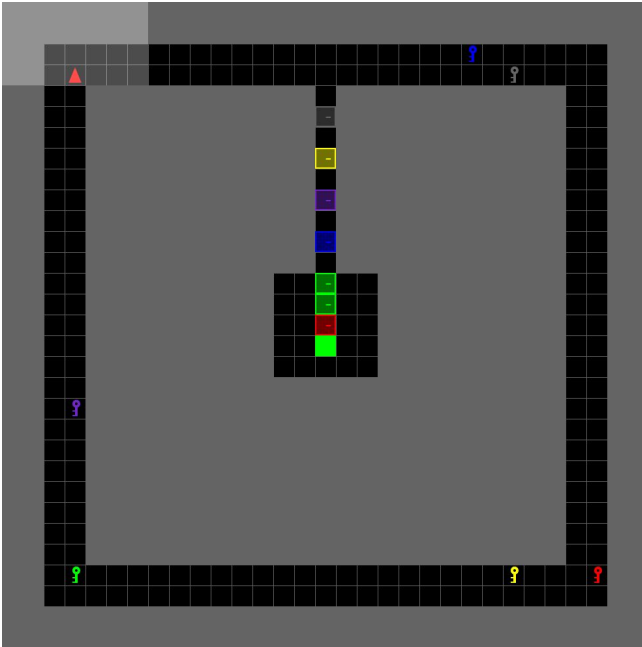


Figure 9: Example of a generated MiniGrid environment (cf. Fig. 3). For this environment, the agent (red arrow) must reach the green goal tile by unlocking the intermediate colored doors using the corresponding keys.

B.2 PyGame Implementation Details

In the PyGame environment, each instance defines a bounded 2D plane in pixel space, with task-specific width and height parameters. The agent is modeled as a circular body with a fixed physical radius of 15 pixels, while the goal zone is a rectangular target area guaranteed to fully contain the agent’s circle upon successful completion.

Obstacles are axis-aligned rectangles with randomly positioned and sized dimensions. Their placement follows strict feasibility constraints:

- Obstacles must not overlap with the goal zone.
- Obstacles must not overlap with each other.
- New obstacles are placed only if their inflated bounding box (expanded by the agent’s radius) does not intersect existing obstacles, ensuring local non-overlap and feasible placement.

Action and Observation Spaces. The agent selects a continuous 2D velocity vector $[dx, dy] \in [-1.0, 1.0]^2$ at each timestep. This vector is scaled by an environment-defined speed factor to determine pixel-wise displacement. Collision detection is performed for each proposed movement; invalid moves that would result in obstacle penetration or leaving environment bounds are rejected, leaving the agent stationary.

Observations are provided as a structured dictionary containing:

- `agent_pos`: The agent’s center coordinates in pixels.
- `objects`: A list describing the goal zone and each obstacle, with entries specifying type, pos, size, and (for the goal zone) purpose.
- `step_count`: The current timestep within the episode.

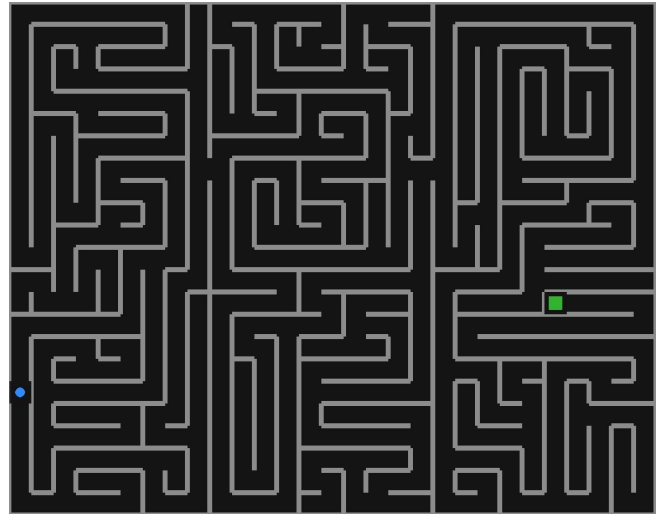


Figure 10: Example of a generated PyGame environment (cf. Fig. 4). In this environment, the agent (blue circle) must navigate the environment (spatially) to reach the goal (green rectangle).

Task Parameters. Task difficulty is progressively scaled by modifying environment parameters, including:

- The number of obstacles, increasing clutter, and requiring more deliberate path planning.
- The environment’s width and height, expanding navigation complexity.
- The agent’s movement speed, reducing maneuverability.
- The minimum agent-goal start distance, forcing longer traversal paths.

These parameters are dynamically adjusted by the environment generator to produce increasingly challenging, yet solvable, task instances.

Episode Termination and Reward. An episode terminates when the agent’s circular body is entirely within the goal zone or when the maximum allowed steps are exhausted.

Feasibility Guarantees. To ensure the agent can navigate to the goal, the environment performs a reachability check using a discretized occupancy grid that inflates obstacle regions by the agent’s radius. This guarantees that all generated tasks are physically feasible for the agent to complete. Invalid placements of obstacles or agent start positions are rejected during generation. This process ensures that every evaluation involves meaningful, solvable navigation challenges with non-trivial spatial reasoning requirements.

B.3 CARLA Implementation Details

Simulator and Map. We use CARLA Town01 in synchronous mode with a fixed time step. The route is a recorded closed polyline. The roadway is a two-way single carriageway: one lane per direction, each ≈ 4 m wide (total ≈ 8 m). Each episode spawns the ego at a fixed start; non-ego vehicles and pedestrians are randomized.



Figure 11: Example of a generated CARLA environment (cf. Fig. 5). From the ego viewpoint, the car perceives pedestrians, traffic lights, and other vehicles as described in Appendix B.3. The red car to the right of the ego vehicle (Tesla v3) was intentionally placed (from the LLM) to confuse the policy.

Episodes terminate on collision, red-light violation, timeout, or loop completion.

Frenet Geometry and Progress. Let the route be a looped polyline $\{P_i\}_{i=1}^N$. For a world point $p \in \mathbb{R}^2$, we project onto each segment

$$t_i = \text{clip}\left(\frac{(p - P_i)^\top (P_{i+1} - P_i)}{\|P_{i+1} - P_i\|_2^2}, 0, 1\right), \quad \hat{p}_i = P_i + t_i(P_{i+1} - P_i).$$

Let $k = \arg \min_i \|p - \hat{p}_i\|_2$, segment length $\ell_k = \|P_{k+1} - P_k\|_2$, and cumulative arclength up to segment k be s_k . We define arclength and lateral offset:

$$s(p) = s_k + t_k \ell_k, \quad \ell_\perp(p) = (p - \hat{p}_k)^\top \mathbf{n}_k, \quad \mathbf{n}_k = \frac{1}{\ell_k} \begin{bmatrix} -d_{k,y} \\ d_{k,x} \end{bmatrix}.$$

Progress from the episode start s_0 wraps on the loop: $\Delta s = (s(p) - s_0) \bmod L$, with loop length L . We express relative positions/velocities in the ego frame via $R_{\text{we}}(\psi) = \begin{bmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{bmatrix}$. The yaw error is $\Delta \psi = ((\psi - \psi_{\text{path}} + \pi) \bmod 2\pi) - \pi$.

Observation Space. We expose only the features the policy needs to drive on a prescribed path while interacting with traffic and pedestrians:

Ego kinematics. Speed `speed_mps`, yaw rate `yaw_rate_rps`, lateral error ℓ_\perp , yaw error $\Delta \psi$. Short histories (length 4) for {speed, lateral error, yaw error, past steer/throttle/brake} stabilize control.

Traffic light. Nearest traffic light ahead on the route: `exists`, `dist_m`, `state` \in {Red, Green, Yellow}. For simplicity, Yellow is treated as Red.

Vehicles. We keep a small, ordered snapshot (top-2) per class with ego-frame gaps and simple surrogates:

$$\text{THW} = \frac{g_x}{\max(0.5, v_{\text{ego}})}, \quad \text{TTC} = \begin{cases} g_x / (-\Delta v_x), & \Delta v_x < 0 \\ \text{null}, & \text{otherwise.} \end{cases}$$

Lead cars are those with Frenet lateral $|\ell_\perp| \leq 2$ m (ego lane). *Opposite cars* fall in $-6 \leq \ell_\perp < -2$ m (oncoming lane).

Pedestrians. Within a forward window along the route, we classify: (i) *in-lane* if $|g_y| \leq 2$ m; (ii) *approaching* if $2 < |g_y| \leq 3$ m and

moving toward the lane ($\Delta v_y g_y < 0$). For approaching walkers, we estimate time-to-enter the near lane edge, $t_{\text{enter}} = (y_\star - g_y) / \Delta v_y$ when $|\Delta v_y|$ is non-negligible, where $y_\star = \pm 2$ m.

Action Space. A continuous 3-vector (steer, throttle, brake) with `steer` $\in [-1, 1]$, `throttle` $\in [0, 1]$, `brake` $\in [0, 1]$.

Notes. This design yields a small, interpretable state while covering path tracking (via $\ell_\perp, \Delta \psi$), car-following and oncoming interactions (via THW/TTC and lane bands), signal compliance (traffic-light snapshot), and pedestrian crossing risk (in-lane vs. approaching with t_{enter}). All constants and implementation details (e.g., horizons, smoothing) are provided in our code release.

C Prompts

We provide the exact prompts used for environment and policy generation in each domain. These are instantiated dynamically at each iteration, reflecting task-specific parameters and environment configurations.

C.1 Environment Generation Prompts

Box 1: MiniGrid Environment Prompt

```

GOAL
Minimize the scalar "Actual Score" in [0,1] evaluated on the Nash-weighted policy
mix:
{Weights}
{Policies}

You will return a SINGLE Python class that replaces the existing:
class CustomEnv(MiniGridEnv):

REQUIREMENTS (MANDATORY)
1) Class:
- Keep "class CustomEnv(MiniGridEnv):" and its public API exactly.
- Do not modify the base class or inheritance.

2) - If you add helpers, define them inside the same file.
- Do not rely on undefined globals or external dependencies.

3) Fixed knobs in __init__
- Set once (and only here):
self.size = {Size}
self.num_doors = min((NumDoors), 6)
- "_gen_grid()" must use these fixed values directly (no dynamic rescaling).

4) Structured placement
- Place perimeter, rooms/corridors, doors, keys, and goal via explicit logic.
- Do not place objects blindly or randomly.

5) Solvability check
- After placements, call "check_solvable(self, start, goal)" exactly as provided.
- Accept the layout only if it returns True.

6) Episode diversity
- "_gen_grid()" must generate different layouts across episodes (vary partitions,
door indices, key pockets) while using the fixed knobs.

7) Termination
- Episode ends when (a) the agent reaches the goal, or (b) max_steps is exceeded.

8) Retry policy
- If unsolvable, retry up to 1000 times.
- If still unsolvable, set "self.failure_feedback" and return.

FEASIBILITY FUNCTIONS (DO NOT MODIFY)
- "_verify_mandatory_door_keys(self)"
- "bfs_ignore_doors(self, start, goal)"
- "bfs_block_locked(self, start, goal)"
- "_find_key_spot_block_locked(self, agent_pos, door_pos, unlocked_colors)"
Use these as-is. You may add helpers, but never replace or alter them.

OUTPUT
Return ONLY the updated "CustomEnv" class (no commentary).
    
```

Box 2: PyGame Environment Prompt

```

GOAL
Minimize the scalar "Actual Score" in [0,1] evaluated on the Nash-weighted policy
mix:
{Weights}
{Policies}

REQUIREMENTS (MANDATORY)
1) Class/API
- Keep "class CustomEnv:" and its public API exactly.
- Do not change the class name or inheritance.

2) Implement only these methods:
- reset(self)
- step(self, action)
- draw_objects(self)
- task_description(self)
- _get_obs(self)
- render(self)
- Any private helpers defined inside the class (e.g., _handle_quit, _sample_pos).

3) task_description(self)
- Must return a plain string describing:
- The task objective (agent must reach the goal zone).
- The action space (continuous [dx,dy] in [-1.0,1.0]^2).
- Key parameters (sizes, margins, speeds).
- The full observation dictionary structure.

4) Episode termination
- Ends if agent reaches the goal (checked externally by _check_done()).
- Ends if max_steps is exceeded.
- Do not call _check_done() inside step().

5) Structured placement
- Place agent, goal, and obstacles via explicit rules.
- Ensure no overlaps; keep all objects inside bounds.
- Guarantee solvability (always at least one valid path).

6) Randomization
- Use structured randomness (np.random.randint, np.random.uniform) in reset().
- Every reset() must produce a distinct environment instance.
- Randomness must contribute to meaningful diversity.

7) Safety
- The goal zone must be large enough: width,height >= 2*agent_radius + margin.

8) Behavior
- step(action): interpret action as 2D continuous move.
- Call self._handle_quit() at the start to process quit events.
- Return (obs, reward, done) where obs=self._get_obs(), reward=0.0, done=self.done.

9) Observations
- _get_obs() must return:
{
  "agent_pos": [x,y],
  "agent_radius": r,
  "objects": [
    { "type":"zone", "pos":[cx,cy], "size":[w,h], "purpose":"goal" },
    { "type":"obstacle", "shape":"rect", "pos":[cx,cy], "size":[w,h] },
    ...
  ],
  "bounds": [W,H],
  "step_count": N,
  "max_steps": M
}

10) Rendering
- draw_objects(): use pygame primitives; only draw if self.render_mode==True.
- render(): create/update PyGame surface, call draw_objects(), flip buffers if
render_mode==True.

TASK EVOLUTION
- Increase distance between agent and goal.
- Add obstacles or tighter passages.
- Increase W,H and proportionally increase max_steps.
- You are free and motivated to introduce new difficulties as long as the task
remains the same: The agent must reach the goal zone.

CONSTRAINTS
- Do NOT add symbolic puzzles (no keys, doors, colors).
- Do NOT use MiniGrid tile logic.
- Do NOT add irrelevant randomness.
- Do NOT remove or rename required methods.
- Do NOT alter the external _check_done().

OUTPUT
Return ONLY the updated "CustomEnv" class (no commentary).
    
```

Box 3: CARLA Environment Prompt

```
GOAL
Minimize the scalar "Actual Score" in [0,1] evaluated on the Nash-weighted policy
mix:
{Weights}
{Policies}
Current environment performance: {ActualScore}

You will return a SINGLE Python class that replaces the existing:
{Actual_Class}
MUTATION CONSTRAINT
- Apply exactly ONE structural change to the class.
- A structural change is defined as:
  (i) adding or removing a single environment component or behavior, OR
  (ii) (ii) modifying the logic of a single existing component (within a single
  method).
- All other code must remain semantically identical.

REQUIREMENTS (MANDATORY)
1) Class
  - Keep the class name "CarlaTown01Env" exactly.
  - Preserve all existing methods; you may add new helpers inside the class.

2) Task identity
  - Ego vehicle is always vehicle.tesla.model3.
  - Start and goal follow the same Town01 loop.
  - Do not spawn oversized vehicles (buses, trucks) or actors that may block
  solvability.

3) task_description(self)
  - Must return a plain string describing:
    - The driving objective (complete the loop without collisions).
    - Key environment elements (traffic, pedestrians, lights, dynamic behaviors).
    - The observation dictionary fields provided to the policy.
  - If new fields are added to obs (via get_obs), they must be explicitly
  documented in this string.
  - Do not remove fields; only extend if needed.

4) Observations
  - get_obs() must remain consistent with the description.
  - The observation dictionary includes ego state, histories, traffic lights, lead
  cars, opposite cars, pedestrians.
  - New factors (e.g., jaywalkers, lane changes) must be added carefully and
  described in task_description.

5) Solvability and safety
  - Always ensure at least one feasible driving strategy exists.
  - Pedestrian side-hit guard must remain intact.
  - Adjust max_steps proportionally if difficulty increases.
  - No actor may force unsolvable collisions.

6) Constraints
  - Do not add global code or side effects outside the class.
  - Do not remove feasibility checks already in place.
  - Do not change the class name.

OUTPUT
Return ONLY the updated CarlaTown01Env class (no commentary).
```

C.2 Policy Generation Prompts

Box 4: PyGame Policy Prompt

```
GOAL
Maximize the scalar "Actual Score" in [0,1] for the current policy by improving the
given function.

YOU MUST RETURN
A SINGLE Python function that replaces the existing one:
def policy(obs): # -> [dx, dy] in [-1.0, 1.0]^2
MUTATION CONSTRAINT
- Apply exactly ONE structural change to the class.
- A structural change is defined as:
  (i) adding or removing a single environment component or behavior, OR
  (ii) (ii) modifying the logic of a single existing component (within a single
  method).
- All other code must remain semantically identical.

INPUTS
- Actual Score = {ActualScore}
- Given Policy = {Policy}
- Observation dictionary schema (exact field names and meaning) = {obs_dict}

FUNCTION CONTRACT
- Keep the exact signature: def policy(obs).
- Return a 2D continuous action [dx, dy] with each component in [-1.0, 1.0].
- Do not use randomness, globals, I/O, or external libs beyond numpy.

OBSERVATION DICTIONARY
- Use only fields provided in obs. The runner supplies {obs_dict}. The typical
  structure is:
  - agent_pos: [x, y] current agent center position in pixels
  - agent_radius: r agent circle radius in pixels
  - objects: list of dicts describing scene items. Each item:
    - type: "zone" or "obstacle"
    - pos: [cx, cy] center position
    - size: [w, h] rectangle width, height
    - purpose: optional string, e.g., "goal" for the goal zone
  - bounds: [W, H] canvas width and height in pixels
  - step_count: N steps elapsed
  - max_steps: M step limit before timeout

STRICT REQUIREMENTS
1) Goal-directed motion
  - Move toward the goal zone whenever a safe direction exists.
  - Prefer progress that reduces distance-to-goal.

2) Obstacle and bounds safety
  - Do not output actions that would cause collision with any obstacle rectangle.
  - Do not move outside [0, W] x [0, H].
  - Respect agent_radius clearance when deciding direction.

3) Action validity
  - Clamp or normalize output so each component stays within [-1.0, 1.0].
  - Avoid jitter or oscillation near the goal.

4) Robustness
  - If the direct route is blocked, choose a safe detour around obstacles.
  - Avoid loops by preferring actions that reduce distance-to-goal over time.

EDGE CASES
- Goal directly reachable: head straight toward the goal center or safe entry edge.
- Narrow passage: align with passage axis and pass through without scraping
  boundaries.
- Stuck against obstacle: choose an alternate heading that increases free-space
  margin.
- Near goal zone edge: reduce overshoot and enter the zone cleanly.

QUALITY TARGETS
- Short time-to-goal.
- Minimal wasted motion and reversals.
- Collision-free trajectories across diverse layouts.

FORBIDDEN
- Changing the function name, arguments, or return type.
- Returning values outside [-1.0, 1.0].
- Ignoring obstacles, bounds, or agent_radius in decisions.
- Using randomness, global state, file or network I/O, or non-numpy libraries.

OUTPUT
Return ONLY the improved function "def policy(obs):"
No explanations, no comments, no extra text.
```

Box 5: MiniGrid Policy Prompt

GOAL
 You are tasked **with** improving an existing policy function **for** navigating MiniGrid environments by applying macro-mutation operator.
 The policy must analyze the grid, reason about objects, plan an optimal path, **and** execute actions efficiently.
 The objective **is** to reach the goal tile (OBJECT_IDX=8).

You are provided **with**:
 - Actual Score = {ActualScore}, a scalar **in** [0,1] that reflects the performance of the given policy.
 - Policy = {Policy}, the current implementation of the policy function.

YOUR TASK
 Analyze the given policy together **with** its score **and** modify it to improve performance.
 The output must be a new version of the same function **with** improvements.

MUTATION CONSTRAINT
 - Apply exactly ONE structural change to the **class**.
 - A structural change **is** defined as:
 (i) adding **or** removing a single environment component **or** behavior, **OR**
 (ii) (ii) modifying the logic of a single existing component (within a single method).
 - All other code must remain semantically identical.

OUTPUT
 Return a single Python function:

```
def policy(obs, agent_pos, agent_dir): # -> int in {0,1,2,3,4,5}
```

ENVIRONMENT FORMAT
 - obs **is** a 2D NumPy array of shape (grid_size, grid_size, 3).
 - Each tile **is** encoded as (OBJECT_IDX, COLOR_IDX, STATE).
 - Indexing **is** (x=row, y=column).

OBJECT_IDX MAP: 0=Unseen, 1=Empty, 2=Wall, 3=Floor, 4=Door, 5=Key, 6=Ball, 7=Box, 8=Goal, 9=Lava, 10=Agent

DOOR STATE: 0=Open (free to **pass**), 1=Closed (requires Toggle action=5 when facing), 2=Locked (requires correct key + Toggle=5)

ACTIONS: 0=Turn Left, 1=Turn Right, 2=Move Forward, 3=Pick Up, 4=Drop, 5=Toggle

STRICT REQUIREMENTS
 1) Goal-Oriented Navigation
 - Always plan **and** execute a valid path to the Goal (OBJECT_IDX=8).
 - Avoid unnecessary detours unless a locked door blocks the path.

2) Door Handling
 - Open doors (STATE=0) act as free space.
 - Closed doors (STATE=1): face the door, Toggle (5) to **open**, then Move Forward (2).
 - Locked doors (STATE=2): only approach after collecting the correct key. Face the door, Toggle (5), then Move Forward (2).

3) Key Handling
 - Keys are only collected **if** required to unlock a blocking door.
 - The agent can hold exactly one key at a time.
 - If already holding a different key, Drop (4) into the front cell (**if** empty) before picking up the new one.
 - Keys must be picked up **with** Pick Up (3) when the agent **is** adjacent **and** facing the key.
 - Dropped keys must remain accessible.

4) Safety **and** Obstacles
 - Never Move Forward (2) into a Wall (2) **or** Lava (9).
 - Treat Unseen tiles (0) as blocked until explored.

5) Orientation
 - Before **any** interaction (Move Forward, Pick Up, Drop, Toggle), ensure the agent **is** facing the correct adjacent cell.
 - Rotate (0=Left, 1=Right) until aligned, then act.

6) Termination
 - The episode ends when the agent reaches the Goal **or** exceeds max_steps.
 - The policy must minimize wasted actions **and** maximize efficiency.

EDGE CASES
 - If the agent needs a key but already holds another, drop the held key before pickup.

FORBIDDEN
 - Changing the function name, arguments, **or** return type.
 - Returning values outside {0,1,2,3,4,5}.
 - Using randomness, **global** state, **or** external libraries.

OUTPUT FORMAT
 - Return only the improved function `def policy(...)` **in** valid Python.
 - No explanations, no comments, no extra text.

Box 6: CARLA Policy Prompt

GOAL
 Maximize the scalar "Actual Score" **in** [0,1] by improving the current driving policy.
 You will **return** a SINGLE Python **class** that replaces the existing:

```
{Actual_Policy}
```

MUTATION CONSTRAINT
 - Apply exactly ONE structural change to the **class**.
 - A structural change **is** defined as:
 (i) adding **or** removing a single environment component **or** behavior, **OR**
 (ii) (ii) modifying the logic of a single existing component (within a single method).
 - All other code must remain semantically identical.

INPUTS
 - Actual Score = {ActualScore}
 - Previous Policy = {Policy}
 - Path = np.ndarray (N,2) lane-center polyline

STRICT REQUIREMENTS
 1) Class/API
 - Keep the **class** name "Policy".
 - Implement `__init__(self)` **and** `compute_action(self, obs, path)`.
 - Return (steering, throttle, brake) as floats.
 - steering **in** [-1,1], throttle **in** [0,1], brake **in** [0,1].
 - If brake > 0 then throttle must equal 0.

2) Determinism **and** smoothness
 - No randomness **or** learning.
 - Ensure gradual changes, avoid jerks.

3) Robustness
 - Handle None **or** NaN conservatively.
 - On invalid input, default to safe stop (steer=0, throttle=0, brake>0).
 - No prints **or** logging.

4) Sign conventions
 - lateral_hist4: right-positive meters.
 - yaw_error_hist4: ego yaw - path yaw, right-positive.
 - yaw_rate_rps: right-positive radians/s.

OBSERVATION FORMAT
 obs **is** a dictionary containing:

- Ego state:
 - speed_mps: current speed **in** m/s
 - yaw_rate_rps: yaw rate **in** rad/s (right-positive)

- Ego histories (arrays of length 4):
 - speed_hist4: past speeds
 - lateral_hist4: lateral errors (m, right-positive)
 - yaw_error_hist4: yaw errors (rad, right-positive)
 - steer_cmd_hist4: previous steering commands
 - throttle_cmd_hist4: previous throttle commands
 - brake_cmd_hist4: previous brake commands

- Traffic light:
 - exists: boolean
 - state: **int** {0=unknown, 1=green, 2=yellow, 3=red}
 - dist_m: distance to stop line (m)

- Lead cars (up to 2, same schema each):
 - gap_long_m: longitudinal gap (m)
 - gap_lat_m: lateral gap (m)
 - rel_long_mps: relative longitudinal speed (m/s)
 - ttc_s: time-to-collision (s)
 - thw_s: time headway (s)
 - Opposite cars (up to 2, same schema as lead cars)

- Pedestrians (variable count):
 - lane: lane index
 - state: **int** encoding motion state
 - gap_long_m: longitudinal gap (m)
 - gap_lat_m: lateral gap (m)
 - rel_lat_mps: relative lateral speed (m/s)
 - t_enter_lane_s: predicted time to enter lane (s)
 - side: which side of road (left/right)
 - All dynamic actors truncated to 35 m ahead of ego

OBJECTIVES
 - Lateral: minimize lateral **and** yaw errors relative to the centerline.
 - Longitudinal: track target speed up to 6.94 m/s (25 km/h) **if** unimpeded.
 - Traffic lights: stop smoothly before stop line on red; never cross on red.
 - Pedestrians: **yield** to pedestrians **in** **or** entering ego lane.
 - Lead vehicles: maintain safe following distance; avoid indefinite blocking.
 - Precedence order: red light stop > pedestrian yielding > lead vehicle following > cruising.
 - Fail-safe: **if** uncertain, perform controlled stop.
 - Comfort: avoid abrupt oscillations; prioritize smooth steering **and** braking.

FORBIDDEN
 - Changing the **class** name **or** method signatures.
 - Returning values outside steering/throttle/brake ranges.
 - Simultaneous throttle **and** brake > 0.
 - Using randomness, logging, prints, **or** external dependencies.

OUTPUT
 Return **ONLY** the new improved **class** "Policy".

D Additional Results

D.1 Generalization Across Environments

We compare *standardized, unseen environments* to those produced during co-evolution. The standardized set comprises MiniGrid DoorKey-16x16-v0 and LockedRoom-v0, and CARLA Town02 (trained on Town01). These differ from our evolved environments in three respects: (i) *structure* (fixed layouts and goal semantics rather than co-evolved variants), (ii) *scale* (grid/world size and path lengths), and (iii) *sequential dependencies* (e.g., key-door ordering and room unlocking). For CARLA, Town02 diverges from Town01 by road-network density and traffic complexity: it has sharper turns, narrower lanes, and more intersections and pedestrian crossings, requiring longer detours and tighter maneuvers compared to the more regular Town01 layout. We evaluate with identical rollout settings.

D.2 Reinforcement Learning Results

D.2.1 MiniGrid Maze-solving. We evaluate two representative algorithms using Stable-Baselines3 [38]: PPO, a policy-gradient method, and QRDQN, a value-based method for discrete domains.

Reward shaping. For reward, we use the default MiniGrid reward function:

$$R(s, a) = \begin{cases} 1 - 0.9 \cdot \frac{t}{T_{\max}}, & \text{if the agent reaches the goal at step } t, \\ 0, & \text{otherwise,} \end{cases}$$

where T_{\max} is the maximum episode length. Thus, faster completion yields a higher return.

Algorithm	Env 0	Env 2	Env 6
PPO	12.0 ± 1.4%	0.0 ± 0.0%	0.0 ± 0.0%
QRDQN	68.5 ± 12.0%	0.0 ± 0.0%	0.0 ± 0.0%

Table 4: Success rates (% , mean ± std over two runs) across MiniGrid environments.

D.2.2 PyGame 2D Navigation. We evaluate two representative algorithms using Stable-Baselines3 [38]: PPO, and SAC.

Reward shaping. In the PyGame environments, reward is sparse with a per-step penalty:

$$R(s, a) = \begin{cases} +1, & \text{if the agent reaches the goal,} \\ -0.01, & \text{otherwise (each step).} \end{cases}$$

This encourages agents to minimize path length while ensuring sparse success feedback.

D.3 Additional seed run

E Best Performing Policies

The final evolved policies are too extensive to analyze line by line. Instead, we provide high-level summaries of their algorithmic structure and key heuristics.

Algorithm	Env 0	Env 2	Env 6
PPO	61.7 ± 2.3%	6.2 ± 0.7%	0.0 ± 0.0%
SAC	88.2 ± 3.0%	22.5 ± 5.9%	0.0 ± 0.0%

Table 5: Success rates (% , mean ± std over two runs) across PyGame environments.

E.1 Best Performing MiniGrid Policy

The best performing policy (`policy_9`, see Figure 6) is a fully model-based planning agent that formulates MiniGrid navigation as a discrete A^* search over agent position, orientation, held key color, and door-open states.

The policy operates as follows: (1) it parses the grid to identify the goal, doors, and keys, (2) computes a relaxed reachability region that ignores door and key semantics to conservatively identify which doors and key colors can lie on a valid start-goal corridor, (3) performs lexicographic-cost A^* *planning* over a factored state space with explicit door toggling, key pickup, key drop, and movement actions, and (4) executes only the first action of the optimal plan, replanning at every timestep.

To reduce the search space without sacrificing correctness, the planner reasons only about *useful doors* (doors that lie in the relaxed start-goal corridor) and *useful key colors* (keys that can open such doors). Keys are treated as blocking cells during planning, and keys are only dropped when the agent is holding one and the cell ahead is empty, ensuring safe and deterministic key management.

The A^* heuristic combines Manhattan distance to the goal with a lower bound on the number of turns required to face a direction that reduces this distance, improving guidance while preserving admissibility.

Together, these components yield a deterministic planning agent that can reliably resolve door-key dependencies, minimize unnecessary interactions, and avoid key-handling loops in complex MiniGrid environments.

E.2 PyGame Policy

The PyGame agent is implemented as a *planning-reactive navigation policy* that combines global path planning with local, feasibility-aware motion selection at every timestep.

- **Global planning:** The agent computes a global path to the goal using A^* search on a coarse occupancy grid. Obstacles are mildly inflated based on the agent radius to ensure collision-free paths. The resulting path is cached and only recomputed when the goal changes or when progress stalls.
- **Waypoint tracking with visibility lookahead:** The agent follows the planned path using waypoints, advancing when sufficiently close. If multiple upcoming waypoints are directly visible, the agent skips intermediate points and targets the farthest visible waypoint.
- **Local motion candidates:** At each step, the policy samples a set of candidate motion directions around the desired path direction, including small angular deviations and obstacle-aligned tangents when near walls.

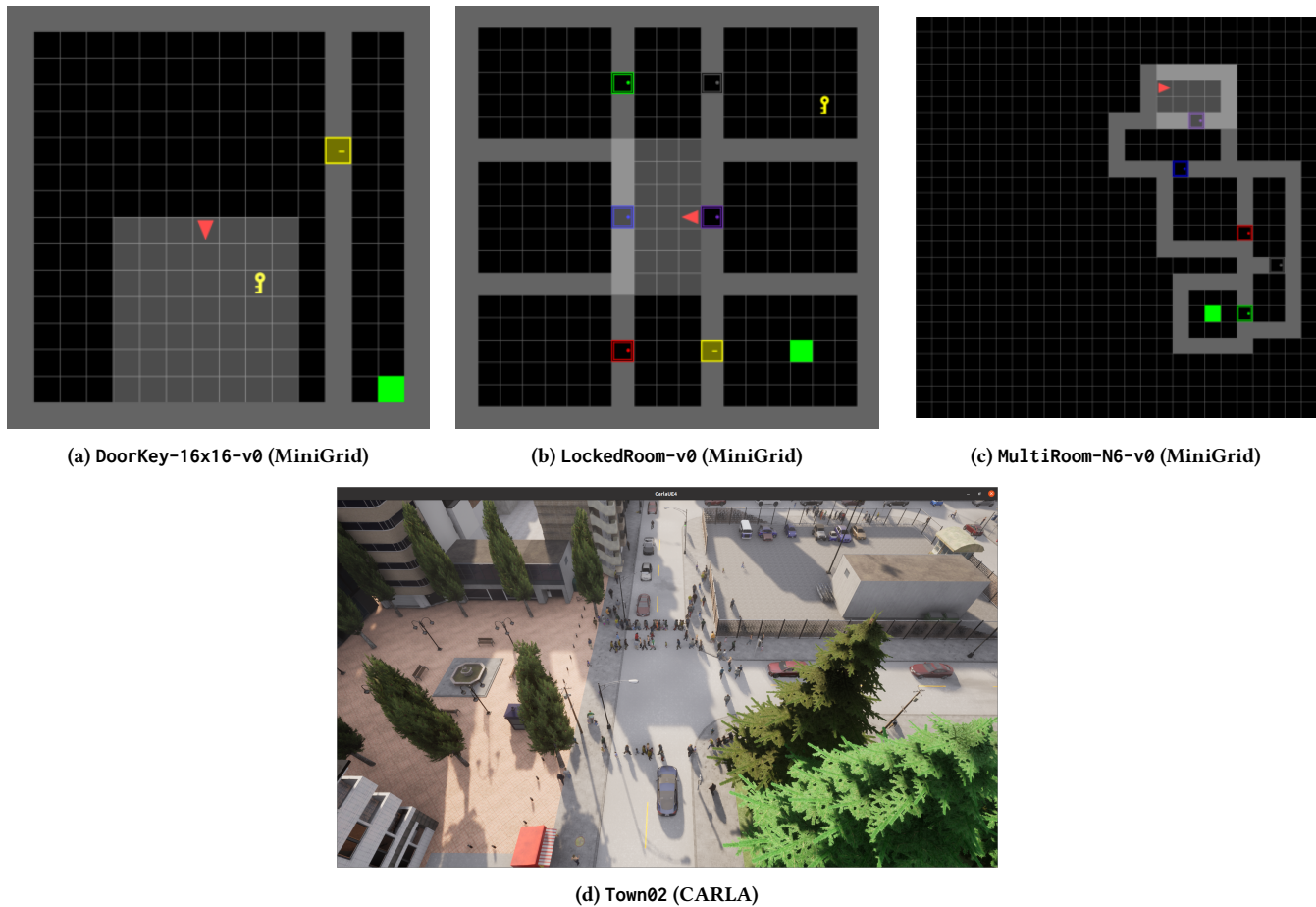


Figure 12: Examples of previously unseen standardized environments used to validate generalization. MiniGrid snapshots (*top*): DoorKey, LockedRoom, and the hardest goal-reaching benchmark ObstructedMaze-Full. CARLA Town02 (*bottom*).

- **Predictive collision checking:** Each candidate direction is validated using one-step forward collision checks that match the environment’s continuous collision model. Infeasible directions are discarded before scoring.
- **Directional scoring and selection:** Feasible candidates are scored based on path-aligned progress, goal alignment, continuity with the previous action, and local obstacle clearance. The highest-scoring direction is selected.
- **Oscillation control:** The policy maintains a short-term memory of recent motion directions and penalizes rapid directional sign changes. A turn-rate limiter further constrains angular changes between successive actions.
- **Gap-centering behavior:** When near obstacles, lateral ray probes estimate free space on either side of the agent, biasing motion toward the center of locally available free space.
- **Execution:** The final action is a normalized 2D velocity direction returned to the environment. If no feasible direction exists, the agent temporarily halts and triggers replanning.

This structure allows the agent to consistently alternate between global path guidance and locally feasible motion execution in continuous PyGame navigation environments.

E.3 Carla Policy

The best performing controller (`policy_9`, see Fig. 6) augments a smooth cruise/follow core with a clearance-aware passing routine and stricter intersection handling.

Four-stage loop. (1) **Signal gating:** strict traffic-light guard (*Yellow = Red*), stop-line latch, and pedestrian holds; approach speed is limited by both stop-line distance and queued-lead gap. (2) **Lead classification:** distinguishes a right-curb parked blocker from an in-lane stopped lead using lateral intrusion and relative speed cues. (3) **Clearance-aware pass:** if the blocker is parked, oncoming is clear, and distance gates are met, the agent enters a bounded left-offset pass. It maintains a minimum offset and a small, gated opposite-lane incursion, *holds* the offset while alongside, and only recenters after front-clearance (with a brief hold if the lead vanishes). (4) **Smooth tracking:** target-speed smoothing with curvature/heading caps

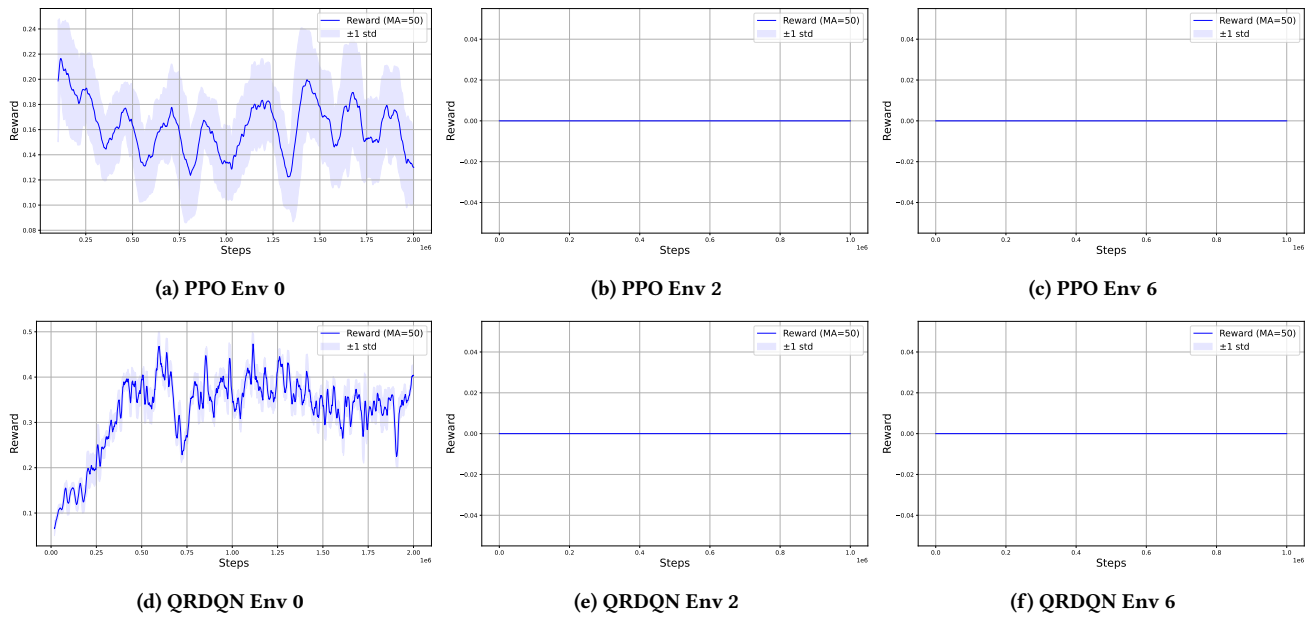


Figure 13: Training curves of PPO (*top*) and QRDQN (*bottom*) across MiniGrid environments. The y-axis represents reward, and the x-axis represents total training steps.

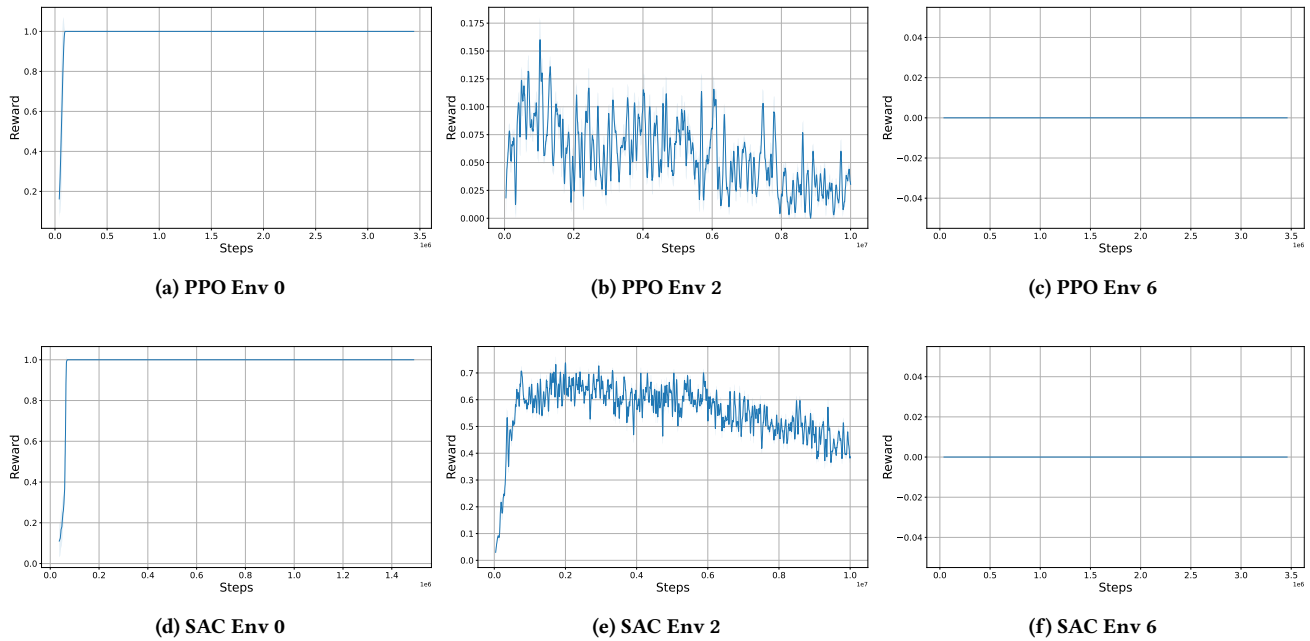


Figure 14: Training curves of PPO (*top*) and SAC (*bottom*) across PyGame environments. The y-axis represents reward, and the x-axis represents total training steps.

and a damped lookahead lateral controller; unstick logic provides a gentle creep when safe.

- **Stop-line priority:** combined stop-line/lead-gap caps and a near-line latch prevent creeping over the line on non-green states.

Key heuristics.

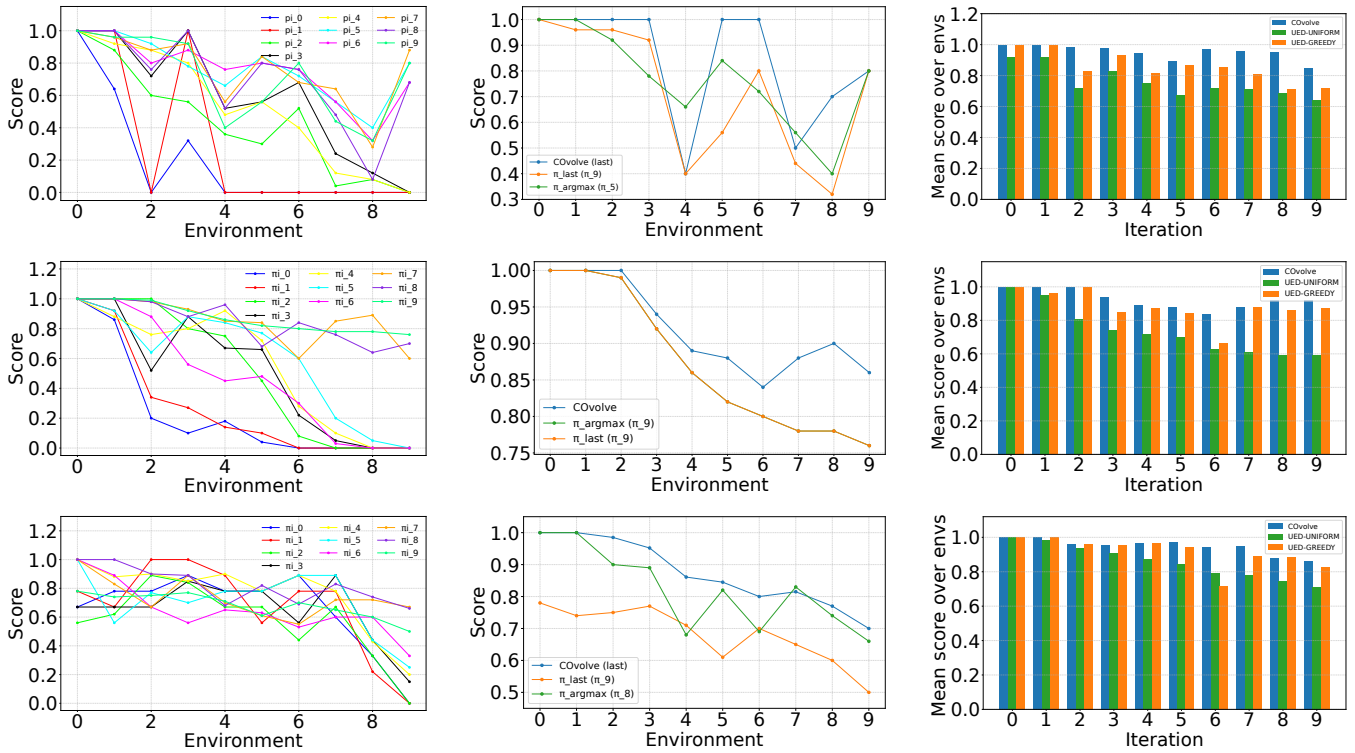


Figure 15: Performance during environment-policy co-evolution for the second seed. Left: Success rates of all discovered policies evaluated on all environments generated during evolution (policy-environment payoff matrix). Center: Comparison between the mixed-strategy Nash equilibrium (MSNE) policy mixture, the best single policy (π_{argmax}), and the latest policy (π_k), evaluated on all environments $\{\theta_0, \dots, \theta_k\}$ for Pygame, π_{argmax} coincides with π_k . Right: Mean success over environments $\{\theta_0, \dots, \theta_k\}$ for three strategies: UED-Greedy (latest policy only), UED-Uniform (uniform mixture over all policies so far), and COvolve (MSNE mixture).

- **Right-edge pass safety:** minimum pass offset, centerline guard, and oncoming no-pass gate; tiny opposite-lane incursion is permitted only when clear.
- **Stability under occlusion:** offset-hold on brief lead dropouts avoids snap-back; post-clear recenter includes a short hysteresis.

- **Hard vs. soft chokepoints:** some doors are reinforced by barrier walls that force strict bottlenecks, while others include short wings or detours that add complexity without fully blocking the corridor.
- **Protected corridors:** a one-cell halo ensures that critical key-door paths remain open even as random obstacles are added, guaranteeing solvability.

F Evolved Environments

F.1 MiniGrid Environment Mutations

In the MiniGrid maze-solving task, the LLM mutates discrete grid-worlds where an agent must navigate to a goal while avoiding obstacles, doors, and keys. Difficulty increases through the following mechanisms:

- **Grid scaling:** larger grids extend path length and increase exploration requirements.
- **Obstacle density:** additional walls create more complex mazes and reduce direct visibility of the goal.
- **Sequential key-door dependencies:** locked doors are introduced along the main corridor, requiring keys to be collected and used in the correct order.

This progression transforms initially trivial layouts into structured mazes that demand multi-step reasoning, ordered dependencies, and long-horizon planning while ensuring that every environment remains solvable by construction.

F.2 PyGame Environment Mutations

In the PyGame navigation task, the LLM mutates a continuous 2D arena where a circular agent must reach a rectangular goal zone while avoiding collisions. While early generations adjust simple parameters such as arena size or obstacle counts, later environments evolve into structured mazes with corridor-like passages and long detours. Key mutation axes include:

- **Corridor formation:** long rectangular bars are placed to partition the arena into corridors, forcing agents to identify traversable passages rather than rely on direct routes.

- **Bottleneck and detour creation:** increasing bar thickness and obstacle density narrows passageways and introduces dead ends, requiring agents to plan long, non-greedy paths.
- **Start-goal separation:** minimum distance constraints push the agent to begin far from the goal, ensuring navigation requires multiple turns and obstacle avoidance.
- **Precision termination:** the goal region remains small relative to agent size, demanding careful alignment to trigger success.
- **Scalable horizons:** enlarging arenas and increasing maximum steps allows environments to grow in complexity without becoming unsolvable.

Unlike gridworlds, these continuous PyGame arenas induce navigation behaviors closer to geometric planning: agents must balance global pathfinding with local collision checks, and later evolved environments present rich mazes with narrow corridors that mimic real-world navigation challenges.

F.3 Carla Environment Mutations

In the *Carla Town01* driving task, the LLM mutates a fixed urban loop with signalized intersections, oncoming traffic, and pedestrians. Difficulty rises from light, compliant flows to dense, heterogeneous traffic with narrow-clearance segments, while remaining solvable by construction.

- **Traffic scaling:** vehicle counts increase from light to heavy urban load; speed variance and lane changes introduce realistic flow heterogeneity.

- **Pedestrian pressure:** higher crossing rates and tighter cadences create frequent curb-to-lane interactions requiring cautious approach and yielding.
- **Intersection strictness:** virtual “second gates” beyond stop lines mirror light states, penalizing early acceleration and forcing disciplined red/yellow behavior.
- **Narrow-clearance segments:** parked or frozen intrusions create lane squeezes that demand bounded lateral offsets and precise, short opposite-lane incursions when clear.
- **Micro-perturbations:** periodic brake-taps on leads and occasional temporary stoppers test following stability without causing deadlocks.
- **Oncoming dynamics:** faster opposite-lane bursts create brief no-pass windows, requiring agents to time passes and maintain centerline guards.
- **Jam watchdog & solvability:** stall detectors inject bounded flow perturbations to unstick traffic; obstacle placements and signal logic are constrained to ensure episodes remain completable.
- **Observation compatibility:** added features (e.g., lane-squeeze indicators, extended stop-line states) are exposed via backward-compatible fields to avoid policy breakage.

This progression turns a benign city loop into a dense, signal-rich scenario with tight margins and bursty interactions, pushing policies to coordinate cautious intersection handling, safe passing, and recovery from transient jams.