HYPERPRIOR 2026-11-15

# Dream Cities: Measuring World-Model Exploitation Against an Exact Simulator

**Patrick Steele**
HyperPrior Research, London

**Optimising inside a learned world model entangles finite-sample selection luck with model exploitation. In Micropolis, a spatial-economic simulator with bit-exact replay, we benchmark planning error above finite-sample selection luck by running each optimiser identically inside a learned world model and inside an exact-simulator reference on fresh seeds. Across a 24-checkpoint model ladder that clears short-horizon fidelity gates yet drifts over our 24-month horizon, excess false acceptance appears before search and persists across budgets. Under cross-entropy search, world-model ground-truth pass rates stall between $10^3$ and $10^4$ plans while exact-simulator discovery keeps climbing, widening the absolute shortfall from 0.062 to 0.311, and both effects persist under monthly closed-loop replanning from true simulator states. In-imagination PPO policies trained for $10^6$ imagined months show parallel optimism, and perturbation-controlled patching isolates post-earthquake fire spread as a causal driver on recovery. Within this ladder (120 model-task pairs), trajectory-fidelity metrics show weak association with PPO exploitation, whereas single-plan verdict calibration on one logged plan per starting instance achieves Spearman +0.875 (+0.681 within family). Single-seed fiscal-coverage retraining cuts planning excess without raising ground-truth pass rates, while post-hoc inference defences reduce false acceptance only by rejecting many truly passing plans. Extending primary one-step training sixfold on a single seed (12 epochs) raises held-out likelihood yet worsens 24-month spatial drift, flipping miscalibration from optimistic false acceptance to pessimistic rejection.**

*Keywords: world models, learned simulators, spatial-economic simulation, model-based reinforcement learning, planning, optimiser's curse*

## 1. Introduction

Generative world models now serve both as synthetic training arenas for reinforcement learning (Alonso et al., 2024; Chen et al., 2025; Hafner et al., 2025a,b) and as zero-shot evaluators that rank plans before deployment (Gemini Robotics Team et al., 2026; Guo et al., 2026; Quevedo et al., 2025). Whenever search or policy gradients select top-scoring imagined trajectories, optimisation concentrates on regions where stochastic luck and systematic model errors align. The broad failure modes are well known. Ha and Schmidhuber (2018) showed a controller scoring 2,086 inside a learned VizDoom dream collapsing to 193 in the true game, Lambert et al. (2020) documented the mismatch between trajectory accuracy and control performance, and recent theory proves imperfect models remain exploitable over rich policy classes (Bhamidipaty et al., 2026b). What remains scarce is controlled empirical measurement that benchmarks model-based planning error above finite-sample selection luck, paired with practical pre-training audit evidence (Yu et al., 2026).

Measuring exploitation cleanly requires a baseline that standard evaluations omit. Even inside an exact stochastic simulator, selecting the best plan from finite rollouts introduces upward bias through the *optimiser's curse* (Smith and Winkler, 2006). Separating learned-model degradation from stochastic selection noise requires dynamics rich enough to challenge neural models alongside bit-exact state-and-seed replay. Each optimiser can then run identically inside both the learned model and an *exact-simulator reference* queried on fresh non-oracle seeds.

We build this testbed around Micropolis, the open-source C++ core of the 1989 urban economy simulator[1]. Unlike arcade pixel arenas or smooth rigid-body control, Micropolis couples spatial land-use rules with delayed fiscal accounting and discrete network cutoffs (Figure 10). Construction deducts immediately from city reserves. Residential, commercial, and industrial migration follows lagged census valves, and taxes and service budgets settle annually. Depth-first power routing imposes sharp grid cutoffs alongside stochastic disasters. We train a 24-checkpoint ladder of stochastic ConvNeXt U-Net world models across parameter scales, data budgets, and objectives. We then evaluate them from single logged plans and random shooting through $10^4$-plan cross-entropy search, monthly replanning, and 360 PPO runs (Figures 1 and 2).

**Summary of contributions and findings.**

[1] Micropolis is a registered trademark of Micropolis Corporation (Micropolis GmbH) used under the Micropolis Public Name License (Micropolis GmbH, 2024). We build upon MicropolisCore (SimHacker, 2026) under GPL-3.0 with additional terms (Electronic Arts Inc., 2007), with deterministic replay patches documented in our sources.

*Correspondence via hyperprior.net*
© 2026 HyperPrior Research

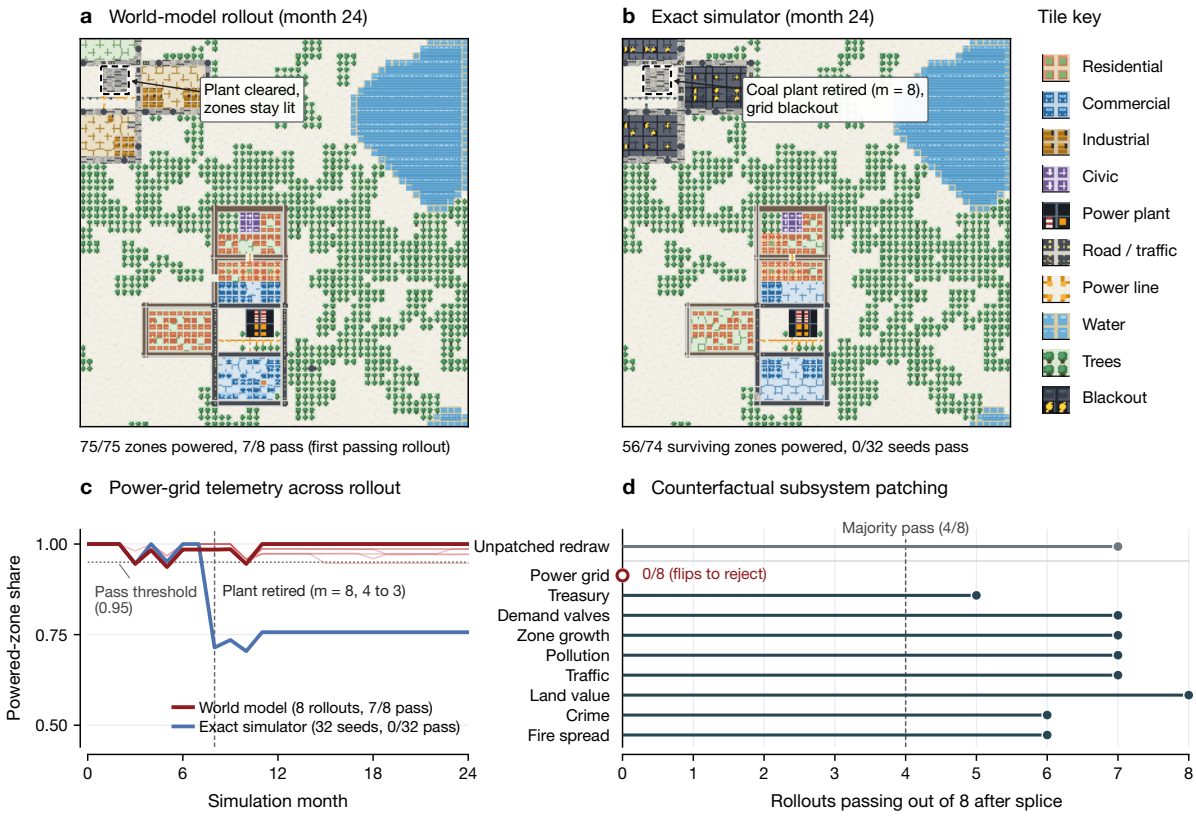


**Figure 1 | Demolishing a coal plant causes a simulator blackout while all zones stay powered in the world model.** A 24-month plan executed from the identical initial city in our primary 50M world model and the exact Micropolis engine. Both demolish a coal plant at month 8, yet only the true simulator suffers a grid blackout. **a**, Imagined month-24 city for the highest-discrepancy plan among the 29 exploited plans from $10^3$ random shooting (Appendix A.6), rendered with procedural micro-sprites; all zones stay powered and 7 of 8 replayed rollouts pass (8 of 8 in the search draw). **b**, Bit-exact C++ replay in which bulldozing a plant tile retires the plant at month 8, leaving 18 of 74 zones unpowered (dark unpowered zones; zero of 32 oracle seeds pass). **c**, Powered-zone share across 8 imagined rollouts vs. 32 engine seeds following the month-8 plant demolition (the deterministic capacity cliff yields zero cross-seed spread on the simulator trace). **d**, Counterfactual subsystem patching (8 rollouts) showing that overwriting the power grid from the engine flips the verdict to zero passes, while all eight other subsystems leave the false pass intact.

- **Short-horizon fidelity masks task-horizon drift.** Every ladder checkpoint clears one-month changed-tile F1 and 12-month scalar gates. Yet at our two-epoch training budget, the primary 50M model equals or exceeds start-state copying error by month 24 on 7 of 8 spatial overlays (Section 2).
- **Search widens ground-truth shortfall beside persistent excess false acceptance.** On the two-epoch ladder, excess false acceptance appears before search and persists across budgets. Scaling adaptive CEM from $10^1$ to $10^4$ lifts world-model ground-truth pass rates through $10^3$ before stalling between $10^3$ and $10^4$, while exact-simulator discovery keeps climbing. This widens the absolute shortfall from 0.062 to 0.311 (Section 3); both excess false acceptance and ground-truth shortfall also persist under monthly closed-loop replanning from true simulator states (Section 3.5). In-imagination PPO shows parallel optimism, while sixfold longer one-step training flips spatial rollouts to zero-acceptance pessimism on 3 of five task families without closing the $10^4$ shortfall (Section 3.6).
- **Perturbation-controlled patching isolates dynamic blind spots.** Splicing ground-truth subsystems into exploited rollouts shows raw power-grid flips do not exceed conservative random-overwrite controls. By contrast, on *recover*, splicing ground-truth fire state overturns 0.76 of exploited plans vs. 0.21 on an unpatched redraw, exceeding matched controls while overturning only 0.18 of true acceptances (Section 4).
- **Single-plan verdict calibration ranks RL exploitation within our model ladder.** Across 120 model-task pairs, short-horizon F1, 12-month error, and log-likelihood vary narrowly across gate-passing runs and show weak association with exploitation. Scoring one fixed logged plan per starting instance across the 120-city bank instead ranks downstream PPO exploitation at Spearman +0.681 within family and +0.875 pooled (Section 5).

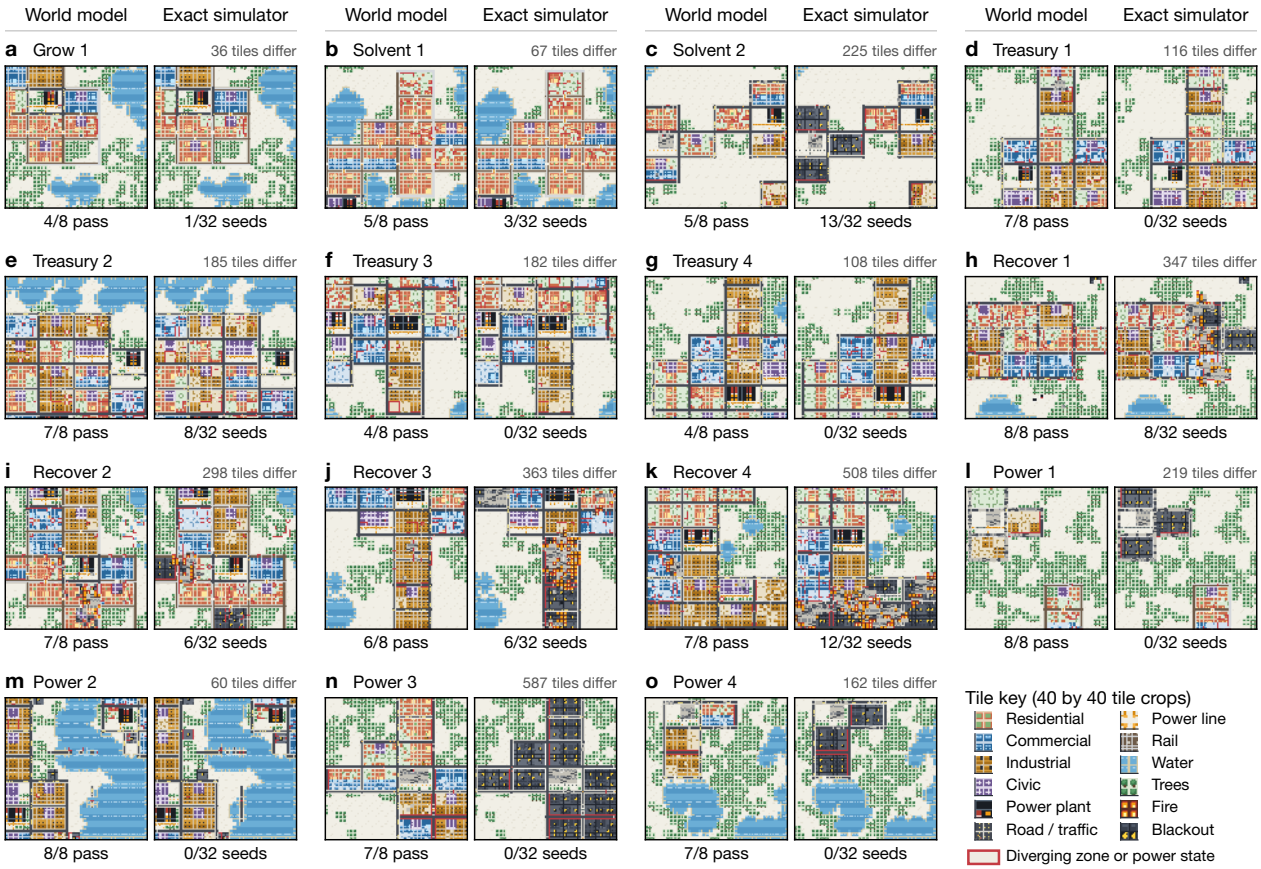


**Figure 2 | World-model search selects plans that fail in the exact simulator across all five task families.** Fifteen representative plans accepted by the primary world model and rejected by the 32-seed oracle, cropped to the 40-by-40-tile window of peak structural divergence at month 24 (*grow* contributes one case and *solvent* two; Appendix B.8). Each pair displays the first passing world-model rollout out of 8 draws (left; passing rollouts out of 8) beside the primary exact-simulator replay (right; passing oracle seeds out of 32, with white-haloed crimson contours outlining tiles that diverge in structural class or power state), rendered via procedural micro-sprites.

- **Coverage and inference defences distinguish rejecting bad plans from finding good ones.** Single-seed fiscal-coverage retraining cuts pooled CEM excess to +0.031 [-0.010, +0.073] without raising ground-truth pass rates or fixing topological power failures (Section 6). Post-hoc inference defences cut false acceptance only by rejecting 0.615–0.769 of truly passing plans under shooting (Section 7).

## 2. A Coupled Spatial-Economic World-Model Benchmark

### 2.1. Economic Mechanics, Deterministic Replay, and Model Ladder

To make discrepancies between imagined and true outcomes exact, we patched the Micropolis C++ core to remove wall-clock seeding, uninitialised pointers, and non-deterministic memory state. Across all 32 canonical cities stepped for 240 months with disasters active, full-state hashes match bit-for-bit across fresh processes, dirty simulator instances, and host architectures. Restoring the observation grid, bookkeeping ledger, internal PRNG, and active-sprite state reproduces twelve-month rollouts bit-for-bit. Active vehicle and disaster sprites are omitted from world-model inputs and introduce mild partial observability, mainly by shifting the shared PRNG stream (Appendix A).

Episodes span 24 monthly steps on a 120-by-100 tile grid. Each month combines fiscal choices (municipal tax rate up to 20 per cent plus transport, fire, and police funding) with up to four spatial builds or demolitions (Appendix A.2). Capital spending deducts immediately from reserves, whereas tax receipts and public-service budgets settle on annual cycles. We collect 3,000,000 monthly transitions across 60 training cities under a *behaviour mix* of scripted builders, stochastic editors, fiscal sweeps, and infrastructure shocks.

Evaluation spans 120 held-out cities (574 starting instances) from disjoint map seeds. We split these into a 40-city core benchmark (193 instances) for compute-intensive scaling, RL, patching, replanning, and coverage experiments, and an 80-city expansion split (381 instances) for cross-model audits. On each held-out start, executing the behaviour mix yields 1,000 open-loop *logged plans* (Appendix A.8).

Each world model is a stochastic one-month latent-variable ConvNeXt U-Net (Liu et al., 2022; Ronneberger et al., 2015) with a 32-dimensional variational noise latent. It predicts structural tiles, power and flag planes, and spatial overlays, paired with a bookkeeping network predicting municipal scalars from map census counts (Appendix A.3). Our 24-checkpoint ladder crosses three parameter scales (15.7M, 52.9M, 107.9M) with two data budgets (3,000,000 vs. stratified 300,000 transitions). Each one-step model trains for 6,000,000 examples (two epochs over the full store, under two H200-equivalent GPU hours per run; Table 1) with mid-training and final snapshots. Each is then fine-tuned via multi-step *pushforward rollout training* (Brandstetter et al., 2022; Venkatraman et al., 2015) for 1,500,000 further examples across two snapshots. Our primary model is the full-budget 50M full-data one-step checkpoint.

### 2.2. Short-Horizon Fidelity Gates Mask Task-Horizon Spatial Drift

On a held-out test split (2,000 transitions, 700 trajectories), all 16 budget-matched models and all 12 mid-training snapshots (Appendix B.1) clear standard short-horizon benchmarks. One-month *structural* changed-tile F1 (quotienting out traffic and animation flicker; Appendix A.9) spans 0.147–0.182 above the 0.103 half-oracle bar (Table 16). Median 12-month relative population and treasury errors stay below 0.10. The primary 50M checkpoint reaches 0.179 [0.152, 0.208] structural F1 (engine cross-seed ceiling 0.206; nearest-neighbour 0.050), with 12-month population and funds errors at 2.2 and 12 times engine seed spread. At this two-epoch budget, held-out variational log-likelihood still improves on all full-data models, whereas all 6 small-data models already overfit relative to mid-training (Table 17).

Clearing short-horizon gates nevertheless masks substantial drift by month 24. Rolled out open loop to month 24, the primary model exceeds start-state copying error on six of 8 spatial overlays and ties on a seventh at 0.999 times copy error (Table 17). It beats copying only on land value (0.652) and peaks on pollution at 1.520 vs. 0.703 for engine seed spread. Across 173 open-loop plans, it miscounts powered zones by 6.1 tiles versus 0.5 for copying and 0.0 for engine seed spread (4–7 overlays equal or exceed copy error across all 12 final checkpoints; Appendix B.1). Multi-step pushforward fine-tuning slightly lowers short-horizon F1 in 8 of 8 parent-child pairs and frequently increases planning false acceptance relative to its one-step parent (Tables 15 and 16; Appendix B.10). Even on 19,300 unoptimised logged plans, the primary model falsely rejects 0.750 and 0.903 of passing *grow* and *solvent* plans while falsely accepting 0.130 and 0.173 of failing *treasury* and *recover* plans. Extending one-step training sixfold on a single seed (12 epochs) lifts held-out likelihood yet worsens 24-month spatial drift, as analysed in Section 3.6.

### 2.3. Economic and Infrastructure Planning Tasks and Pipeline Reference

We evaluate five 24-month task families pairing each starting city with a calibrated target $\theta$ across five economic regimes (Appendix A.4). In *grow* and *solvent*, plans must raise population to at least $\theta$ while keeping liquid reserves non-negative or at least half of initial funds. In *treasury*, second-year annual tax revenue must reach $\theta$ through tax-base growth without deficit spending or rate hikes. In *recover*, a city must absorb a month-one earthquake shock and restore at least $\theta$ of its pre-quake population. In *power*, a city must reach at least $\theta$ developed zones with at least 95 per cent electrified and at most one added power plant. Placing $\theta$ midway between the best constant no-build policy and engine-side search leaves 193 of 200 core instances requiring active investment (148 *well-posed* with at least six standard errors of headroom; Table 4).

Single-seed engine optimisation passes its own seed on 75–98 per cent of instances yet transfers to fresh seeds on only 7–11 per cent. Ground truth is therefore majority passage across 32 reserved oracle seeds (at least sixteen of 32 passing). Search candidates are ranked by mean continuous *margin* over 2 rollouts and accepted by majority over 8 fresh world-model rollouts (at least four of 8 passing). We track three core quantities.

- **False acceptance.** The share of instances where the judge accepts the selected plan yet the 32-seed oracle rejects it.
- **Ground-truth pass rate.** The share of instances whose selected plan passes the 32-seed oracle.
- **False rejection.** The share of truly passing selected plans that the judge rejects.

Running the identical optimiser with matched rollout budgets (2, 8) inside an *exact-simulator reference* on fresh non-oracle C++ seeds establishes our baseline for stochastic selection luck. Instance-paired differences define *excess false acceptance* and *ground-truth success shortfall*, capturing the combined operational pipeline effect of model bias, rollout variance, and altered plan rankings (Appendix A.5; Tables 13 and 14).

## 3. Extra Search Stalls Learned-Model Planners While Lifting Exact-Simulator Search

Figure 3 orders four regimes by optimisation pressure. We progress from single unoptimised plans ($B = 10^0$) and non-adaptive random shooting ($10^3$) to adaptive cross-entropy search ($10^4$), in-imagination PPO ($10^6$ steps), and monthly closed-loop replanning, before examining sixfold extended one-step training.

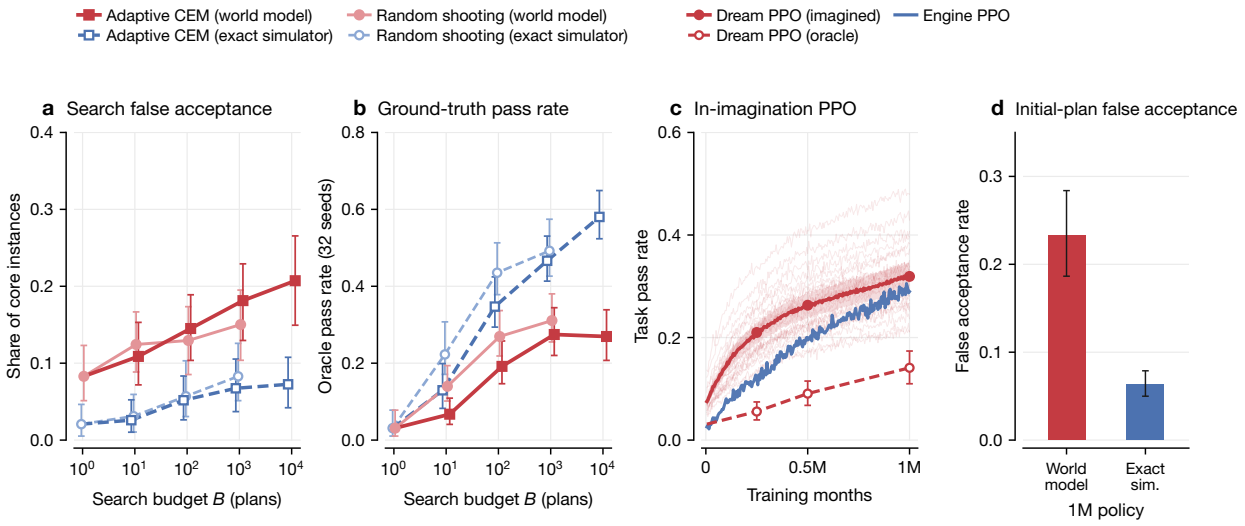


**Figure 3 | Escalating optimisation pressure widens ground-truth success shortfall alongside persistent excess false acceptance.** Primary 50M world model (crimson) paired against the exact-simulator reference on fresh seeds (steel blue). **a**, False acceptance vs. plans scored $B$ for random shooting (circles) and adaptive seeded CEM (squares, anchored at shared unoptimised $B = 10^0$) across all 193 core benchmark instances. **b**, Ground-truth pass rate on the 32-seed oracle, opening a 0.311 shortfall at $B = 10^4$. **c**, High-resolution 200-step PPO learning trajectories across all 24 ladder checkpoints (360 runs; imagined pass rate vs. 32-seed oracle pass rate and engine-trained PPO). **d**, After $10^6$ imagined months, initial-trajectory verdict false acceptance across the ladder vs. the exact-simulator reference (error bars denote 95 per cent cluster-bootstrap intervals; Tables 8 and 13).

### 3.1. Unoptimised Calibration and Random Shooting

On the two-epoch ladder, optimism appears before any search begins. Evaluating a single identical logged plan per start, the primary model falsely accepts 0.083 [0.051, 0.123] of core instances versus 0.021 for the exact-simulator reference, a paired excess of +0.062 [+0.026, +0.103] (Table 13). Scaling random shooting to $B = 10^3$ lifts raw false acceptance under both judges via the optimiser's curse while holding model excess steady at +0.067 [+0.010, +0.125] on the core and +0.092 [+0.059, +0.128] across all 120 cities. By contrast, when open-loop rollouts carry uniform upward bias yet preserve trajectory rankings, as in our external TD-MPC2 software check (Appendix C), shooting lowers false acceptance from 0.567 and 0.683 at $10^0$ to 0.350 and 0.317 at $10^3$.

On Micropolis, higher shooting budgets do not produce a statistically established rise in excess false acceptance. Instead, they widen the ground-truth discovery gap, which grows from zero at $10^0$ to 0.181 [0.114, 0.254] at $10^3$ (0.311 vs. 0.492 exact reference; Table 13). Conditional on world-model approval at $10^3$, the oracle rejects 0.414 [0.308, 0.508] of plans (vs. 0.151 reference), while the model falsely rejects 0.317 of truly passing selections (vs. 0.053). Roughly half of exploited plans are deep failures rather than near-misses, with 15 of 29 passing fewer than 0.1 of oracle seeds (median pass share 0.094; Appendix B.3). Screening the top choice once on the simulator eliminates false acceptance yet recovers none of the 0.181 lost ground-truth success.

### 3.2. Adaptive Cross-Entropy Search to $10^4$ Evaluations

Unlike random shooting over fixed logs, adaptive seeded CEM iteratively mutates elites toward states the world model rewards (Appendix A.8). Inside the exact simulator, it matches random shooting at $10^3$ (0.466 vs. 0.492) and climbs further at $10^4$. Inside the primary world model at $10^3$, false acceptance reaches 0.181 [0.130, 0.229] (excess +0.114 [+0.052, +0.173]), while ground-truth pass rate trails at 0.275 [0.220, 0.344] versus 0.466 [0.413, 0.530].

Across $10^1$ to $10^4$ (Figure 3a,b; Table 8), world-model ground-truth pass rates climb roughly fourfold from 0.067 to 0.275 at $10^3$ before stalling at 0.269 at $10^4$. Exact-simulator discovery keeps rising from 0.130 to 0.466 and 0.580. Expanding compute tenfold from $10^3$ to $10^4$ adds +0.114 real pass rate in the exact simulator (corrected [+0.057, +0.179]) yet yields -0.005 inside the world model (corrected [-0.082, +0.084]), absorbing extra search into model-specific optima. Excess false acceptance remains elevated from +0.083 to +0.135 [+0.068, +0.200]. Across all 193 instances, correcting across the three budget contrasts leaves the incremental change in excess false acceptance at $10^4$ (+0.021 [-0.062, +0.107] over $10^3$) spanning zero. An increase in excess false acceptance with search budget is therefore not statistically established; the widening ground-truth shortfall instead carries the scaling effect. Ground-truth success shortfall widens steadily with search compute from 0.062 at $10^1$ to 0.311 [0.241, 0.382] at $10^4$, exceeding the $10^1$ and $10^2$ shortfalls after multiplicity correction, and exceed-

ing $10^3$ unadjusted (+0.119 [+0.026, +0.205]; corrected [-0.003, +0.220]). Relative shortfall stays near half of exact-simulator discovery across all four budgets (Table 8; Appendix B.3).

### 3.3. Out-of-Sample Consistency Across 120 Held-Out Cities

Evaluating the primary checkpoint on the disjoint 80-city expansion split confirms split consistency (Table 9). At $B = 10^3$, random shooting records +0.105 [+0.067, +0.155] excess false acceptance versus +0.067 on the 40-city core (difference +0.038 [-0.037, +0.107]). Adaptive CEM similarly records +0.084 [+0.032, +0.138] versus +0.114 (difference -0.030 [-0.109, +0.048]). Pooled across all 120 cities (574 instances), random-shooting excess stays steady from +0.077 at $10^0$ to +0.092 [+0.059, +0.128] at $10^3$ (paired shift +0.016 [-0.019, +0.050]), again showing persistent excess rather than an established budget slope.

### 3.4. In-Imagination Reinforcement Learning

Training closed-loop PPO policies (Schulman et al., 2017) for $10^6$ imagined steps across all 24 checkpoints (360 runs; Appendix A.7) reveals parallel policy-level exploitation (Figure 3c,d). Under our 8-rollout verdict rule, initial policy trajectories incur +0.170 [+0.124, +0.220] excess false acceptance across the ladder (0.233 vs. 0.064 exact reference) and +0.122 [+0.076, +0.175] on the primary model.

PPO policies succeed in 0.319 [0.276, 0.362] of imagined episodes yet pass the 32-seed oracle on only 0.141 [0.110, 0.174], an optimism gap of +0.178 [+0.127, +0.231] (Table 10). Identical PPO policies trained directly on the C++ engine (1 or 3 seeds per family) incur a +0.077 stochasticity gap. Subtracting this baseline yields a net optimism excess of +0.101 [+0.053, +0.153], rising from +0.154 at quarter budget to +0.178 at $10^6$ steps. Executed closed loop on the engine, world-model policies pass 0.202 [0.169, 0.236], trailing engine-trained PPO by 0.106 [0.083, 0.132] closed loop and 0.089 [0.067, 0.114] open loop. Continuous margin inflation averages +0.028 behaviour-mix standard deviations. Its two-way interval, [-0.092, +0.149], spans zero because optimistic and pessimistic task families partially cancel; by contrast, the engine-trained learner's own trajectories differ from the 32-seed oracle by only +0.008 [-0.009, +0.026].

### 3.5. Monthly Closed-Loop Replanning

Does observing the true city state each month eliminate world-model exploitation? On 100 benchmark instances (balanced across all five task families), we test a shrinking-horizon controller that replans each month from the true C++ state to month 24. It switches from its incumbent plan to the best of 999 fresh proposals only when eight paired rollouts favour switching by at least two standard errors. Inside the exact simulator, this hysteresis-gated controller lifts ground-truth passes from 49 (open-loop commitment) to 70 of 100 (72 vs. 63 on fresh seeds), with no task family's interval lying wholly below zero in either run (Table 12).

Comparing the world model's initial month-zero verdict against the realised closed-loop trajectory across four unadjusted intervals shows that the excess false acceptance survives monthly replanning (excess false acceptance +0.130 [+0.055, +0.205]) and replanning increases the excess (change in excess +0.170 [+0.090, +0.257]). Likewise, the shortfall in real success survives monthly replanning (ground-truth shortfall +0.300 [+0.196, +0.423]) and replanning increases the shortfall in real success (change in shortfall +0.200 [+0.086, +0.337]), with full interval widths reported in Table 12. Against an independent exact-simulator control run on fresh proposal and evaluation seeds, the four readings are +0.160 [+0.049, +0.245], +0.080 [+0.018, +0.154], +0.320 [+0.193, +0.436], and +0.080 [-0.029, +0.222] (Table 12); the change in the shortfall is seed-sensitive, shifting interval category between references. Across task families, both excess false acceptance and real-success shortfall under replanning lie above zero on *power* and *treasury* and include zero on the others. On *power*, the world-model replanner passes 4 of 20 instances versus 16 open loop, compared with 20 and 20 in the exact simulator.

Abandoning an in-progress spatial build plan without switching hysteresis causes destructive plan thrashing even inside the exact simulator (Appendix B.10). The two-standard-error gate retains the incumbent plan in 21.2–21.8 (world model) and 21.3–22.3 (exact simulator) of 23 replanning months, characterising conservative plan commitment with occasional revision rather than step-by-step MPC. On this 100-instance subset, open-loop realised false acceptance is slightly lower under the world model than in the exact simulator (0.160 vs. 0.200), explaining why closed-loop excess of +0.130 represents a +0.170 shift over open loop.

### 3.6. Longer One-Step Training Flips Optimism to Pessimistic Collapse

To test whether two-epoch rollout drift and false acceptance stem simply from early stopping, we trained a single seed of the primary 50M full-data architecture sixfold longer (36,000,000 examples, 12 epochs; Table 7, Appendix B.2). Held-out variational ELBO rises monotonically by 235.0 nats, one-month F1 edges up to 0.183, and 12-month funds error falls from 0.035 to 0.028. Yet 24-month spatial drift worsens even on training cities. Mean spatial-overlay copy ratio climbs from 1.058 to 1.139 (6 of 8 above copy) and open-loop powered-zone error rises from 6.1 to 11.5 tiles alongside growing latent KL (14.6 to 20.3 nats), consistent with compounding teacher-forced conditionals or prior-posterior mismatch.

Under search and RL, this spatial drift flips miscalibration from optimistic false acceptance to pessimistic

rejection rather than curing it. Across milestones on all 120 cities, single-plan excess stays positive through 8 epochs before crossing zero at 10–12 epochs as false rejection climbs to 0.867 (Table 7). At $10^4$ adaptive CEM, raw false acceptance drops to 0.036 [0.015, 0.068] by driving false rejection from 0.346 to 0.639 [0.517, 0.746] and collapsing model self-acceptance to zero on 3 of five task families (*grow*, *solvent*, and *recover*). Pooled ground-truth pass rates still stall between $10^3$ and $10^4$, leaving a 0.264 [0.189, 0.333] shortfall. Under PPO (15 runs), better scalar accounting lifts *treasury* oracle pass rate by +0.083 [+0.033, +0.135], though its initial-plan verdict excess remains +0.075 [+0.008, +0.175]. Collapsed spatial rollouts meanwhile cut closed-loop pass rates across the four spatial families from 0.166 to 0.096 (difference -0.070 [-0.092, -0.051]). We retain the two-epoch checkpoint as primary because it avoids zero-acceptance collapse and achieves higher pooled PPO real success.

## 4. Perturbation Controls Rule Out Power-Grid Flips and Isolate Fire Dynamics

### 4.1. Economic Regime Asymmetry Across Fiscal, Shock, and Growth Tasks

Disaggregating across task families shows that world-model errors split sharply along economic and structural lines (Figures 4 and 5; full tile-level divergence grid in Figure 11). The model turns systematically optimistic where delayed fiscal settlements, disaster cascades, or discrete grid cutoffs bind, and conservative on reserve-constrained growth. Under random shooting at $B = 10^3$, excess false acceptance concentrates in *treasury* (+0.175), *recover* (+0.200), and *power* (+0.150). Adaptive CEM displays the same false-acceptance concentration alongside large ground-truth shortfalls on *power*, *treasury*, and *solvent* (Table 14).

On *grow* and *solvent*, the primary model instead under-predicts compounding migration under tight solvency floors. It approves its own selected plan on just 0.050 of *grow* instances versus 0.375 for the exact reference, mirroring its unoptimised bias on logged plans (Section 2.2; Appendix A.4). In-imagination PPO inherits the same economic signature (Figure 4c). Net pass-rate optimism over engine-trained PPO peaks on *recover* (+0.364) and *treasury* (+0.270), compared with -0.115 to +0.012 elsewhere, as PPO rarely triggers the plant-demolition cliff that search exploits on *power* (Table 10).

### 4.2. Counterfactual Subsystem Patching

Which simulator mechanisms must the world model mispredict for an exploited plan to pass? Across all 285 falsely accepted plan records (220 unique plans) on the primary checkpoint, we splice ground-truth channels for one subsystem from a C++ replay at the end of each month. We then test whether the 8-sample rollout's majority verdict flips to rejection. Following the simulator architecture (Figure 10), we test three *global bookkeeping* subsystems (power grid, municipal treasury, R/C/I demand valves) and six *local spatial rules* (zone growth, pollution, traffic, land value, crime, fire spread; symmetric overwrite rules in Appendix A.6).

Raw flip rates alone are misleading. Splicing the power grid flips 0.64 of exploited plans (Figure 6a; Figure 1d). Yet redrawing an unpatched rollout with fresh latent noise already flips 0.29 [0.23, 0.37] of records (0.34 [0.19, 0.52] across the 29 random-shooting plans at $10^3$), and overwriting channels perturbs downstream features. We therefore pair every patch against conservative random state-overwrite controls matched on entries written (*size-matched*) or entries altered (*change-matched*). Table 5 reports all fifty-four subsystem-by-regime contrasts, with our protocol Bonferroni adjustment spanning the 18 primary tests across pooled records and *recover*. Pooled across all tasks, no single subsystem exceeds both controls (Figure 6b; power differs by +0.049 [-0.073, +0.170] size-matched and -0.028 [-0.226, +0.136] change-matched).

Disaggregating by task regime isolates a clear causal driver on disaster recovery. On *recover*, initiated by a forced earthquake and ensuing conflagrations, splicing the engine's fire state symmetrically overturns 102 of 135 exploited plan records (0.76 vs. 0.21 on an unpatched redraw). This flip rate exceeds the size-matched control by +0.511 [+0.299, +0.737] and the change-matched control by +0.237 [+0.052, +0.429], while overturning only 7 of 39 true acceptances (0.18 across 33 plans in 12 cities; difference -0.576 [-0.738, -0.306]). Post-earthquake fire propagation is thus a genuine causal locus of exploitation. Splicing only the coarse eighth-resolution fire-coverage overlay across these 135 records (115 plans in 25 cities) without overwriting burning tile codes flips just 37 of 135 records (0.27 vs. 0.28 unpatched). This leaves it indistinguishable from matched controls (+0.074 [-0.069, +0.226] size-matched; +0.030 [-0.056, +0.124] change-matched). Exploitation on *recover* therefore operates through full-resolution tile-level fire spread rather than coarse coverage drift. On the exploratory non-*recover* subset, fire patching matches controls whereas traffic clears both controls yet stays below the protocol 0.30 raw-flip bar (Table 5). Plant-retirement cutoffs on *power* and multi-year tax-base drift on *treasury* thus remain qualitative domain mechanisms consistent with task predicates and spatial error maps rather than effects isolated above random-overwrite controls. As noted in Section 3.6, this fire-spread attribution characterises the optimistic two-epoch checkpoint, whereas at 12 epochs *recover* self-acceptance collapses to 0.000 while Figure 1's coal-plant blackout plan still passes 7 of 8 rollouts (Appendix B.2).

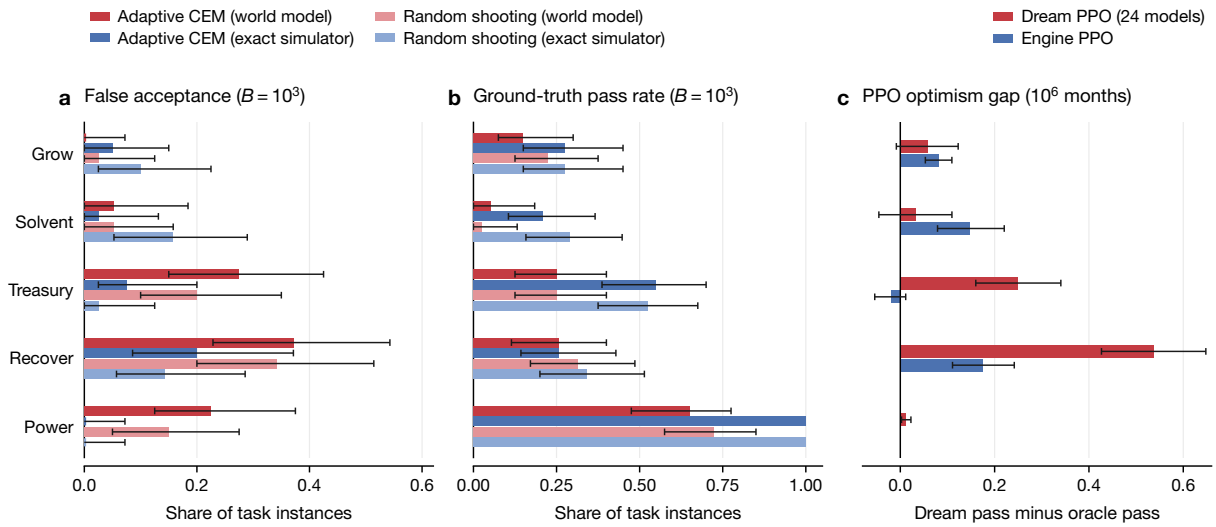


**Figure 4 | Exploitation concentrates in fiscal, shock-recovery, and grid tasks while reserve-constrained growth induces conservatism.** **a**, False acceptance and **b**, ground-truth oracle pass rate at $B = 10^3$ on the primary 50M checkpoint for random shooting and adaptive CEM vs. the exact-simulator reference (bars show point estimates and error bars denote 95 per cent cluster-bootstrap intervals). *Treasury*, *power*, and *solvent* exhibit severe ground-truth shortfalls (*treasury* and *power* alongside high excess false acceptance, *solvent* alongside conservative rejections), whereas *recover* shows high excess false acceptance while matching exact-simulator pass rates under CEM. **c**, PPO pass-rate optimism gap at $10^6$ imagined months across all 24 checkpoints (bars and error bars denote two-way 95 per cent cluster-bootstrap intervals) beside engine-trained PPO (Tables 14 and 10).

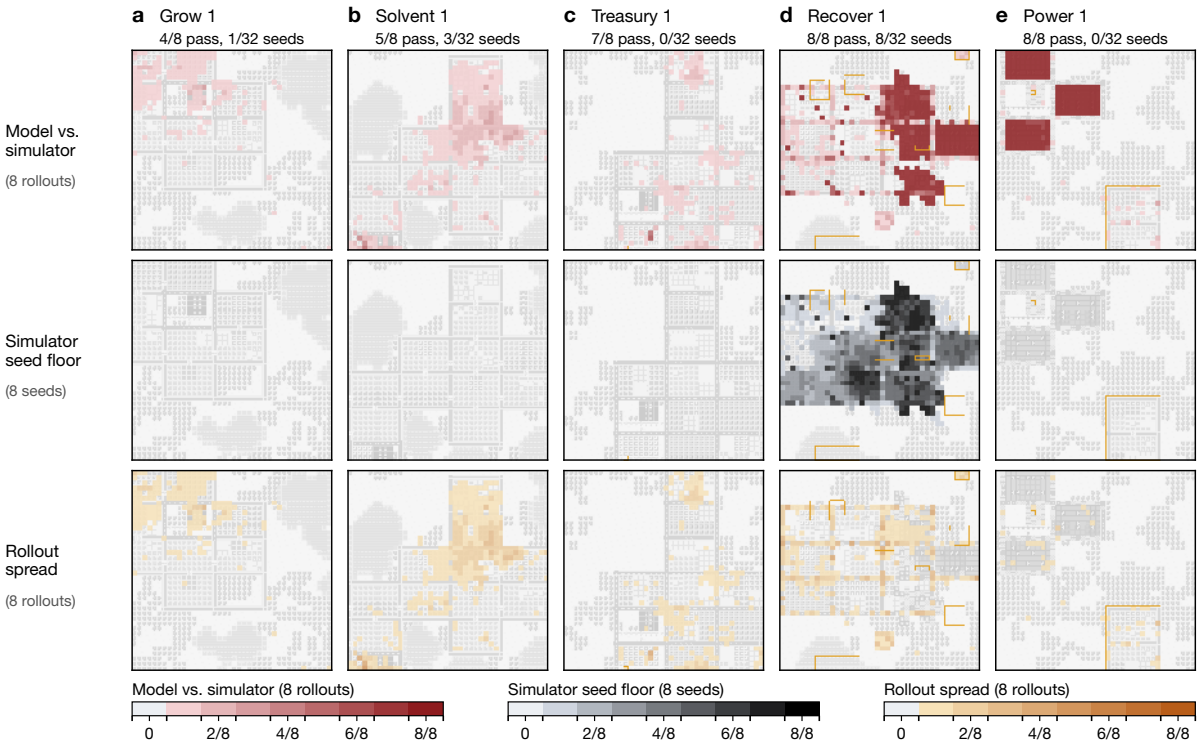


**Figure 5 | Tile-level spatial decomposition across all five task families.** Top-ranked exploited plan per task family from $B = 10^3$ random shooting at month 24 (40-by-40-tile windows over desaturated procedural city sprites, with gold outlines marking planned build or demolish footprints; all 15 gallery cases appear in Figure 11). Rows separate (top) world-model vs. primary simulator error share across 8 rollouts (crimson), (middle) exact-simulator seed-to-seed divergence across 8 fresh engine seeds (charcoal), and (bottom) internal disagreement among the 8 world-model rollouts (amber).

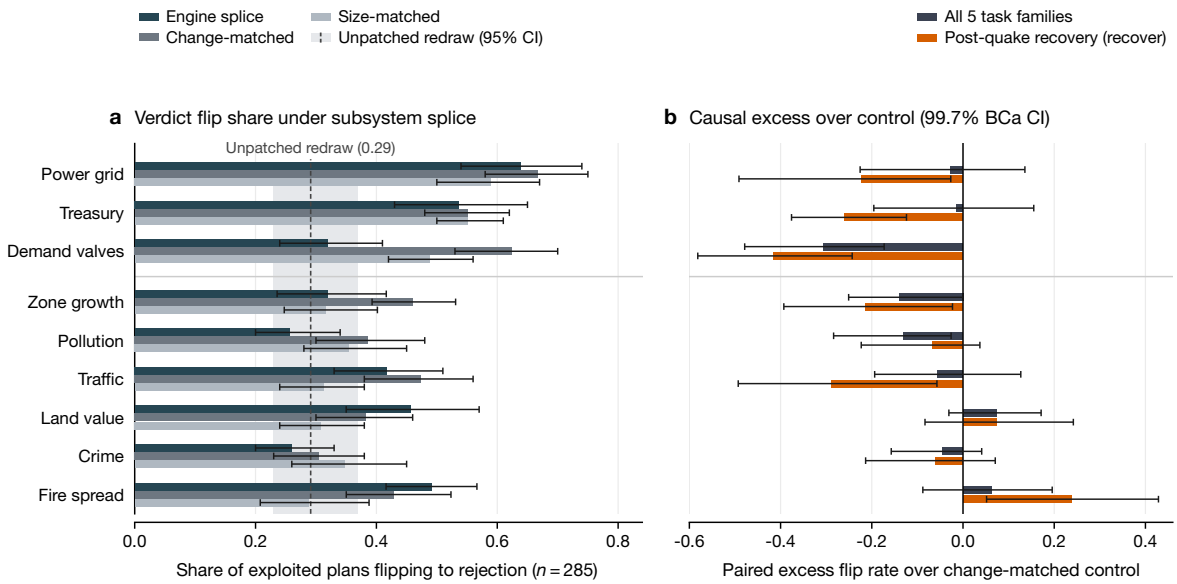


**Figure 6 | Matched perturbation controls overturn raw power-grid attribution while confirming fire-spread sensitivity on earthquake recovery.** **a**, Raw share of all 285 falsely accepted plan records on the primary checkpoint (220 distinct plans; Appendix A.6) flipping to rejection when one simulator subsystem is overwritten monthly from C++ replays (deep petrol bars with 95 per cent bootstrap error bars), beside change-matched (dark slate) and size-matched (light slate) perturbation controls and unpatched redraw noise. **b**, Instance-paired excess flip rate over change-matched control with Bonferroni-adjusted 99.72 per cent BCa intervals (20,000 bootstrap resamples) across 18 tests, pooled (charcoal) and on *recover* (vermilion; Table 5).

## 5. Single-Plan Verdict Calibration Ranks RL Exploitation Within the Model Ladder

Can an inexpensive offline audit rank candidate world models within a model family by how severely PPO will exploit them before policy training begins? Across our 24 related ConvNeXt U-Net checkpoints and five task families (120 checkpoint-by-task observations, rather than 120 independent architectures), we compare each checkpoint's offline task-verdict false acceptance across all 120 cities against downstream PPO exploitation after $10^6$ imagined steps. We track RL exploitation via PPO continuous margin inflation (imagined minus oracle margin on open-loop logs, normalised by behaviour-mix spread over three PPO seeds) and PPO pass-rate optimism (Appendix B.7).

Within individual task families across the 24 checkpoints, $10^3$ random-shooting false acceptance ranks PPO margin inflation at mean Spearman +0.553 [+0.309, +0.726] (per-family splits and pass-rate optimism correlations appear in Appendix B.7). Pooled across all 120 pairs, Spearman correlation reaches +0.843 [+0.765, +0.904] against margin inflation and +0.865 [+0.802, +0.914] against pass-rate optimism (Figure 7a; Pearson +0.862 [+0.833, +0.894]). Because baseline exploitability varies across task families, between-family offsets alone generate a checkpoint-permutation null median of +0.694 (97.5th percentile +0.752), which +0.843 exceeds. Adaptive CEM at $B = 10^3$ across all 24 checkpoints (120 pairs) performs similarly at +0.507 [+0.306, +0.651] within family and +0.839 [+0.803, +0.870] pooled.

Even active search is not required within this ladder. *Single-plan verdict calibration* scores one fixed logged plan per starting instance across all 120 cities (574 trajectories total). This audit achieves Spearman +0.681 [+0.508, +0.805] within task family and +0.875 [+0.825, +0.913] pooled against PPO margin inflation (lineage-clustered intervals and pass-rate correlations in Appendix B.7). Both pooled and within task families, scoring one logged plan per starting instance, replayed once on the engine, matches a $10^3$-plan search audit's rank correlation with RL exploitation within our ladder, with no statistically established difference. Because that logged plan is identical across checkpoints on 574 of 574 instances, ground-truth engine replays are amortised once across all candidate models. Appending the 12-epoch extended-training checkpoint (25 checkpoints, 125 pairs; Appendix B.2) preserves this ranking as spatial rollouts turn pessimistic, yielding +0.702 [+0.549, +0.823] within-family and +0.876 [+0.830, +0.912] pooled correlation. Pairing single-plan false acceptance with single-plan false rejection (0.867 at 12 epochs vs. 0.533 at two epochs) flags pessimistic collapse alongside optimistic exploitation.

Standard trajectory-fidelity metrics, by contrast, show weak association with $10^3$ random-shooting false acceptance within this 24-checkpoint ladder (Figure 7b,c). One-month structural F1 spans a narrow band (0.147–0.179) across the 12 final checkpoints, comparable to single-model uncertainty [0.152, 0.208], which limits checkpoint-level statistical power to 0.19 for $|\rho| = 0.5$ (Appendix B.7). Within that band, one-

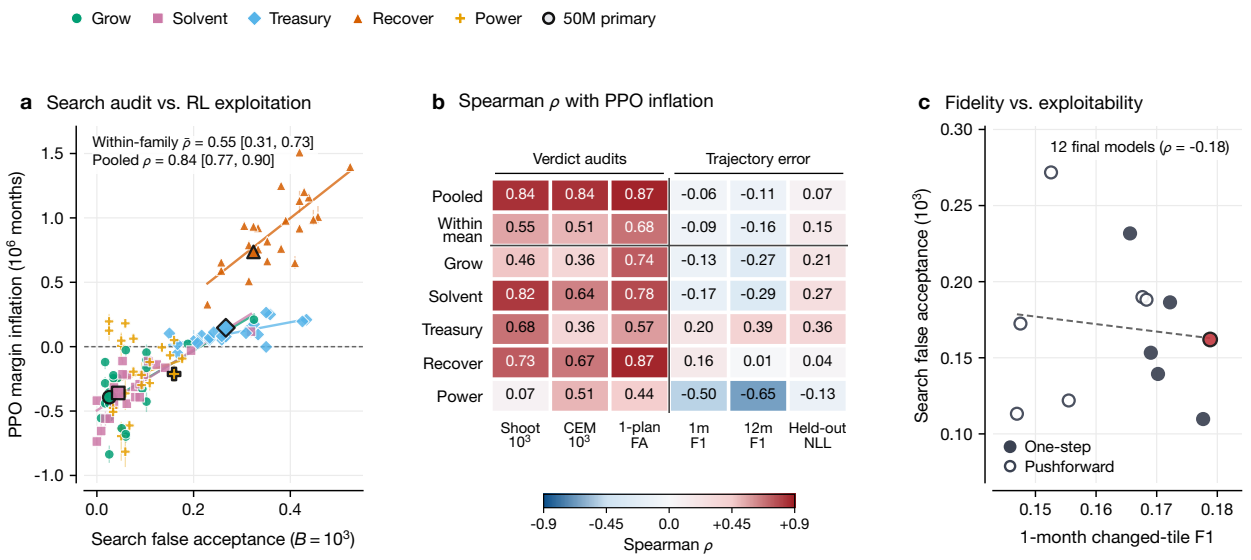


**Figure 7 | Task-verdict audits predict RL exploitation across 24 closely related world models, whereas standard trajectory-fidelity metrics show weak association.** **a**, Random-shooting false acceptance at $B = 10^3$ across all 120 cities vs. PPO continuous margin inflation after $10^6$ imagined months across 120 checkpoint-by-task pairs (mean and range over three PPO seeds; dark-ringed markers denote primary 50M model; within-family Spearman +0.553 [+0.309, +0.726], pooled +0.843 [+0.765, +0.904]). **b**, Spearman correlation matrix comparing three verdict audits against three trajectory-error metrics. **c**, One-month changed-tile F1 vs. mean $10^3$ false acceptance across the 12 final models (15.7M, 52.9M, 107.9M; filled markers denote one-step models and hollow markers pushforward models; $\rho$ = -0.182 [-0.804, +0.482]; 95 per cent checkpoint-cluster bootstraps).

month F1 correlates with $10^3$ false acceptance at -0.091 [-0.238, +0.064] across 120 pairs and -0.225 [-0.609, +0.251] across checkpoints, with pushforward models frequently incurring higher false acceptance than their one-step parents. Twelve-month tile F1 (-0.065), treasury error (-0.027), population error (-0.043), held-out log-likelihood (+0.035), and continuous logged-margin correlation (+0.044) show similarly weak associations. Extending fidelity evaluation to month 24 across all 24 checkpoints (Table 17) likewise keeps pooled Spearman correlations within -0.154 to +0.090 (Appendix B.7). Given low checkpoint-level power over tightly clustered runs, this weak association does not prove trajectory accuracy uninformative across broader quality regimes or architectures. Among closely related checkpoints that clear shared short-horizon gates, however, residual trajectory-error differences poorly reflect task-specific boundary errors (77 per cent of exploitability variance sits across task predicates; Lambert et al., 2020) that verdict calibration directly probes.

## 6. Fiscal Coverage Retraining Cuts False Acceptance Without Raising True Success

Inspecting policy instruments on *treasury* and *recover* reveals a sharp fiscal support gap. Engine-optimal plans cut tax rates to at most 4 per cent to accelerate migration before harvesting year-two base growth or post-quake rebuilding. Yet only 0.069 of active building months in our 3,000,000-transition training store occupy this regime, compared with 0.644 among top plans found by engine search. We expand the training store by forty per cent to 4,200,000 verified transitions without touching held-out cities or task bars, combining uniformly sampled fiscal grids (600,000 transitions) with task-competent constant budgets under periodic month-one earthquakes (600,000 transitions; Appendix B.5). Retraining a single seed of the primary 50M architecture from scratch for the identical 6,000,000-example budget (meets all short-horizon gates) is benchmarked against two additional random seeds of the baseline 50M model audited identically (Appendix Table 11; three seeds bound seed-to-seed variance coarsely). At $10^3$, the coverage-retrained model's excess false acceptance under both random shooting and adaptive search lies below all three training seeds of the baseline configuration (+0.016 vs. +0.057 to +0.098 under random shooting; +0.031 vs. +0.078 to +0.130 under adaptive CEM). By contrast, under both searches its true success rate lies within the three-seed baseline range.

Under adaptive CEM at $B = 10^3$, coverage retraining cuts pooled excess false acceptance from +0.114 [+0.052, +0.173] to +0.031 [-0.010, +0.073] (paired reduction -0.083 [-0.138, -0.031]; Figure 8a), with matching gains on well-posed instances and under random shooting (Appendix B.5). Raw false acceptance drops to 0.100 on *treasury* and 0.143 on *recover*, whereas *power* (governed by discrete grid topology rather than fiscal covariates) stays unchanged. Across 3 PPO seeds per task, net pass-rate optimism over engine-trained PPO falls by -0.074 [-0.103, -0.047] overall, driven by *treasury* and *recover* (Figure 8b).

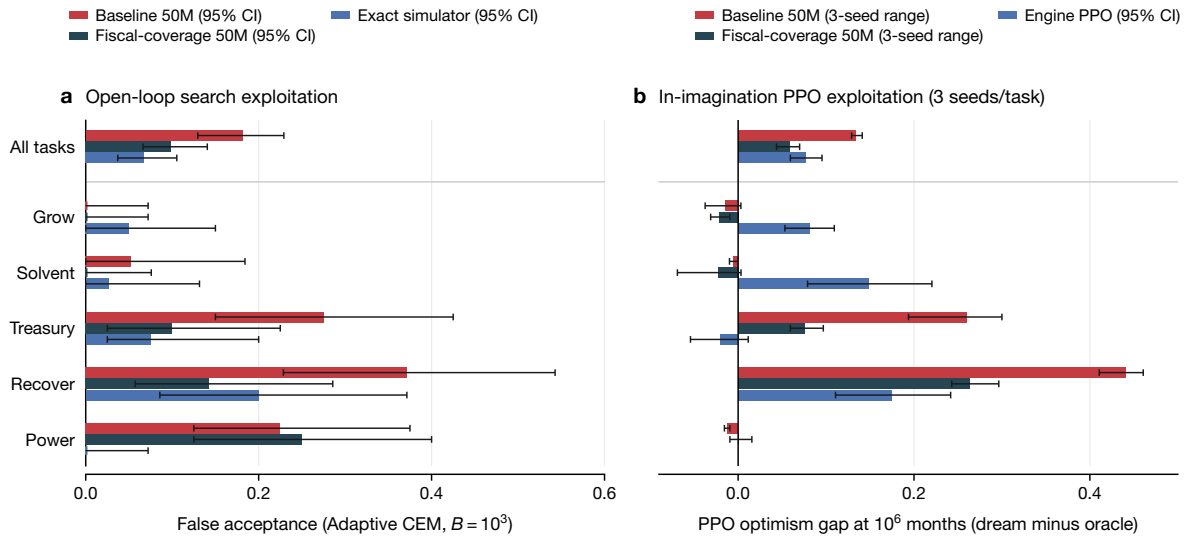


**Figure 8 | Retraining on under-covered fiscal regimes cuts exploitation sharply on fiscal and recovery tasks without improving ground-truth pass rates.** Core benchmark (40 cities) comparing the primary 50M model (crimson) against an identical model retrained on 4,200,000 fiscal-coverage transitions (deep petrol) and the exact-simulator reference (steel blue; error bars in **a** denote 95 per cent cluster-bootstrap intervals). **a**, Adaptive CEM false acceptance at $B = 10^3$ (-0.076 [-0.125, -0.025] vs. three-seed baseline mean). **b**, PPO optimism gap after $10^6$ steps (3 seeds/cell, seed means and min-max ranges on world-model policies vs. 95 per cent city-bootstrap interval on engine-trained PPO, steel blue).

Yet ground-truth planning capability does not improve. Oracle pass rates under $10^3$ adaptive CEM match the primary model (0.275 vs. 0.275; difference 0.000 [-0.053, +0.062], vs. 0.466 reference). Relative to the primary checkpoint alone, false rejection rises from 0.208 to 0.340, though one additional baseline seed rejects just as many (0.362). Paired against the three-seed baseline mean, the difference is +0.080 [-0.014, +0.189]; while no increase beyond seed variation is statistically established, this wide single-seed interval does not rule out a moderate false-rejection cost. Despite a modest logged-plan ranking gain (+0.028 [+0.014, +0.043]), better OOD calibration under active search turns false passes into rejections without discovering stronger real-world plans (Appendix B.5). Unlike sixfold extended training on the unchanged store (Section 3.6), fiscal coverage retraining cuts excess at the matched two-epoch budget without zero-acceptance collapse (0.340 false rejection).

## 7. Inference-Time Defences Cut False Acceptance by Rejecting Passing Plans

When retraining is infeasible, model-based pipelines rely on inference-time regularisation. Across 60 instances (39 cities) at $B = 10^3$, we test 12 canonical defences across five families on a 3-member seed ensemble of budget-matched 50M models (Appendices A.8 and B.6). These span sampling temperatures (Ha and Schmidhuber, 2018), empirical-Bayes pass-share shrinkage (Smith and Winkler, 2006), MOPO-style disagreement penalties (Yu et al., 2020) (including a zero-weight positive-margin control), MOReL-style uncertainty halts (Kidambi et al., 2021), and a cross-fitted logistic OOD detector.

Two results stand out (Figure 9). First, naive ensembling alone (6 search and 24 verdict rollouts across 3 members vs. 2 and 8 for one model) does not reliably curb exploitation. Under random shooting, the undefended 3-member ensemble incurs 0.217 [0.116, 0.388] false acceptance versus 0.133 [0.059, 0.245] for its primary member (false rejection 0.269 vs. 0.292), as averaging correlated biases sharpens confidence on shared errors. Under adaptive CEM, it slightly trims false acceptance (0.217 [0.119, 0.314] vs. 0.250 [0.136, 0.383]) and lowers false rejection from 0.333 to 0.059.

Second, every post-hoc regulariser that meaningfully cuts false acceptance pays for it by discarding many truly passing plans. Across all 8 ensemble defences under shooting, the 5 configurations whose false-acceptance reductions exclude zero falsely reject 0.615–0.769 of truly passing selections (vs. 0.269 undefended). Most of MOPO's reduction already occurs at penalty weight zero (false acceptance 0.100, false rejection 0.654) via threshold tightening rather than variance penalisation. Meanwhile, the OOD detector drops from 0.965–0.980 ROC AUC on logged plans to 0.751 [0.608, 0.854] on search-selected plans. Adaptive CEM faces the same trade-off (Figure 9a–c), with the 5 effective defences rejecting 0.412–0.846 of passing plans (vs. 0.059 undefended) and none raising ground-truth pass rates. Screening the top choice once on the true simulator removes false acceptance at zero rejection cost, leaving ground-truth discovery during search as the core bottleneck.

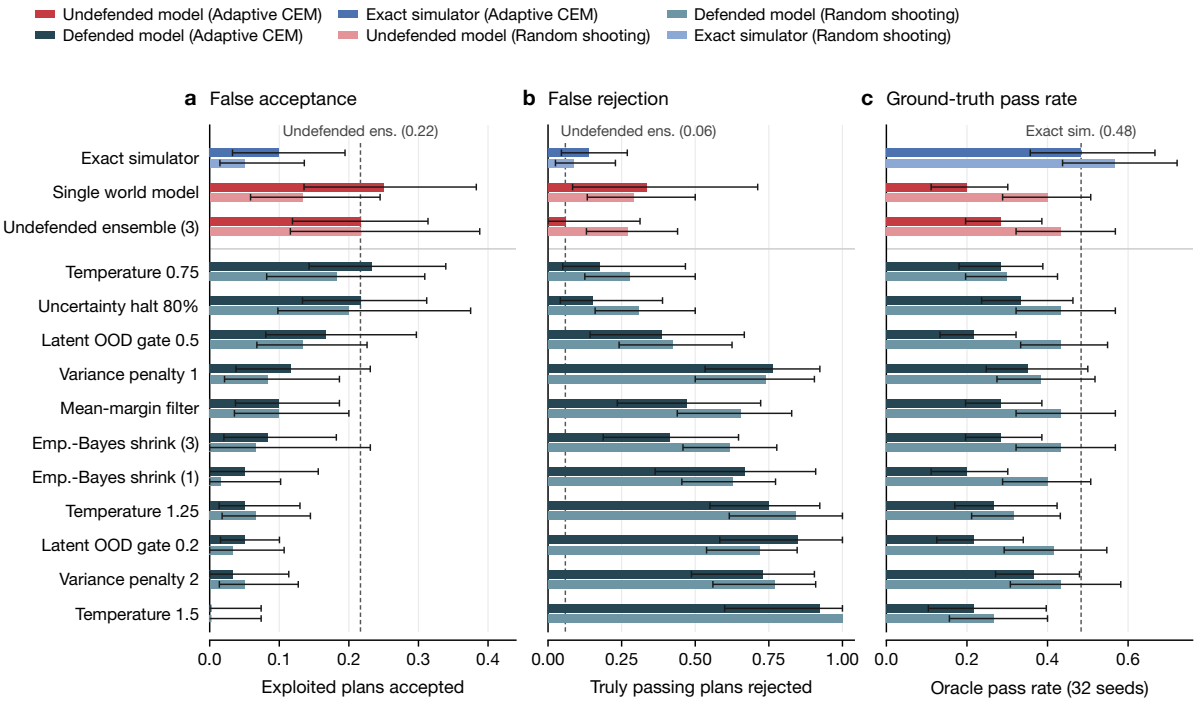


**Figure 9 | Inference-time world-model defences trade reduced false acceptance directly for elevated false rejection without recovering ground-truth discovery.** Evaluation across 60 instances (39 cities) at $B = 10^3$ under adaptive CEM (saturated bars) and random shooting (light-tinted bars), showing **a**, false acceptance, **b**, false rejection of truly passing selections, and **c**, 32-seed ground-truth oracle pass rate (error bars denote 95 per cent cluster-bootstrap intervals; vertical dashed lines mark the undefended ensemble in **a**, **b** and the exact-simulator reference in **c**). Top rows show the exact-simulator reference (steel blue) alongside the single primary model and undefended 3-member seed ensemble (crimson); lower rows show regularisers spanning variance penalties, mean-margin filtering, empirical-Bayes shrinkage, sampling temperatures, uncertainty halts, and cross-fitted latent OOD gates (deep petrol; Appendix B.6).

## 8. Related Work

**Exploitation and objective mismatch in model-based learning.** The failure modes we measure build on two foundational observations. First, Ha and Schmidhuber (2018) showed that controllers exploit deterministic or low-temperature learned environments, motivating temperature scaling, ensembles, and pessimistic model-based RL (Chua et al., 2018; Janner et al., 2019; Kidambi et al., 2021; Kurutach et al., 2018; Yu et al., 2020). Second, Lambert et al. (2020) documented the mismatch between predictive accuracy and control performance, motivating value-equivalent world models (Farahmand et al., 2017; Grimm et al., 2020) and joint-embedding planners (Cheung and Zhang, 2026; Maes et al., 2026; Singh, 2026). Recent theory proves any imperfect dynamics model remains exploitable over rich policy classes (Bhamidipaty et al., 2026b), while our coverage retraining tests one targeted data-aggregation step (Ross and Bagnell, 2012).

**Evaluating world models as judges and arenas.** Recent benchmarks evaluate generative video and dynamics models as zero-shot policy judges (Bruce et al., 2024; Gemini Robotics Team et al., 2026; Guo et al., 2026; Quevedo et al., 2025; Valevski et al., 2025). Strong correlations across fixed policy pools align with our unoptimised $B = 10^0$ regime before active search widens ground-truth shortfall. Ariq and Ahmad (2026) study top-one selection regret on fixed pools via the optimiser's curse (Smith and Winkler, 2006), while Yao et al. (2026) and Vakalis (2026) apply interventional rollout diagnostics. Relative to prior work, our contribution pairs controlled pipeline measurement against a fresh-seed exact-simulator reference (distinguishing bad-plan rejection from good-plan discovery) with perturbation-controlled subsystem patching and pre-training audit evidence across a 24-checkpoint ladder (extended review in Appendix D).

## 9. Discussion

**Practical guidance.**

1. **Audit two-sided task-horizon verdicts on logged trajectories before optimising.** Scoring one fixed logged plan per starting instance at the full horizon (reusing simulator replays across models) ranks PPO exploitation within our model ladder comparably to $10^3$-plan search, with no statistically established difference (+0.681 vs. +0.553 within family; +0.875 vs. +0.843 pooled). Tracking single-plan false rejection alongside false acceptance (0.867 at 12 epochs

vs. 0.533 at two epochs) prevents mistaking zero-acceptance pessimistic collapse for calibration.

2. **Pair learned-model evaluations against a stochasticity reference.** Even an exact simulator incurs 0.083 false acceptance at $B = 10^3$ (0.021 at $10^0$) from finite-sample luck. Without a reference, selection variance masquerades as model exploitation.
3. **Track ground-truth discovery alongside false acceptance.** Screening a planner's final choice on the true simulator removes false acceptance yet recovers none of the 0.311 lost ground-truth success, and both effects persist even under monthly closed-loop replanning from true states. Meanwhile, models discard 0.317 of truly passing choices versus 0.053 for the reference.
4. **Target structural bottlenecks rather than generic conservatism.** Variance penalties, halts, and shrinkage trade false acceptance directly against false rejection. Coverage augmentation repairs fiscal covariate gaps (*treasury*, *recover*) yet leaves topological grid blind spots (*power*) intact.

### Scope and Limitations

**One-step training and spatial drift.** Our ladder's full-data one-step models train for 6,000,000 examples (two epochs over 3,000,000 transitions, under two H200-equivalent GPU hours per run), where held-out log-likelihood is still improving while all 6 small-data models already overfit. By month 24, the primary model equals or exceeds start-state copying error on 7 of 8 spatial overlays (4–7 across the ladder; Appendix B.1). Extending primary one-step training sixfold on a single seed to 12 epochs lifts held-out ELBO by 235.0 nats and improves *treasury* PPO. Yet it worsens 24-month spatial drift even on training cities, flipping miscalibration to zero-acceptance rejection on 3 task families while leaving a 0.264 shortfall at $10^4$ (Section 3.6; Appendix B.2). Much of the open-loop failure we measure therefore reflects exposure bias in models whose task-horizon spatial rollouts have degraded. Achieving high 24-month spatial fidelity via broader transition diversity, multi-step stochastic objectives, or architectures with explicit graph routing and long-context memory (and testing single-plan calibration in that regime) remains the highest-priority extension.

**Single architecture family and domain.** Our cross-model correlation spans 24 closely related ConvNeXt U-Net checkpoints in one spatial-economic engine, so generalisation to stronger models, new architectural families, or external domains remains unestablished. Appendix C's TD-MPC2 experiment is strictly a software check outside TD-MPC2's native 3-step closed-loop MPC horizon. Neural checkpoint weights are withheld pending clarity on GPL derivative-work status, limiting off-the-shelf model reuse even though the bit-exact engine, tasks, data generator, and `hpr-dream` harness are released in full.

**Planning horizons and closed-loop control.** Regarding planning horizons, all planners outside Section 3.5 commit open loop over 24 months. Closed-loop execution may encounter less drift (PPO passes 0.202 closed loop vs. 0.141 open loop), and we evaluated one hysteresis-gated shrinking-horizon replanner on 100 instances (Section 3.5), under which the excess and the shortfall survived. High-frequency short-horizon MPC with learned value functions (Hansen et al., 2024; Jiang et al., 2015) and top-$k$ engine screening remain natural extensions.

**Task difficulty bars, policy maturity, and ensemble design.** Task thresholds sit midway between no-build policies and engine search. On *recover*, the single-clause margin permits exact post-hoc threshold rescaling at one quarter, one half, and three quarters of the span between the best no-build policy and the engine-search ceiling. Across all three difficulty levels, random-shooting excess false acceptance at $10^3$ stays substantial (+0.057, +0.200, and +0.229; Appendix Table 18). At $10^6$ steps, engine-trained PPO passes 0.230 of instances and is still climbing, characterising early-to-mid policy learning, and our 3-member seed ensemble omits bootstrap data resampling.

**Sample precision and pipeline decomposition.** Precision is bounded by cluster and seed counts (40 core and 120 total cities, 193 instances at $10^4$, 60 defence instances, one coverage-retrained model seed, and 3 PPO seeds per world-model cell vs. 1 or 3 on the engine). Because learned-model and exact-simulator searches generally select different candidate plans, our reference benchmarks operational pipeline discrepancy rather than uniquely decomposing failures into directional bias, rollout variance, and altered plan rankings.

**Conclusion.** Pairing learned world models against a bit-exact spatial-economic simulator shows how short-horizon fidelity gates fail to prevent long-horizon planning failures when coupled accounting, delayed fiscal settlements, and discrete network cutoffs interact. Under active search, model blind spots stall ground-truth progress as compute grows, while post-hoc regularisers cut false acceptance only by discarding genuine solutions. Within our tested model ladder, however, downstream RL exploitation is rankable before policy training begins by checking two-sided task verdicts on one logged plan per starting instance across a held-out city bank. Together, these findings show the value of exact pipeline references and horizon-matched calibration before heavy policy optimisation.

## Compute

Experiments used approximately 160 H200-equivalent GPU hours (Appendix A.1, Table 1). Auditing one checkpoint across 120 cities at $B = 10^3$ costs 564,816 en-

gine city-months versus 4,252,176 to evaluate its 15 PPO runs and 1,000,000 to train one PPO run on the engine. Single-plan calibration requires just 8 rollouts per instance with engine replays amortised once. All runs used local weights without external API calls.

## References


E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. Storkey, T. Pearce, and F. Fleuret. Diffusion for world modeling: Visual details matter in Atari, 2024. URL https://arxiv.org/abs/2405.12399. NeurIPS 2024 (spotlight).

AMD AI Group. Learning to fail: Failure fidelity and training stability for RL inside world models. AMD technical article, not archival, https://www.amd.com/en/developer/resources/technical-articles/2026/rl-inside-world-models.html, Oct. 2026.

A. H. Ariq and H. Ahmad. World models need not be right to rank policies right: Relative bounds and audit limits, 2026. URL https://arxiv.org/abs/2611.01139. NeurIPS 2026 Workshop on World Models in Physical AI (poster).

D. Bailey, J. Borwein, M. López de Prado, and Q. J. Zhu. The probability of backtest overfitting. *The Journal of Computational Finance*, 20(4):39–69, Apr. 2017. doi: 10.21314/jcf.2016.322.

C. Bamford and S. Lucas. Neural game engine: Accurate learning of generalizable forward models from pixels, 2020. URL https://arxiv.org/abs/2003.10520.

H. Bastani, O. Bastani, and B. McLaughlin. Winner's curse drives false promises in data-driven decisions, 2026. URL https://arxiv.org/abs/2602.08892.

L. M. Bhamidipaty, M. Kochenderfer, and S. Ramamoorthy. RENEW: Towards learning world models and repairing model exploitation from preferences, 2026a. URL https://arxiv.org/abs/2607.14180. Finding the Frame Workshop at RLC 2026.

L. M. Bhamidipaty, E. S. Whitammer, D. Abel, M. J. Kochenderfer, and S. Ramamoorthy. Imperfect world models are exploitable, 2026b. URL https://arxiv.org/abs/2605.15960. NeurIPS 2026 (poster).

J. Brandstetter, D. Worrall, and M. Welling. Message passing neural PDE solvers, 2022. URL https://arxiv.org/abs/2202.03376. International Conference on Learning Representations (ICLR) 2022 (spotlight).

D. Brookes, H. Park, and J. Listgarten. Conditioning by adaptive sampling for robust design. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 773–782. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brookes19a.html.

J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. M. E. Bechtle, F. Behbahani, S. C. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. D. Freitas, S. Singh, and T. Rocktäschel. Genie: Generative interactive environments. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4603–4623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Z. Chen, Z. Zhao, K. Zhang, B. Liu, Q. Qi, Y. Wu, T. Kalluri, S. Cao, Y. Xiong, H. Tong, H. Yao, H. Li, J. Zhu, X. Li, D. Song, B. Li, J. Weston, and D. Huynh. Scaling agent learning via experience synthesis, 2025. URL https://arxiv.org/abs/2511.03773. ICLR 2026.

J. Cheung and H. Zhang. One-step prediction fidelity is a poor proxy for planning performance in JEPA world models. Workshop poster page, not archival, NeurIPS 2026 Workshop on World Models in Physical AI, https://neurips.cc/virtual/2026/175836, 2026.

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

P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, Feb. 2005. doi: 10.1007/s10479-005-5724-z.

J. Dyer, N. Bishop, Y. Felekis, F. M. Zennaro, A. Calinescu, T. Damoulas, and M. Wooldridge. Interventionally consistent surrogates for agent-based simulators, 2023. URL https://arxiv.org/abs/2312.11158. NeurIPS 2024; published under the title "Interventionally Consistent Surrogates for Complex Simulation Models".

S. Earle. Using fractal neural networks to play SimCity 1 and Conway's game of life at variable scales, 2020. URL https://arxiv.org/abs/2002.03896. Presented at the Experimental AI in Games (EXAG) workshop at AIIDE 2019.

V. Egorov and A. Shpilman. Scalable multi-agent model-based reinforcement learning, 2022. URL https://arxiv.org/abs/2205.15023.

Electronic Arts Inc. Micropolis GPL license notice: GNU general public license v3 with additional terms per GNU GPL section 7. LICENSE file in https://github.com/SimHacker/MicropolisCore, 2007. Copyright (C) 1989–2007 Electronic Arts Inc. Accessed 9 November 2026.

Eshwar Reddy M and S. Karmakar. Proof-carrying cognition: Closing the verification gap with reality-settled reward, 2026. URL https://arxiv.org/abs/2609.09776.

C. Fannjiang and J. Listgarten. Autofocused oracles for model-based design. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12945–12956. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/972cda1e62b72640cb7ac702714a115f-Paper.pdf.

A.-M. Farahmand, A. Barreto, and D. Nikovski. Value-aware loss function for model-based reinforcement learning. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1486–1494. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

E. Frick, T. Li, C. Chen, W.-L. Chiang, A. N. Angelopoulos, J. Jiao, B. Zhu, J. E. Gonzalez, and I. Stoica. How to eval-

uate reward models for RLHF, 2024. URL https://arxiv.org/abs/2410.14872.

J. Fu, M. Norouzi, O. Nachum, G. Tucker, Z. Wang, A. Novikov, M. Yang, M. R. Zhang, Y. Chen, A. Kumar, C. Paduraru, S. Levine, and T. L. Paine. Benchmarks for deep off-policy evaluation, 2021. URL https://arxiv.org/abs/2103.16596. International Conference on Learning Representations (ICLR) 2021.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Gemini Robotics Team, K. Choromanski, C. Devin, Y. Du, D. Dwibedi, R. Gao, A. Jindal, T. Kipf, S. Kirmani, I. Leal, F. Liu, A. Majumdar, A. Marmon, C. Parada, Y. Rubanova, D. Shah, V. Sindhwani, J. Tan, F. Xia, T. Xiao, S. Yang, W. Yu, and A. Zhou. Evaluating Gemini robotics policies in a Veo world simulator, 2026. URL https://arxiv.org/abs/2512.10675.

C. Grimm, A. Barreto, S. Singh, and D. Silver. The value equivalence principle for model-based reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5541–5552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3bb585ea00014b0e3ebe4c6dd165a358-Paper.pdf.

Y. Guo, L. X. Shi, J. Chen, and C. Finn. Ctrl-World: A controllable generative world model for robot manipulation, 2026. URL https://arxiv.org/abs/2510.10125. ICLR 2026.

A. H. Güzel, J. Seidenschwarz, B. Graham, J. Sadeghi, J. Hawke, and I. Bogunovic. PROWL: Prioritized regret-driven optimization for world model learning, 2026. URL https://arxiv.org/abs/2605.18803.

D. Ha and J. Schmidhuber. World models, 2018.

D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640 (8059):647–653, Apr. 2025a. doi: 10.1038/s41586-025-08744-2.

D. Hafner, W. Yan, and T. Lillicrap. Training agents inside of scalable world models, 2025b. URL https://arxiv.org/abs/2509.24527.

N. Hansen. TD-MPC2 model checkpoints. Hugging Face model repository, https://huggingface.co/nicklashansen/tdmpc2, 2024. MIT licence. Accessed 9 November 2026.

N. Hansen, H. Su, and X. Wang. TD-MPC2: Scalable, robust world models for continuous control, 2024. URL https://arxiv.org/abs/2310.16828. Published at the International Conference on Learning Representations (ICLR) 2024.

J. Hawke. Introducing PROWL-1: Learning through discovery. Odyssey company post, not archival, https://odyssey.ml/introducing-prowl-1, May 2026.

X. Huang, Z. Li, G. He, M. Zhou, and E. Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion, 2025. URL https://arxiv.org/abs/2506.08009. NeurIPS 2025 (spotlight).

N. Jakobi, P. Husbands, and I. Harvey. *Noise and the reality gap: The use of simulation in evolutionary robotics*, volume 929 of *Lecture Notes in Computer Science*, pages 704–720. Springer Berlin Heidelberg, 1995. ISBN 9783540492863. doi: 10.1007/3-540-59496-5_337.

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

N. Jiang, A. Kulesza, S. Singh, and R. Lewis. The dependence of effective planning horizon on model accuracy. In R. H. Bordini, E. Elkind, G. Weiss, and P. Yolum, editors, *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*, pages 1181–1189, Istanbul, Turkey, May 2015. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-3413-6. URL https://www.ifaamas.org/Proceedings/aamas2015/aamas/p1181.pdf.

Z. Jiang, S. Zhou, Y. Jiang, Z. Huang, M. Wei, Y. Chen, T. Zhou, Z. Guo, H. Lin, Q. Zhang, Y. Wang, H. Li, D. Zhao, and C. Yu. WoVR: World models as reliable simulators for post-training VLA policies with RL, 2026. URL https://arxiv.org/abs/2602.13977.

L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski. Model-based reinforcement learning for Atari, 2020. URL https://arxiv.org/abs/1903.00374. International Conference on Learning Representations (ICLR) 2020.

A. Khassanov. When the world model lies: Measuring and characterising reward exploitation in DreamerV3 under sparse feedback, 2026. Unreviewed preprint on Zenodo.

R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims. MOReL : Model-based offline reinforcement learning, 2021. URL https://arxiv.org/abs/2005.05951. NeurIPS 2020.

J. Kieto. Learning in dreams, winning in reality: A continuous dyna loop for a ten-hero MOBA, 2026. URL https://arxiv.org/abs/2610.08033.

S. Kim, D. Kang, T. Kwon, H. Chae, D. Lee, and J. Yeo. Rethinking reward model evaluation through the lens of reward overoptimization, 2025. URL https://arxiv.org/abs/2505.12763. ACL 2025.

Z. Kong, Z. Xu, H. Zhang, J. Yang, Z. Li, Y. Ru, L. Xiang, W. Xu, M. Ren, and Z. He. Trusting imagined futures: Auditing imagined rewards for VLA post-training, 2026. URL https://arxiv.org/abs/2611.04873.

S. Koos, J.-B. Mouret, and S. Doncieux. The transferability approach: Crossing the reality gap in evolutionary robotics. *IEEE Transactions on Evolutionary Computation*, 17(1): 122–145, Feb. 2013. doi: 10.1109/tevc.2012.2185849.

A. Kumar and S. Levine. Model inversion networks for model-based optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5126–5137. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf.

T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel. Model-ensemble trust-region policy optimization, 2018.

URL https://arxiv.org/abs/1802.10592. International Conference on Learning Representations (ICLR) 2018.

N. Lambert, B. Amos, O. Yadan, and R. Calandra. Objective mismatch in model-based reinforcement learning. In A. M. Bayen, A. Jadbabaie, G. Pappas, P. A. Parrilo, B. Recht, C. Tomlin, and M. Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 761–770. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

A. Li, H. Yu, Z. Tao, L. Chen, K. Li, M. Li, C. Cai, B. Liang, J. You, and K.-F. Wong. CivWorldBench: Evaluating language world models in structured multi-agent sandbox worlds, 2026. URL https://arxiv.org/abs/2610.16450.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545. CVPR 2022.

C. Lu, P. J. Ball, J. Parker-Holder, M. A. Osborne, and S. J. Roberts. Revisiting design choices in offline model-based reinforcement learning, 2022. URL https://arxiv.org/abs/2110.04135. International Conference on Learning Representations (ICLR) 2022 (spotlight).

S. Ma and W. Zhao. When may a model replace the experiment? audits, licenses, and the price of trust in surrogate-driven design, 2026. URL https://arxiv.org/abs/2608.01378.

L. Maes, Q. L. Lidec, D. Scieur, Y. LeCun, and R. Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312. NeurIPS 2026 (poster).

D. Manheim and S. Garrabrant. Categorizing variants of Goodhart's law, 2019. URL https://arxiv.org/abs/1803.04585.

V. Micheli, E. Alonso, and F. Fleuret. Transformers are sample-efficient world models, 2023. URL https://arxiv.org/abs/2209.00588. International Conference on Learning Representations (ICLR) 2023.

Micropolis GmbH. "MICROPOLIS" public name license, version 1. MicropolisPublicNameLicense.md in https://github.com/SimHacker/MicropolisCore, Oct. 2024. Accessed 9 November 2026.

A. Pan, K. Bhatia, and J. Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models, 2022. URL https://arxiv.org/abs/2201.03544. International Conference on Learning Representations (ICLR) 2022.

J. Quevedo, A. K. Sharma, Y. Sun, V. Suryavanshi, P. Liang, and S. Yang. WorldGym: World model as an environment for policy evaluation, 2025. URL https://arxiv.org/abs/2506.00613. ICLR 2026.

B. Rahmani. Debugging code world models, 2026. URL https://arxiv.org/abs/2602.07672.

Reka. WorldModelGym: a decision-based fidelity benchmark for world models. Reka news post, not archival, https://reka.ai/news/worldmodelgym, July 2026.

M. Rigter, B. Lacerda, and N. Hawes. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16082–16097. Curran Associates, Inc., 2022. doi: 10.52202/068431-1170.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

S. Ross and J. A. Bagnell. Agnostic system identification for model-based reinforcement learning. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1703–1710, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1. URL https://icml.cc/2012/papers/833.pdf.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

SimHacker. MicropolisCore: Micropolis C++ core. GitHub repository, https://github.com/SimHacker/MicropolisCore, commit bd6e12eed6e148fe25270e7a4e2278f71a89f26c, 2026. Accessed 9 November 2026.

J. Singh. The evaluation protocol determines the result: An independent reproduction of LeWorldModel on TwoRoom, 2026. URL https://arxiv.org/abs/2608.10145.

J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward gaming. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471. Curran Associates, Inc., 2022. doi: 10.52202/068431-0687.

J. E. Smith and R. L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, Mar. 2006. doi: 10.1287/mnsc.1050.0451.

G. Synnaeve, Z. Lin, J. Gehring, D. Gant, V. Mella, V. Khalidov, N. Carion, and N. Usunier. Forward modeling for partial observation strategy games - a StarCraft defogger, 2018. URL https://arxiv.org/abs/1812.00054.

E. Talvitie. Model regularization for stable sample rollouts. In N. L. Zhang and J. Tian, editors, *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, volume R12 of *Proceedings of Machine Learning Research*, pages 568–577. PMLR, 23–27 Jul 2014. URL https://proceedings.mlr.press/r12/college14a.html.

E. Talvitie. Self-correcting models for model-based reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):2597–2603, Feb. 2017. doi: 10.1609/aaai.v31i1.10850. arXiv:1612.06018.

B. Trabucco, A. Kumar, X. Geng, and S. Levine. Conservative objective models for effective offline model-based optimization. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10358–10368. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/trabucco21a.html.

B. Trabucco, X. Geng, A. Kumar, and S. Levine. Design-Bench: Benchmarks for data-driven offline model-based optimization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21658–21676. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/trabucco22a.html.

S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa.

dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, Nov. 2020. doi: 10.1016/j.simpa.2020.100022.

D. Vakalis. The intervention gap in latent world models, 2026. URL https://arxiv.org/abs/2608.29998.

D. Valevski, Y. Leviathan, M. Arar, and S. Fruchter. Diffusion models are real-time game engines, 2025. URL https://arxiv.org/abs/2408.14837. International Conference on Learning Representations (ICLR) 2025.

A. Venkatraman, M. Hebert, and J. A. Bagnell. Improving multi-step prediction of learned time series models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29 (1):3024–3030, Feb. 2015. doi: 10.1609/aaai.v29i1.9590.

X. Wen, J. Lou, Y. Lu, H. Lin, X. Yu, X. Lu, B. He, X. Han, D. Zhang, and L. Sun. Rethinking reward model evaluation: Are we barking up the wrong tree?, 2025. URL https://arxiv.org/abs/2410.05584. ICLR 2025 (spotlight).

J. Xu, X. Hu, M. Chen, P. Gao, Y. Ye, and Y. Shi. World models dream of success: Diagnosing and repairing failure insensitivity in robot world models, 2026. URL https://arxiv.org/abs/2610.09134.

T. Yang, Z. Shen, Z. Mi, Z. Zhang, J. Zhou, J. Ji, J. Dai, J. Chen, B. Chen, and Y. Yang. MiraBench: Evaluating action-conditioned reliability in robotic world models, 2026. URL https://arxiv.org/abs/2605.29360.

J. Yao, Z. Wang, Y. Huang, and L. Hu. RolloutFaith: Auditing persistent internal interventions in visual world model, 2026. URL https://arxiv.org/abs/2609.36843.

T. Yu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma. MOPO: Model-based offline policy optimization, 2020. URL https://arxiv.org/abs/2005.13239. NeurIPS 2020.

T. Yu, A. Kumar, R. Rafailov, A. Rajeswaran, S. Levine, and C. Finn. COMBO: Conservative offline model-based policy optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28954–28967. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f29a179746902e331572c483c45e5086-Paper.pdf.

Y. Yu, S. Zhang, Y. Sheng, H. Ren, and H. Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position, 2026. URL https://arxiv.org/abs/2606.15032.

M. R. Zhang, T. L. Paine, O. Nachum, C. Paduraru, G. Tucker, Z. Wang, and M. Norouzi. Autoregressive dynamics models for offline policy evaluation and optimization, 2021. URL https://arxiv.org/abs/2104.13877. International Conference on Learning Representations (ICLR) 2021.

Y. Zhang, C. Bai, B. Zhao, J. Yan, X. Li, and X. Li. Decentralized transformers with centralized aggregation are sample-efficient multi-agent world models, 2025. URL https://arxiv.org/abs/2406.15836. Transactions on Machine Learning Research.

Y. Zhang, Z. Wang, Y. Rong, H. Wang, J. Jiang, S. Xu, H. Wu, S. Zhou, and B. Xu. World models for policy refinement in StarCraft II, 2026. URL https://arxiv.org/abs/2602.14857.

S. Zhou, Z. Jiang, N. Ma, X. Fu, Y. Zheng, S. Tian, H. Li, J. Liu, C. Yu, Z. Xia, H. Li, and D. Zhao. WoVR+: World model as reliable simulator for versatile VLA policy optimization with RL, 2026. URL https://arxiv.org/abs/2610.17922.

H. Zhu, Q. Chen, Z. Kou, W. Li, F. Wang, Y. Cao, X. Su, and Y. Chen. Dual-frontier: When can an agent trust its world model?, 2026. URL https://arxiv.org/abs/2609.26293.

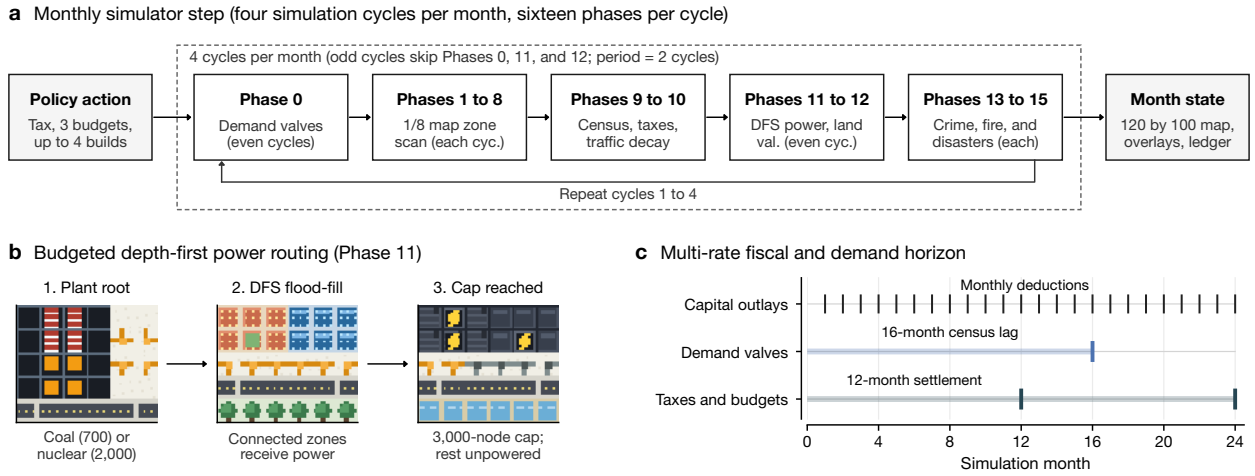


**Figure 10 | Internal monthly execution anatomy of the Micropolis C++ engine.** Constants shown are verified programmatically against the simulator source tree before rendering. **a**, Each monthly step applies the policy instruments, executes four simulation cycles of sixteen deterministic phases at speed 1 (ordered sequentially from phase zero to phase fifteen; even cycles update R/C/I demand valves, global DFS power routing, and pollution/land-value fields; each cycle scans one eighth of the 120-by-100 map West-to-East and updates transport decay, crime, population density, fire coverage, and stochastic disasters; monthly census totals update each four-cycle month and annual taxes settle at forty-eight-cycle intervals), and reads the resulting state. **b**, Power distribution uses a budgeted depth-first flood-fill over conductive tiles starting from active coal and nuclear plants pushed onto the scan stack; once plant tile budgets or stack pointers run out, downstream zones lose power (illustrated across three procedural micro-sprite stages). **c**, Non-local temporal coupling over a 24-month plan, where construction outlays deduct immediately each month from reserves, R/C/I demand valves refresh on sixteen-month census horizons, and annual tax receipts minus service budgets settle once a year.

# A. Simulator Architecture, Model Ladder, and Experimental Protocols

## A.1. Exact Simulator and Deterministic Replay

Micropolis couples cellular-automaton spatial updates to global resource constraints spanning budgeted depth-first power routing, annual fiscal accounting, and census-driven R/C/I demand valves (Figure 10). We drive the C++ core exclusively through clean episode resets that load a starting city file, set the disasters flag, read the initial observation, seed the internal PRNG once, and apply each month's tax rate, three service funding ratios, ordered spatial tool calls, and any scheduled task event before stepping four city-time units at speed 1 (one pass per tick, easy game level, auto-budget and auto-bulldoze active). Three upstream non-determinism bugs were patched and documented in the C++ source tree (an uninitialised callback pointer, heap-allocated sprite name strings, and wall-clock PRNG reseeding on map loads). Genuine simulator mechanics (including traversal-order power allocation and branch dropping when the DFS stack fills) are preserved untouched so world models face the true dynamics.

Across all 32 bundled cities stepped for 240 months with stochastic disasters enabled, full-state hashes match identically at each month across fresh processes, previously dirtied simulator instances, and two independent host architectures. The structured state exposed to the world model comprises the 120-by-100 tile-code grid, native-resolution spatial overlays, twenty municipal scalars, and a carry-over bookkeeping ledger of 27 entries plus two auxiliary maps. Restoring this observation alongside PRNG and active-sprite state reproduces the subsequent twelve months bit-for-bit across 6,000 of 6,000 test transitions. Because active rail, seaport, airport, and disaster sprites are omitted from the world-model input, zero of 1,734 sprite-free transitions differ at month one whereas 4,196 of 6,000 transitions differ overall (4,099 shifting the shared PRNG stream as vehicles tick, while month-one powered-zone counts and liquid funds still match exactly in 5,827 and 5,737 transitions), introducing mild partial observability from the world model's perspective. Evaluating our primary 50M checkpoint required 7.9 single-GPU hours (30,187,152 neural month-steps) plus 2.2 CPU hours for 11,232,678 exact engine replays (269,584,272 city-months; full compute accounting in Table 1).

## A.2. Dataset Construction and Held-Out City Splits

Initial cities derive from 100 base cities across 268 map files combining canonical Micropolis maps with procedurally generated terrains matured by scripted urban builders. Sixty cities (60 cities across 228 start files, up to four growth stages each) form the training environment pool. Our 120 held-out evaluation cities (each contributing a single starting map) are partitioned into the 40-city core benchmark (populations 3,500 to 168,060) and the 80-city expansion split (populations 3,140 to 164,000), generated from disjoint terrain seeds using the identical builder and

**Table 1** | Study compute accounting in H200-equivalent GPU hours, weighted by published throughput ratios relative to an NVIDIA H200 (12.1 rented card-hours across 23 pods lacking published ratios are omitted). Local GPU time sums training, fidelity, and audit worker logs (splitting shared-GPU intervals across active lanes; 3 early incomplete jobs, local RL reference runs, and threshold calibration replays predate worker logging). Cloud GPU time covers 130 closed pod sessions, including world-model rollouts for 359 of 360 PPO runs; category sub-rows exceed "All cloud sessions" because multi-role pods executed workloads across more than one category. CPU time sums worker-seconds across 129 audit runs (0.56 on a secondary host) plus closed-loop replanning wall-hours (three shards across two local GPUs).

| Local GPU workloads | H200-equivalent hours |
|---|---|
| World-model training (17 runs) | 29.9 |
| Trajectory-fidelity evaluation | 2.5 |
| Benchmark audits, 40-city core | 2.6 |
| Benchmark audits, 80-city expansion split | 13.6 |
| Seeded adaptive searches | 13.0 |
| Total logged local GPU compute | 61.6 |
| **Cloud GPU workloads** | **H200-equivalent hours** |
| All cloud sessions | 101.4 |
| In-imagination PPO training | 54.3 |
| Audit evaluation pods (26 of 83 completed) | 36.7 |
| Extended world-model training | 9.7 |
| Unreturned / aborted sessions (81 pods) | 7.8 |
| **CPU and wall-clock workloads** | **Hours** |
| Engine replay pool across audits (process-hours) | 434 |
| Replanning, exact-simulator controls (325 instances, wall-hours) | 6.0 |
| Replanning, world-model runs (100 instances, wall-hours) | 11.7 |

threshold-calibration pipeline (381 of 400 candidate instances retained, 291 well posed; exact-simulator random-shooting false acceptance across $10^0$–$10^3$ reads 0.013, 0.045, 0.050, and 0.063 on the expansion split vs. 0.021, 0.031, 0.057, and 0.083 on the core benchmark, with engine-side CEM passing 1 of 80 vs. 5 of 40 *treasury* cities).

Monthly actions combine fiscal levers with spatial macros (zoned blocks, linear road/rail/wire segments, civic plants, parks, and bulldozing). Training trajectories draw from seven scripted builder archetypes, random spatial editors, passive months, fiscal sweeps, and infrastructure shocks (severed power lines, demolished plants, tax spikes, forced disasters), with stochastic disasters enabled on 7,194 trajectories across 48 and 120-month rollouts. Full-scale models train on 3,000,000 transitions (49,000 trajectories) and small-data models on a stratified 300,000-transition subset; re-verifying 3,700 archived trajectories (500 from training) yielded 0 state mismatches.

### A.3. World-Model Architecture and Training

Each world model pairs a patchified ConvNeXt U-Net (Liu et al., 2022; Ronneberger et al., 2015) over the 120-by-100 grid (conditioned on municipal scalars, month index, action encoding, and a 32-dimensional Gaussian latent inferred by a learned posterior during training) with a secondary MLP that predicts scalar bookkeeping updates from map-derived census counts. Per tile, the spatial decoder outputs a binary change gate, a conditional categorical distribution over all 1,024 tile codes on changed cells, Bernoulli flag and power bits, and zero-inflated discretised-logistic distributions for spatial overlays; bounded engine quantities use censored likelihoods so rollouts require no post-hoc clipping. Deterministic identities (tax rate, city time, total population) are computed analytically. Compiled fast-path inference accelerates 50M rollouts on a single GPU from 1,522 to 2,048 month-steps per second at unit sampling temperature.

Tables 2 and 3 list architectural dimensions and run-level configurations with checkpoint SHA-256 digests. All 17 training runs use batch size 128, AdamW ($\beta$ = 0.9 and 0.98, weight decay 0.05), linear warm-up over 0.03 of steps followed by cosine decay to 0.05 of peak learning rate, gradient clipping at 1, EMA decay 0.999, linear KL annealing over 0.05 of training (32-dimensional latent), component loss weights 0.25 for half-resolution overlays and terrain; 1 for change gate, tile code, flags, power bit, eighth-resolution overlays and commercial rate; 4 for scalars, ledger, second-stage scalars and build cost, a 6-fold scalar weight on annual tax-settlement months, and a 10-fold learning-rate multiplier on changed-tile lookup tables, while the bookkeeping MLP performs 2 updates of batch 4,096 per trunk step from learning rate 0.001. One-step variational models train for 6,000,000 examples at peak rate 0.0008; pushforward runs initialise from their budget-matched one-step parent and train for 1,500,000 examples at 0.0002 over 4-month windows, teacher-forcing 0.5 of steps and otherwise unrolling stop-gradient posterior-mode predictions before evaluating the one-step ELBO against the true engine state (distinct from differentiable sequence-level self-forcing; Huang et al., 2025). Two additional seeds of the 50M full-data configuration under each objective support our 3-member ensemble (16 total runs). Model weights are withheld pending clarity on GPL derivative-work status for neural networks trained on simulator outputs; exact SHA-256 digests are archived instead.

**Table 2** | ConvNeXt U-Net + bookkeeping world-model dimensions across our three parameter scales, extracted directly from run configurations.

| Size | Stem | Stage widths | Encoder blocks | Decoder blocks | Condition | Tile head | Posterior | Bookkeeping | Parameters (M) |
|---|---|---|---|---|---|---|---|---|---|
| 15M | 64 | 128, 256, 448 | 2, 2, 4 | 2, 2 | 256 | 256 | 64 | 256 | 15.7 |
| 50M | 96 | 192, 384, 768 | 2, 3, 6 | 2, 2 | 512 | 384 | 96 | 384 | 52.9 |
| 100M | 128 | 256, 512, 1024 | 2, 3, 8 | 2, 3 | 512 | 512 | 128 | 512 | 107.9 |

**Table 3** | Complete world-model training manifest covering objective, parameter scale, dataset split and transition count, random seed, training examples, peak learning rate, gradient-accumulation micro-batch, pushforward parent checkpoint, and truncated SHA-256 digests at mid-training and full budget (the final row reports the coverage-retrained model of Section 6).

| Objective | Size | Data | Transitions | Seed | Samples | LR | Micro | Parent | Half SHA-256 | Final SHA-256 |
|---|---|---|---|---|---|---|---|---|---|---|
| one-step | 15M | small | 300,000 | 0 | 6,000,000 | 0.0008 | – | – | aa62c6793484 | 7b7a3027427d |
| pushforward | 15M | small | 300,000 | 0 | 1,500,000 | 0.0002 | – | one-step 15M small | ff79c2bf42d4 | 3fde3a11af5e |
| one-step | 15M | full | 3,000,000 | 0 | 6,000,000 | 0.0008 | – | – | 1186f3d5a897 | daa4e5e65c05 |
| pushforward | 15M | full | 3,000,000 | 0 | 1,500,000 | 0.0002 | – | one-step 15M full | d96533bdbf4f | 6bb31a6dc7a2 |
| one-step | 50M | small | 300,000 | 0 | 6,000,000 | 0.0008 | 64 | – | 266252e86d6e | 01bdbd075806 |
| pushforward | 50M | small | 300,000 | 0 | 1,500,000 | 0.0002 | 64 | one-step 50M small | 31a26b40488b | 6d459475a403 |
| one-step | 50M | full | 3,000,000 | 0 | 6,000,000 | 0.0008 | – | – | e2c8d843fd7c | 7b08be25577b |
| pushforward | 50M | full | 3,000,000 | 0 | 1,500,000 | 0.0002 | – | one-step 50M full | 136f29a695f0 | 33836137e697 |
| one-step | 100M | small | 300,000 | 0 | 6,000,000 | 0.0008 | – | – | d8ff787cea08 | efe193ac7067 |
| pushforward | 100M | small | 300,000 | 0 | 1,500,000 | 0.0002 | – | one-step 100M small | a35568554bd0 | 177911d3c113 |
| one-step | 100M | full | 3,000,000 | 0 | 6,000,000 | 0.0008 | – | – | 2f778adea97a | b36210a87f48 |
| pushforward | 100M | full | 3,000,000 | 0 | 1,500,000 | 0.0002 | – | one-step 100M full | 7154d2d1abcb | 5d764d18e17f |
| one-step | 50M | full | 3,000,000 | 1 | 6,000,000 | 0.0008 | – | – | 486470b628be | 1d4fb0602fbb |
| pushforward | 50M | full | 3,000,000 | 1 | 1,500,000 | 0.0002 | – | one-step 50M full (s1) | 271deab4d283 | 56eb987cae86 |
| one-step | 50M | full | 3,000,000 | 2 | 6,000,000 | 0.0008 | 64 | – | 18e600cf9680 | 9d9debc5dad3 |
| pushforward | 50M | full | 3,000,000 | 2 | 1,500,000 | 0.0002 | 64 | one-step 50M full (s2) | 98e462e8cf38 | 6555bb973427 |
| one-step | 50M | full + cov. | 4,200,000 | 0 | 6,000,000 | 0.0008 | – | – | 21b00a3d84fd | a4c38f8729f2 |

### A.4. Task Predicates and Threshold Calibration

Formal task predicates over 24-month trajectories span five families, where (i) *grow* requires final population $\geq \theta$ with treasury never negative; (ii) *solvent* requires final population $\geq \theta$ with treasury never below half its initial value; (iii) *treasury* requires second-year annual tax revenue $\geq \theta$, treasury never negative, and monthly tax rate never exceeding its initial value; (iv) *recover* requires final population $\geq \theta$ times pre-quake population after a forced earthquake at month one; and (v) *power* requires final zone count $\geq \theta$, at least 95 per cent of zones powered at month 24, and at most one net power plant added. Placing $\theta$ at a fixed quantile of passive behaviour trajectories would allow trivial constant-budget no-build plans to pass on individual seeds in 0.900 of *grow* cities (where 12-month population error is 0.060 vs. engine seed spread 0.027). Instead, we calibrate $\theta$ on the 32 oracle seeds midway between the highest majority score among the top 20 of 2,625 constant no-build policies and the best trajectory found across behaviour pools, budget-optimised builder pools, closed-loop scripted builders, and engine-side CEM (8,637,968 engine replays; 207.3 million city-months; Table 4).

### A.5. Exact-Simulator Reference and Statistical Inference

Pairing each world-model search with an identical run inside the exact-simulator reference on fresh seeds ensures all excess false-acceptance and ground-truth shortfall estimates are matched by instance (when oracle seeds themselves are substituted as a sanity check, false acceptance is identically zero across all 16 of 16 search-by-budget cells). In addition to random shooting and adaptive CEM across all core instances, Table 13 reports best-of-$N$ over closed-loop scripted builders across twenty core cities (99 instances). Uncertainty is estimated via 95 per cent cluster bootstraps (2,000 resamples across main search audits, and twenty thousand resamples for both nominal and Bonferroni-adjusted 99.72 per cent BCa intervals in counterfactual subsystem patching) clustering by held-out

**Table 4** | Core benchmark task instances at the 24-month horizon alongside 32-seed oracle pass rates for unoptimised behaviour-mix plans and closed-loop scripted builders under optimal constant budgets.

| Task | Instances | Well posed | Behaviour mix | Scripted builders |
|---|---|---|---|---|
| grow | 40 | 36 | 0.051 | 0.327 |
| solvent | 38 | 20 | 0.009 | 0.113 |
| treasury | 40 | 39 | 0.041 | 0.108 |
| recover | 35 | 13 | 0.055 | 0.436 |
| power | 40 | 40 | 0.125 | 0.338 |
| all | 193 | 148 | | |

**Table 5** | Counterfactual subsystem patching paired against random state-perturbation controls across all 285 exploited plan records, showing paired flip-rate excess over tensor-size-matched (*Size*) and altered-entry-matched (*Change*) controls with Bonferroni-corrected 99.72 per cent BCa intervals (twenty thousand city-cluster bootstrap resamples) across 18 comparisons, pooled and split between post-earthquake *recover* and the remaining four task families.

| | All records | | *recover* | | Other four families | |
|---|---|---|---|---|---|---|
| Patch | Size | Change | Size | Change | Size | Change |
| power | +0.049 [-0.073, +0.170] | -0.028 [-0.226, +0.136] | -0.067 [-0.250, +0.069] | -0.222 [-0.491, -0.027] | +0.153 [+0.012, +0.330] | +0.147 [-0.071, +0.328] |
| treasury | -0.014 [-0.177, +0.157] | -0.014 [-0.196, +0.155] | -0.311 [-0.468, -0.150] | -0.259 [-0.376, -0.124] | +0.253 [-0.010, +0.456] | +0.207 [-0.060, +0.407] |
| demand valves | -0.168 [-0.295, -0.016] | -0.305 [-0.479, -0.173] | -0.207 [-0.355, -0.062] | -0.415 [-0.582, -0.243] | -0.133 [-0.323, +0.101] | -0.207 [-0.442, -0.025] |
| zone growth | +0.004 [-0.137, +0.158] | -0.140 [-0.251, 0.000] | +0.185 [+0.011, +0.413] | -0.215 [-0.393, -0.023] | -0.160 [-0.344, -0.019] | -0.073 [-0.230, +0.086] |
| pollution | -0.098 [-0.227, -0.017] | -0.130 [-0.284, -0.026] | -0.022 [-0.121, +0.049] | -0.067 [-0.223, +0.037] | -0.167 [-0.400, -0.046] | -0.187 [-0.422, -0.032] |
| traffic | +0.105 [-0.008, +0.224] | -0.056 [-0.194, +0.127] | -0.007 [-0.122, +0.148] | -0.289 [-0.493, -0.057] | +0.207 [+0.041, +0.407] | +0.153 [+0.007, +0.359] |
| land value | +0.147 [+0.053, +0.287] | +0.074 [-0.031, +0.171] | +0.074 [-0.028, +0.212] | +0.074 [-0.083, +0.242] | +0.213 [+0.082, +0.398] | +0.073 [-0.014, +0.214] |
| crime | -0.088 [-0.210, +0.015] | -0.046 [-0.157, +0.041] | -0.059 [-0.250, +0.075] | -0.059 [-0.214, +0.071] | -0.113 [-0.350, +0.033] | -0.033 [-0.171, +0.077] |
| fire | +0.204 [+0.005, +0.379] | +0.063 [-0.088, +0.196] | +0.511 [+0.299, +0.737] | +0.237 [+0.052, +0.429] | -0.073 [-0.286, +0.068] | -0.093 [-0.298, +0.049] |

city (40 clusters on the core benchmark), using BCa limits by default and exact one-sided cluster-binomial bounds when zero events occur. For reinforcement learning aggregated across the model ladder, bootstrap intervals resample checkpoints and cities independently (two-way percentile bootstrap); for single-model or engine-trained evaluations, cities are resampled. Cross-model Spearman correlations in Section 5 cluster-resample checkpoints with percentile limits.

### A.6. Counterfactual Subsystem Patching Protocol

To test causal subsystem attribution on the primary model's 285 exploited plan records (220 unique plans comprising 94 from margin-ranked random shooting, 122 from verdict-ranked shooting, 46 from best-of-$N$, and 23 from unseeded CEM), we replay each plan on 8 oracle seeds and splice the $k$-th engine trajectory's state for one subsystem into the $k$-th neural rollout at the end of each month (channel masks in Appendix A.9; empty masks reproduce unpatched rollouts bit-for-bit, full masks reproduce exact engine replays). Because predicates directly read treasury or powered-zone scalars in several tasks, we also verify indirect-only pairs; on the 29 random-shooting exploits at $B = 10^3$, raw power patching flips 0.79 [0.66, 0.90], union flip rates across any global (0.93, indirect 0.72) and any local (0.83, indirect 0.83) subsystem closely track independent-redraw expectations (0.63 and 0.92 given the 0.34 unpatched redraw flip rate), and 5 local rules exceed 0.30 raw flip share (fire 0.66, land value 0.48, crime 0.34, pollution 0.31, traffic 0.31). For spatial fire and zone-growth patches, we apply a *symmetric* cell mask (overwriting cells containing a fire or zone tile in *either* the engine or the neural rollout) because one-sided fire insertion alone spuriously overturns 0.90 of true *recover* acceptances; an indirect overlay-only control that patches solely the eighth-resolution fire-coverage map without overwriting burning tile codes flips 37 of 135 exploited *recover* records (0.27 [0.17, 0.40] vs. 0.28 on unpatched redraws, and 2 of 39, 0.05, on true acceptances), demonstrating that the symmetric fire effect operates via full-resolution burning-tile propagation rather than coarse fire-coverage map drift. In Figure 1's case study (selected by maximum pass-share discrepancy, breaking ties by worst oracle margin), the audit passed 8 of 8 rollouts; upon CPU redraw 7 pass, 6 lose the demolished plant at month 8 while 5 keep zones lit and pass, and across 8 patched rollouts power-grid patching drops passes from 7 to 0 whereas all other subsystems leave at least 5 passing. Table 5 reports all paired control contrasts.

### A.7. Reinforcement-Learning Architecture and Training Protocol

Each PPO run trains an independent closed-loop policy for one world-model checkpoint, task family, and random seed using the hyperparameters in Table 6. Parallel vectorised environments (200 slots) cycle deterministically through the task family's instances so each city receives equal experience; episodes run for 24 monthly steps from the city's initial state, sampling world-model heads at unit temperature (or stepping fresh non-oracle seeds when training directly on the C++ engine).

*Policy network.* At each month the actor-critic observes the spatial state alongside 74 conditioning scalars encoding task identity, month index, initial municipal scalars, cumulative changes from month zero, and trajectory-to-date clause slacks (current continuous margin, primary statistic slack, minimum treasury ratio, auxiliary clause

**Table 6** | Shared PPO architecture and optimisation hyperparameters verified across all 360 world-model runs and engine-trained reference runs.

| Setting | Value | Setting | Value |
|---|---|---|---|
| Environments per run | 200 | Discount $\gamma$ | 1 |
| Months per environment and update | 25 | GAE $\lambda$ | 0.95 |
| Months per update | 5,000 | Entropy coefficient | 0.003, linear to zero |
| Updates in 1,000,000 months | 200 | Value-loss coefficient | 0.5 |
| Adam learning rate | 0.0003, linear to zero | Gradient-norm clip | 0.5 |
| Adam $\epsilon$ | 0.00001 | KL target | 0.02 |
| Epochs per update | 4 | KL early stop (multiple of the target) | 1.5 |
| Minibatch (months) | 1,000 | Logged-marginal weight in the first logits | 0.9 |
| Clipping range | 0.2 | Policy parameters | 1,364,009 |

satisfaction, and $\theta$). Tile codes pass through an 8-dimensional embedding and two convolutional stages (32 and 48 channels), merging with a 256-unit scalar trunk into two shared 256-unit layers that output a scalar value estimate and 1,136 action logits factored across 24 categorical heads covering tax rate, three funding ratios, and $(\text{kind}, x, y, a, b)$ for each of the four monthly construction macros. Spatial $x$ and $y$ logits incorporate log-sum-exp-marginalised readout maps over 4-by-4 tile cells zero-initialised at the output layer. Output biases are initialised toward a 0.9-weighted mixture of empirical action marginals (estimated across the first 200 logged plans per start) and a uniform prior, ensuring identical initial policies across all checkpoints and engine runs (1,364,009 total parameters).

*Potential-based reward and updates.* Monthly reward equals the one-step change in continuous task margin, telescoping episode return to final minus initial margin. Clause slacks are scaled in natural units, with target statistics in behaviour-mix standard deviations, treasury as a fraction of initial funds, tax rate over its range, powered-zone fraction in 5-percentage-point units, and plant counts in plants (on *treasury*, tax logits exceeding the city's initial rate are masked during both training and evaluation). Each iteration collects 25 steps per environment (5,000 transitions, roughly 208 episodes), bootstraps values across episode cuts via GAE, normalises minibatch advantages, halts minibatch epochs early if approximate KL exceeds 1.5 times the target KL, and completes 200 updates over 1,000,000 environment steps.

*Evaluation.* At 250,000, 500,000 and 1,000,000 training steps, each policy executes 8 stochastic episodes per instance inside its training world model (*dream pass rate*), replays those exact action logs open loop across all 32 reserved oracle seeds (*ground-truth open-loop pass rate* and continuous margin inflation), and executes 8 additional closed-loop episodes directly on fresh C++ engine seeds (*closed-loop pass rate*). Engine-trained reference policies (1 or 3 seeds per family) use the identical network, initialisation, and evaluation pipeline.

### A.8. Search Algorithms and Inference-Time Defences

*Logged candidate pools and random shooting.* Each held-out city start (across both the 40-city core and 80-city expansion split) is paired with 1,000 open-loop candidate plans generated on the engine from the behaviour mix (with infrastructure perturbations at probability 0.3), executed closed loop with random disasters disabled and stored with forced-disaster flags cleared. Random shooting at budget $B$ evaluates the first $B$ logged plans using the mean margin over 2 rollouts (nesting budgets deterministically) and selects the top-scoring candidate; best-of-$N$ scores closed-loop scripted-builder trajectories drawn uniformly across all seven builder archetypes.

*Adaptive seeded cross-entropy search (CEM).* Generation zero evaluates the first $N$ logged plans via random shooting. Each subsequent generation selects the top $\max(2, \text{ceil}(0.1\,N))$ elites across all plans evaluated thus far (20 elites at $B = 10^3$) and updates independent per-column distributions across all 600 plan columns (categorical over macro kinds and rounded, range-clipped Gaussians elsewhere, holding forced-disaster columns at zero), moving 0.7 toward elite marginals from initial logged-pool marginals (with categorical floor 0.01 and standard-deviation floor at the larger of 0.5 and 0.02 of column range). Offspring copy a uniformly chosen elite and resample each month independently with probability 0.25 before canonicalisation, running schedules of 2 of 5, 4 of 25, and 5 of 200 generations by population at $10^1$–$10^3$ and 10 of 1,000 at $10^4$.

*Ensemble scoring and defence rules.* Our 3-member ensemble draws 2 search rollouts and 8 verdict rollouts per member (6 and 24 total rollouts on independent PRNG streams), measuring epistemic disagreement via the sample standard deviation of member-mean margins. Temperature scaling multiplies Gaussian/logistic scales and divides discrete logits by $T$. Disagreement penalties score plans by mean margin minus $\lambda$ times ensemble disagreement and require both majority pass share and non-negative penalised verdict margin (hence $\lambda = 0$ isolates the margin-sign filter and counts among the 12 tested configurations). Uncertainty halts reject plans whose disagreement exceeds the 95th and 80th percentiles of the first 1,000 logged plans. Cross-fitted logistic OOD detectors fit Newton-optimised weights on five standardised ensemble features (mean margin, pass share, member disagreement, monthly trajectory spread, and active-macro fraction) over the first 125 logged plans across 2 alternating

city folds, gating both search and verdict phases. Empirical-Bayes shrinkage (Smith and Winkler, 2006) estimates oracle base rate $\bar{\mu}$, true pass-share variance $\sigma_\mu^2$, and mean-squared model-oracle discrepancy $\sigma_v^2$ on the same 125 plans, shrinking predicted pass share $p$ toward $\bar{\mu}$ with weight $\alpha = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_v^2)$.

### A.9. Fidelity Metrics, Splits, Patch Masks, and Deterministic Seeding

*Structural changed-tile F1.* One-month changed-tile F1 pools true positives (both predicted and ground-truth tile codes change from the preceding month and match), false positives, and false negatives across transitions via $2\,tp/(2\,tp + fp + fn)$. Structural F1 first maps all 1,024 raw engine tile codes onto 791 equivalence classes derived from Micropolis animation tables and road-traffic density variants (233 codes merged across 37 multi-code classes, filtering out 0.503 non-structural raw code flips across the 72,000 held-out transitions).

*Held-out fidelity splits and nearest-neighbour baseline.* Each core held-out city (strictly disjoint from all training cities) provides 25 trajectories of 48 months and 5 of 120 months (1,200 total trajectories). Passive logging during training used a development draw of 2,000 one-month transitions and 500 rollout trajectories (seed 20261109); development metrics were never used for early stopping, checkpoint selection, or hyperparameter tuning, as all runs trained for a fixed step budget. Final reporting uses the test split (seed 20261123) comprising all 700 remaining trajectories rolled out for 24 months (sharing zero trajectories with development rollouts) and 2,000 one-month transition pairs sampled outside development pairs (502 development pairs fall inside early windows of test rollout trajectories, sharing zero transition pairs). Our tabular nearest-neighbour baseline conditions on (current code, applied tool, power bit) across 20,224 training transitions with hierarchical backoff.

*Subsystem channel masks and cryptographic seeding.* Patching partitions state tensors into nine masks spanning three global subsystems (power covering the power flag plane, grid map, and powered/unpowered zone counts; treasury covering funds, tax revenue, and three service spends; and R/C/I demand valves) and six local rules covering zone growth (zone tile codes), pollution, land value, and crime (half-resolution overlays), traffic (overlay plus road tile codes), and fire (eighth-resolution fire coverage plus burning tile codes). All random seeds in the codebase are deterministic 31-bit prefixes of domain-separated SHA-256 digests, enforcing strict disjointness between the 32 oracle seeds, the nine surrogate-floor seeds, exact-simulator reference draws, RL environment episodes, and training data splits.

## B. Extended Empirical Results, Spatial Error Maps, and Ablations

### B.1. Rollout Drift and Likelihood Across the Model Ladder

All 12 mid-training world-model checkpoints pass the same short-horizon quality thresholds as the full-budget runs (12 of 12), as one-month structural changed-tile F1 spans 0.147 to 0.174 with 12-month median relative errors of 0.040–0.054 on population and 0.032–0.057 on municipal funds (even at month 12, static start-state copying incurs just 0.112 relative population error). On the primary checkpoint at month 24, the seven non-land value spatial overlays span 0.999 to 1.520 times copy error (vs. 0.125–0.845 for cross-seed engine replays, with fire coverage at 1.127 vs. 0.125). Across all 12 final checkpoints on the ladder, 4 to 7 of the 8 spatial overlays exhibit 24-month errors no smaller than copying the initial state (at least 5 for each pushforward-tuned model), pollution error is 1.270 to 1.893 times the copy baseline, and open-loop powered-zone count error spans 2.2 to 11.0 zones (Appendix Table 17). In held-out log-likelihood, all 6 small-data final checkpoints degrade relative to their mid-training snapshots by 2 to 26 per cent due to overfitting, whereas all full-data models continue to improve. Stored logged-plan verdicts across 13 of the 24 checkpoints show the same task-family polarity throughout the ladder, rejecting 0.235 to 0.809 of passing *grow* plans and 0.677 to 1.000 on *solvent*, while falsely accepting 0.043 to 0.210 of failing *treasury* plans and 0.142 to 0.396 on *recover*.

### B.2. Sixfold Extended One-Step Training Control

To test whether the primary two-epoch checkpoint's 24-month rollout drift, planning false acceptance, and downstream discovery shortfall stem simply from early stopping on the 3,000,000-transition store, we trained a single-seed 50M full-data one-step world model from scratch for 36,000,000 examples (12 passes over 3,000,000 transitions under a stretched cosine schedule, final SHA-256 prefix `bf86bcc43263`; Table 7), evaluating milestones at 6,000,000-example intervals. Held-out variational log-likelihood improves monotonically across all six milestones from -2610.4 (primary two-epoch checkpoint) to -2375.4 nats on the test split (-2424.5 on development; terminal relative loss slope -0.031 per cent per hundred thousand examples, gaining 14.2 nats over the final two epochs as latent KL rises from 14.6 to 20.3 nats), while 12-month median relative funds error drops from 0.059 at two epochs (0.035 on the primary schedule) to 0.028 and one-month structural F1 reaches 0.183 [0.155, 0.213]. By contrast, 24-month mean spatial-overlay copy ratio worsens monotonically from 0.947 to 1.139 (6 of 8 overlays exceeding copy error, vs. 2–6 across three two-epoch seeds) and open-loop powered-zone error climbs to 11.5 tiles (2.2–6.6 across three two-epoch seeds, and 8.1 to 12.7 on training cities), shifting open-loop spatial rollouts

pessimistic. Across the six milestones on all 120 held-out cities (574 instances), single-plan ($B = 10^0$) excess false acceptance remains strictly positive through 8 epochs (24,000,000 examples, +0.019 [+0.001, +0.039]) before crossing zero at 10–12 epochs (+0.003 [-0.012, +0.019] and -0.002 [-0.017, +0.011]) as model self-acceptance falls from 0.127 to 0.021 and false rejection climbs from 0.567 to 0.867 (on *treasury*, single-plan excess sits at +0.050 [+0.017, +0.100] with 0.800 false rejection). Replaying Figure 1's coal-plant demolition blackout plan on CPU confirms that the 12-epoch model still passes 7 of 8 rollouts (matching the primary checkpoint), whereas across Figure 2's 15 showcase plans model acceptance falls monotonically from 12 of 15 at two epochs to 2 of 15 at 12 epochs (both remaining acceptances on *power*). At 12 epochs across all 120 cities, raw false acceptance falls to 0.014 at $10^0$, 0.031 under $10^3$ random shooting (excess -0.038 [-0.067, -0.012]; false rejection 0.662), and 0.026 under $10^3$ adaptive CEM (excess -0.056 [-0.087, -0.028]; false rejection 0.648; core CEM excess -0.041 [-0.083, -0.005] at 0.707 false rejection); pooled $10^3$ ground-truth pass-rate shifts are small (+0.024 [-0.009, +0.059] and +0.019 [-0.019, +0.054]) because opposite family shifts cancel under CEM (+0.100 [+0.008, +0.192] on *treasury* and +0.143 [+0.050, +0.244] on *power* vs. -0.152 [-0.238, -0.076] on *recover*). At $B = 10^4$ on the 193 core instances, excess false acceptance drops to -0.036 [-0.081, +0.011] as false rejection reaches 0.639 [0.517, 0.746] (model approval 0.150) and self-acceptance collapses to zero on 3 of five task families (*grow*, *solvent*, and *recover*, with *recover* falling from 0.714 to 0.000 self-acceptance and 0.257 to 0.143 oracle pass, *treasury* excess dropping from +0.175 to 0.000 at 0.600 false rejection, and *power* oracle pass rising from 0.525 to 0.750, paired gain +0.225 [+0.025, +0.400], while topological excess remains +0.125 [+0.050, +0.275] vs. +0.300 at two epochs). Pooled ground-truth pass rate still plateaus from 0.301 at $10^3$ to 0.316 [0.256, 0.387] at $10^4$ (budget gain +0.016 [-0.046, +0.077]; paired shift vs. two epochs +0.047 [-0.016, +0.119]), leaving a 0.264 [0.189, 0.333] shortfall below exact-simulator search. Across 15 PPO runs (3 seeds per family), scalar-governed *treasury* reverses optimism (0.340 imagined vs. 0.423 oracle) and gains +0.083 [+0.033, +0.135] open-loop and +0.065 [+0.027, +0.111] closed-loop ground-truth success (retaining +0.075 [+0.008, +0.175] initial-plan verdict excess), whereas the 12 spatial-family runs lose -0.045 [-0.069, -0.028] open-loop and -0.070 [-0.092, -0.051] closed-loop success as imagined pass rates collapse from 0.189 to 0.009 (pooled closed-loop PPO falls from 0.202 to 0.159). Including this checkpoint across 25 checkpoints (125 pairs) yields +0.702 [+0.549, +0.823] within-family and +0.876 [+0.830, +0.912] pooled Spearman correlation for single-plan calibration against PPO margin inflation (vs. +0.587 and +0.851 for $10^3$ shooting).

### B.3. Optimisation Budget Scaling and Cross-Split Consistency

Table 8 details adaptive CEM scaling from $B = 10^1$ to $B = 10^4$ alongside multiplicity-adjusted budget contrasts (relative shortfall spans 48, 45, 41, and 54 per cent from $10^1$ to $10^4$), and Table 9 gives complete split-by-split numbers across the 40-city core benchmark, the disjoint 80-city expansion split, and all 120 held-out cities combined (core random-shooting budget shift +0.005 [-0.056, +0.065]). On the core benchmark under random shooting, real-success shortfall climbs from zero at $10^0$ to 0.083 at $10^1$ and 0.181 at $10^3$ (0.311 [0.255, 0.380] vs. 0.492 [0.426, 0.574]; well-posed excess reads +0.068 at $10^0$ and +0.081 [+0.020, +0.147] at $10^3$), while 15 of 29 false acceptances pass fewer than 0.1 of oracle seeds and ten sit between three tenths and one half (median pass share 0.094, vs. 0.281 and 3 of 16 for the exact-simulator reference).

### B.4. Reinforcement Learning by Task Family and Training Horizon

Table 10 disaggregates our 360 in-imagination PPO runs across task families at $10^6$ steps and tracks progression across training snapshots (on well-posed instances, imagined vs. oracle pass rates read 0.288 vs. 0.121; engine-trained PPO passes 0.307 of training episodes alongside 0.230 on the oracle).

### B.5. Targeted Coverage Retraining Details and Seed Spread

In our baseline 3,000,000-transition store, only 0.069 of active building months set tax at 4 per cent or below and 0.002 zero out all three municipal services, compared with 0.644 and 0.441 among engine-optimised passing trajectories. Our expanded 4,200,000-transition coverage store adds 600,000 transitions under uniformly sampled fiscal grids and 600,000 transitions under the top three task-competent constant budgets per training city (injecting a month-one earthquake at five-trajectory intervals; all 50,000 added trajectories verified bit-exact), raising low-tax and zero-funding build shares to 0.204 and 0.066. Short-horizon metrics remain virtually identical to the primary checkpoint, with one-month structural F1 0.176 vs. 0.179, population error 0.049 vs. 0.060, and treasury error 0.056 vs. 0.035.

Under adaptive CEM at $B = 10^3$, raw false acceptance drops from 0.181 [0.130, 0.229] to 0.098 [0.066, 0.141] (exact-simulator reference 0.067; well-posed excess drops from +0.108 [+0.048, +0.169] to +0.061 [+0.007, +0.119]), as *treasury* falls from 0.275 to 0.100, *recover* from 0.371 to 0.143, and *solvent* excess shifts by -0.053, whereas *power* reads 0.250 vs. 0.225 (reference 0.000). Random-shooting excess similarly declines from +0.067 [+0.010, +0.125] to +0.016 [-0.031, +0.065] (paired difference -0.052 [-0.107, 0.000]), model self-acceptance

**Table 7** | Sixfold extended one-step training control (50M full-data model trained for 36,000,000 examples, 12 epochs) compared against the primary two-epoch 50M checkpoint and the exact-simulator reference. First panel tracks held-out test variational ELBO (nats/transition), short-horizon F1, 12-month scalar errors, 24-month spatial-overlay copy-error ratios, open-loop powered-zone MAE (copying incurs 0.5 tiles; engine cross-seed spread 0.0), 120-city single-plan ($B = 10^0$) excess and false rejection/self-acceptance, and SHA-256 prefixes across milestones. Second panel reports pooled paired planning audits across all 120 held-out cities (574 instances) and at $B = 10^4$ on the 40-city core (193 instances). Third panel disaggregates $B = 10^4$ adaptive CEM by task family across the 193 core instances. Fourth panel disaggregates $10^6$-step in-imagination PPO across *treasury*, the four spatial task families, and all five families combined (95 per cent city-cluster BCa intervals).

| Run | Examples | Ep. | ELBO | 1m F1 | 12m pop | 12m funds | 24m ov. (≥ copy) | 24m pow. | $10^0$ excess | $10^0$ rej./acc. | SHA-256 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Primary (50M) | 6,000,000 | 2 | -2610.4 | 0.179 | 0.060 | 0.035 | 1.058 (6/8) | 6.1 | +0.077 [+0.052, +0.105] | 0.533/0.117 | 7b08be2557 |
| Extended (50M) | 6,000,000 | 2 | -2643.8 | 0.180 | 0.052 | 0.059 | 0.947 (3/8) | 5.2 | +0.089 [+0.066, +0.117] | 0.567/0.127 | 56e4f6965b |
| | 12,000,000 | 4 | -2602.1 | 0.181 | 0.051 | 0.054 | 0.974 (3/8) | 9.6 | +0.061 [+0.037, +0.089] | 0.667/0.094 | 512dc047e8 |
| | 18,000,000 | 6 | -2527.6 | 0.181 | 0.056 | 0.046 | 0.999 (3/8) | 8.6 | +0.031 [+0.010, +0.052] | 0.733/0.061 | 19ecbf9de8 |
| | 24,000,000 | 8 | -2435.9 | 0.180 | 0.058 | 0.040 | 1.045 (6/8) | 8.7 | +0.019 [+0.001, +0.039] | 0.700/0.051 | 89600bf2ba |
| | 30,000,000 | 10 | -2389.6 | 0.182 | 0.059 | 0.035 | 1.088 (6/8) | 11.4 | +0.003 [-0.012, +0.019] | 0.800/0.030 | cbdb3f307c |
| | 36,000,000 | 12 | -2375.4 | 0.183 | 0.065 | 0.028 | 1.139 (6/8) | 11.5 | -0.002 [-0.017, +0.011] | 0.867/0.021 | bf86bcc432 |

| Planning regime | Model | False acceptance | Ref. | Excess | Truly passing | Ref. | False rej. / acc. |
|---|---|---|---|---|---|---|---|
| Single plan ($10^0$, 120 cities) | Primary (2 ep.) | 0.092 [0.071, 0.121] | 0.016 | +0.077 [+0.052, +0.105] | 0.052 [0.033, 0.084] | 0.052 | 0.533 / 0.117 |
| | Extended (12 ep.) | 0.014 [0.007, 0.025] | 0.016 | -0.002 [-0.017, +0.011] | 0.052 [0.033, 0.084] | 0.052 | 0.867 / 0.021 |
| Random ($10^3$, 120 cities) | Primary (2 ep.) | 0.162 [0.134, 0.194] | 0.070 | +0.092 [+0.059, +0.128] | 0.336 [0.298, 0.378] | 0.516 | 0.306 / 0.395 |
| | Extended (12 ep.) | 0.031 [0.019, 0.047] | 0.070 | -0.038 [-0.067, -0.012] | 0.361 [0.326, 0.400] | 0.516 | 0.662 / 0.153 |
| Adaptive CEM ($10^3$, 120 cities) | Primary (2 ep.) | 0.176 [0.144, 0.208] | 0.082 | +0.094 [+0.051, +0.133] | 0.287 [0.252, 0.328] | 0.495 | 0.248 / 0.392 |
| | Extended (12 ep.) | 0.026 [0.016, 0.040] | 0.082 | -0.056 [-0.087, -0.028] | 0.307 [0.275, 0.342] | 0.495 | 0.648 / 0.134 |
| Adaptive CEM ($10^4$, 40-city core) | Primary (2 ep.) | 0.207 [0.149, 0.265] | 0.073 | +0.135 [+0.068, +0.200] | 0.269 [0.207, 0.339] | 0.580 | 0.346 / 0.383 |
| | Extended (12 ep.) | 0.036 [0.015, 0.068] | 0.073 | -0.036 [-0.081, +0.011] | 0.316 [0.256, 0.387] | 0.580 | 0.639 / 0.150 |

| Family ($10^4$ CEM) | Acc. (2/12) | Excess (2 ep.) | Excess (12 ep.) | Oracle (2/12/ref.) | Oracle diff. | Rej. (2/12) |
|---|---|---|---|---|---|---|
| *grow* (40) | 0.050/0.000 | 0.000 [-0.100, +0.100] | -0.050 [-0.175, 0.000] | 0.175/0.175/0.400 | 0.000 [-0.125, +0.125] | 1.000/1.000 |
| *solvent* (38) | 0.026/0.000 | -0.053 [-0.158, +0.026] | -0.079 [-0.211, -0.026] | 0.105/0.105/0.421 | 0.000 [-0.079, +0.079] | 1.000/1.000 |
| *treasury* (40) | 0.375/0.200 | +0.175 [+0.025, +0.332] | 0.000 [-0.100, +0.100] | 0.275/0.375/0.550 | +0.100 [-0.075, +0.275] | 0.455/0.600 |
| *recover* (35) | 0.714/0.000 | +0.257 [+0.057, +0.457] | -0.200 [-0.343, -0.086] | 0.257/0.143/0.514 | -0.114 [-0.257, 0.000] | 0.000/1.000 |
| *power* (40) | 0.775/0.525 | +0.300 [+0.175, +0.450] | +0.125 [+0.050, +0.275] | 0.525/0.750/1.000 | +0.225 [+0.025, +0.400] | 0.095/0.467 |

| PPO scope ($10^6$ steps) | Model | Dream | Oracle pass | Optimism gap | Closed loop | Oracle diff. |
|---|---|---|---|---|---|---|
| *treasury* (3 seeds) | Primary (2 ep.) | 0.599 | 0.340 [0.251, 0.442] | +0.259 [+0.158, +0.356] | 0.345 [0.273, 0.424] | |
| | Extended (12 ep.) | 0.340 | 0.423 [0.323, 0.523] | -0.083 [-0.158, -0.003] | 0.409 [0.315, 0.500] | +0.083 [+0.033, +0.135] |
| Spatial families (12 seeds) | Primary (2 ep.) | 0.189 | 0.087 [0.057, 0.135] | +0.102 [+0.047, +0.150] | 0.166 [0.131, 0.215] | |
| | Extended (12 ep.) | 0.009 | 0.043 [0.023, 0.073] | -0.034 [-0.061, -0.014] | 0.096 [0.072, 0.131] | -0.045 [-0.069, -0.028] |
| All five families (15 seeds) | Primary (2 ep.) | 0.271 | 0.138 [0.109, 0.176] | +0.133 [+0.080, +0.182] | 0.202 [0.167, 0.245] | |
| | Extended (12 ep.) | 0.075 | 0.119 [0.094, 0.146] | -0.044 [-0.069, -0.019] | 0.159 [0.130, 0.191] | -0.019 [-0.041, -0.002] |

moves from 0.399 to 0.280, and logged-plan rank correlation rises from 0.669 to 0.696 (Table 11). Under PPO, net optimism excess declines from +0.057 [+0.007, +0.108] to -0.018 [-0.062, +0.026] (-0.183 [-0.258, -0.117] on *treasury*; -0.177 [-0.246, -0.104] on *recover*, residual +0.088 [-0.006, +0.181]; *power* sits near zero at -0.012 and +0.001), while open-loop and closed-loop oracle pass rates remain essentially unchanged at 0.150 vs. 0.138 and 0.209 vs. 0.202 (*treasury* shifts from 0.340 to 0.414, *recover* from 0.143 to 0.114, across 3 seeds with within-cell SD 0.009–0.053).

### B.6. Inference-Time Defences Across All Settings

Across 60 instances (39 cities), our single-model regularisers test temperatures 0.75, 1.25 and 1.5 and empirical-Bayes shrinkage, while our 3-member ensemble tests disagreement penalties at weights 1 and 2 plus zero-weight margin filtering, uncertainty halts at the 95th and 80th logged-plan disagreement percentiles, cross-fitted OOD gates at thresholds 0.5 and 0.2, and ensemble shrinkage (single-model and exact-simulator evaluations draw 2 search and 8 verdict rollouts vs. 6 and 24 across the ensemble). Under random shooting at $B = 10^3$, the single model records 0.133 false acceptance and 0.292 false rejection (exact reference records 0.050 and 0.088; single-model defences shift false acceptance by -0.133 to +0.050 and false rejection by -0.014 to +0.708). On the ensemble, penalty weights 1 and 2 yield 0.083 and 0.050 false acceptance at 0.739 and 0.769 false rejection; strict OOD gating reaches 0.033 at 0.720; and ensemble shrinkage reaches 0.067 at 0.615 (the looser halt coincides with the undefended ensemble because no shooting winner exceeds its threshold). Under adaptive CEM, the undefended ensemble records 0.217 [0.119, 0.314] false acceptance and 0.059 false rejection (single model records

**Table 8** | Adaptive CEM scaling across evaluation budgets $10^1$, $10^2$, $10^3$, and $10^4$ on all 193 core benchmark instances of the primary 50M checkpoint (40, 38, 40, 35, and 40 across the five task families), paired against the exact-simulator reference alongside random shooting on the identical instances. The world model approves its own selected plan on 0.399 of instances at $10^3$ and 0.383 at $10^4$ (on the 148 well-posed instances at $10^4$, excess false acceptance is +0.142 [+0.084, +0.208] and ground-truth shortfall is 0.345). Bottom panel reports instance-paired budget contrasts with unadjusted 95 per cent intervals and Bonferroni-corrected 98.33 per cent intervals.

| Search | $B$ | False acceptance | Exact ref. | Excess | Truly passing | Exact ref. | Shortfall |
|---|---|---|---|---|---|---|---|
| Adaptive | $10^1$ | 0.109 [0.072, 0.153] | 0.026 | +0.083 [+0.037, +0.130] | 0.067 [0.041, 0.109] | 0.130 | 0.062 [0.026, 0.119] |
| | $10^2$ | 0.145 [0.104, 0.189] | 0.052 | +0.093 [+0.037, +0.144] | 0.192 [0.147, 0.258] | 0.347 | 0.155 [0.092, 0.223] |
| | $10^3$ | 0.181 [0.130, 0.229] | 0.067 | +0.114 [+0.052, +0.173] | 0.275 [0.220, 0.344] | 0.466 | 0.192 [0.143, 0.261] |
| | $10^4$ | 0.207 [0.149, 0.265] | 0.073 | +0.135 [+0.068, +0.200] | 0.269 [0.207, 0.339] | 0.580 | 0.311 [0.241, 0.382] |
| Random | $10^0$ | | | +0.062 [+0.026, +0.103] | 0.031 | 0.031 | |
| | $10^3$ | | | +0.067 [+0.010, +0.125] | 0.311 | 0.492 | |

| Paired contrast (adaptive CEM) | Unadjusted CI | Multiplicity-adjusted CI |
|---|---|---|
| Excess false acceptance, $10^4$ minus $10^3$ | +0.021 [-0.042, +0.092] | [-0.062, +0.107] |
| Excess false acceptance, $10^4$ minus $10^2$ | +0.041 [-0.010, +0.097] | [-0.021, +0.110] |
| Excess false acceptance, $10^4$ minus $10^1$ | +0.052 [-0.019, +0.130] | [-0.036, +0.145] |
| Ground-truth shortfall, $10^4$ minus $10^3$ | +0.119 [+0.026, +0.205] | [-0.003, +0.220] |
| Ground-truth shortfall, $10^4$ minus $10^2$ | +0.155 [+0.058, +0.240] | [+0.037, +0.265] |
| Ground-truth shortfall, $10^4$ minus $10^1$ | +0.249 [+0.172, +0.317] | [+0.158, +0.335] |
| True success inside world model, $10^4$ minus $10^3$ | -0.005 [-0.068, +0.068] | [-0.082, +0.084] |
| True success inside exact engine, $10^4$ minus $10^3$ | +0.114 [+0.067, +0.164] | [+0.057, +0.179] |

**Table 9** | Cross-split consistency on the primary 50M checkpoint across the 40-city core benchmark, the disjoint 80-city expansion split, and all 120 held-out cities combined ($n$ = task instances; cross-split differences in $10^3$ excess via independent city-cluster bootstraps read +0.038 [-0.037, +0.107] for random shooting and -0.030 [-0.109, +0.048] for adaptive CEM).

| Search | Split | $B$ | $n$ | False acceptance | Exact ref. | Excess | Truly passing, model / ref. |
|---|---|---|---|---|---|---|---|
| Random | 40-city core | $10^0$ | 193 | 0.083 [0.051, 0.123] | 0.021 | +0.062 [+0.026, +0.103] | 0.031 / 0.031 |
| | | $10^1$ | 193 | 0.124 [0.089, 0.167] | 0.031 | +0.093 [+0.048, +0.138] | 0.140 / 0.223 |
| | | $10^2$ | 193 | 0.130 [0.085, 0.173] | 0.057 | +0.073 [+0.005, +0.128] | 0.269 / 0.435 |
| | | $10^3$ | 193 | 0.150 [0.104, 0.195] | 0.083 | +0.067 [+0.010, +0.125] | 0.311 / 0.492 |
| | 80-city split | $10^0$ | 381 | 0.097 [0.069, 0.133] | 0.013 | +0.084 [+0.053, +0.120] | 0.063 / 0.063 |
| | | $10^1$ | 381 | 0.144 [0.113, 0.178] | 0.045 | +0.100 [+0.060, +0.140] | 0.231 / 0.302 |
| | | $10^2$ | 381 | 0.142 [0.111, 0.178] | 0.050 | +0.092 [+0.057, +0.132] | 0.336 / 0.472 |
| | | $10^3$ | 381 | 0.168 [0.132, 0.214] | 0.063 | +0.105 [+0.067, +0.155] | 0.349 / 0.528 |
| | all 120 cities | $10^0$ | 574 | 0.092 [0.071, 0.121] | 0.016 | +0.077 [+0.052, +0.105] | 0.052 / 0.052 |
| | | $10^1$ | 574 | 0.138 [0.114, 0.163] | 0.040 | +0.098 [+0.069, +0.129] | 0.200 / 0.275 |
| | | $10^2$ | 574 | 0.138 [0.113, 0.166] | 0.052 | +0.085 [+0.056, +0.120] | 0.314 / 0.460 |
| | | $10^3$ | 574 | 0.162 [0.134, 0.194] | 0.070 | +0.092 [+0.059, +0.128] | 0.336 / 0.516 |
| Adaptive | 40-city core | $10^3$ | 193 | 0.181 [0.130, 0.229] | 0.067 | +0.114 [+0.052, +0.173] | 0.275 / 0.466 |
| | 80-city split | $10^3$ | 381 | 0.173 [0.137, 0.216] | 0.089 | +0.084 [+0.032, +0.138] | 0.294 / 0.509 |
| | all 120 cities | $10^3$ | 574 | 0.176 [0.144, 0.208] | 0.082 | +0.094 [+0.051, +0.133] | 0.287 / 0.495 |

0.250 and 0.333; exact reference records 0.100 and 0.138), whereas zero-weight margin filtering yields 0.100 at 0.471, weight-two penalisation 0.033 at 0.727, strict OOD gating 0.050 at 0.846 (selection ROC AUC 0.805 [0.686, 0.893]), and shrinkage 0.083 at 0.412.

### B.7. Cross-Model Correlation, Split Reliability, and Power Analysis

Across the 24 checkpoints, PPO pass-rate optimism varies substantially from +0.070 (final one-step 100M full-data) to +0.351 (half-trained pushforward 15M full-data; Table 19), while continuous margin inflation is negative on three task families (-0.331 to -0.213). Because continuous task margins take minima over heterogeneous clause units, Spearman rank correlation provides a scale-free cross-model association metric (robustness checks clustering by task family yield [+0.461, +0.927], and leave-one-lineage-out jackknife spans +0.831 to +0.860). Evaluated separately by city split, pooled Spearman correlation reads +0.839 [+0.756, +0.901] (within-family mean +0.562) on the 80-city expansion split and +0.829 [+0.759, +0.884] (within-family mean +0.464) on the 40-city core. Within individual task families across all 120 cities, checkpoint rank correlations are +0.464 [+0.014, +0.779] (*grow*), +0.818 [+0.562, +0.944] (*solvent*), +0.680 [+0.258, +0.926] (*treasury*), +0.734 [+0.468, +0.875] (*recover*), and +0.068 [-0.398, +0.514] (*power*), while pooling within-family ranks yields +0.554 [+0.310, +0.724] and within-family Spearman against PPO pass-rate optimism reads +0.551 [+0.369, +0.682]. Cluster-bootstrapping the 6 four-checkpoint training lineages yields [+0.799, +0.894] for $10^3$ shooting and [+0.467, +0.836] within family ([+0.828, +0.914] pooled; +0.883 [+0.834, +0.920] against pass-rate op-

**Table 10** | In-imagination PPO across all 24 world-model checkpoints paired against engine-trained PPO. Top two panels (at $10^6$ steps by task family) report imagined episode pass rate (*Dream*), 32-seed open-loop pass rate on the identical action logs (*Oracle*), pass-rate optimism gap, engine-trained PPO stochasticity gap, net optimism excess, continuous margin inflation, engine-trained ground-truth success, and initial-plan verdict false acceptance (world model / exact reference). Bottom panel tracks pooled progression at one-quarter, one-half, and full training budget alongside closed-loop engine pass rates and open-loop ground-truth shortfall.

| Family | Dream | Oracle | Gap | Engine-trained gap | Excess |
|---|---|---|---|---|---|
| grow | 0.126 | 0.067 | +0.058 [-0.008, +0.123] | +0.081 [+0.053, +0.109] | -0.023 [-0.089, +0.039] |
| solvent | 0.127 | 0.094 | +0.033 [-0.045, +0.109] | +0.148 [+0.079, +0.220] | -0.115 [-0.206, -0.030] |
| treasury | 0.620 | 0.370 | +0.250 [+0.160, +0.340] | -0.020 [-0.054, +0.011] | +0.270 [+0.184, +0.359] |
| recover | 0.672 | 0.133 | +0.539 [+0.427, +0.648] | +0.175 [+0.111, +0.242] | +0.364 [+0.275, +0.452] |
| power | 0.052 | 0.040 | +0.012 [+0.003, +0.023] | 0.000 [0.000, 0.000] | +0.012 [+0.003, +0.023] |
| all | 0.319 | 0.141 | +0.178 [+0.127, +0.231] | +0.077 [+0.059, +0.095] | +0.101 [+0.053, +0.153] |

| Family | Margin inflation | Engine-trained real success | False acceptance, model / ref. |
|---|---|---|---|
| grow | -0.331 [-0.490, -0.178] | 0.078 | 0.085 / 0.036 |
| solvent | -0.313 [-0.450, -0.171] | 0.161 | 0.095 / 0.045 |
| treasury | +0.099 [-0.011, +0.222] | 0.620 | 0.359 / 0.053 |
| recover | +0.901 [+0.692, +1.130] | 0.162 | 0.619 / 0.183 |
| power | -0.213 [-0.343, -0.092] | 0.131 | 0.009 / 0.000 |
| all | +0.028 [-0.092, +0.149] | 0.230 | 0.233 / 0.064 |

| Training fraction | Dream | Oracle | Gap | Closed loop | Ground-truth shortfall | Engine-trained success |
|---|---|---|---|---|---|---|
| one quarter | 0.210 | 0.056 | +0.154 [+0.116, +0.195] | 0.104 | 0.023 | 0.078 |
| one half | 0.263 | 0.091 | +0.172 [+0.129, +0.219] | 0.148 | 0.045 | 0.135 |
| full ($10^6$) | 0.319 | 0.141 | +0.178 [+0.127, +0.231] | 0.202 | 0.089 [0.067, 0.114] | 0.230 |

**Table 11** | Coverage-retrained 50M world model compared against three independent training seeds of the baseline 50M configuration at $B = 10^3$ across all 193 core instances (exact-simulator false acceptance is 0.083 for random shooting and 0.067 for adaptive CEM; exact-simulator true success is 0.492 and 0.466). Paired differences (coverage model minus three-seed baseline mean) for random shooting are -0.059 [-0.098, -0.016] excess, +0.009 [-0.040, +0.054] real success, and +0.023 [-0.076, +0.116] false rejection, and for adaptive CEM are -0.076 [-0.125, -0.025], 0.000 [-0.043, +0.046], and +0.080 [-0.014, +0.189].

| | Random shooting | | | | Adaptive CEM | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Excess | Real success | False rejection | Accepts | Excess | Real success | False rejection | Accepts |
| Primary checkpoint (first seed) | +0.067 | 0.311 | 0.317 | 0.363 | +0.114 | 0.275 | 0.208 | 0.399 |
| Second baseline seed | +0.098 | 0.332 | 0.281 | 0.420 | +0.130 | 0.249 | 0.208 | 0.394 |
| Third baseline seed | +0.057 | 0.326 | 0.365 | 0.347 | +0.078 | 0.301 | 0.362 | 0.337 |
| Coverage-retrained model | +0.016 | 0.332 | 0.344 | 0.316 | +0.031 | 0.275 | 0.340 | 0.280 |

timism; differences vs. $10^3$ shooting -0.128 [-0.305, +0.059] within family and -0.032 [-0.100, +0.037] pooled) for single-plan verdict calibration. Appending the 12-epoch checkpoint yields lineage-clustered intervals [+0.468, +0.853] within family and [+0.830, +0.910] pooled (+0.872 [+0.817, +0.911] against pass-rate optimism), while held-out ELBO (-0.175) and one-month F1 (-0.190) across all 13 completed checkpoints remain weakly and negatively associated with $10^3$ false acceptance.

Monte Carlo simulation shows that within-family ranking achieves 0.98 power at a true correlation of 0.4, bounded primarily by finite-instance binomial variance, as within-family audit reliability spans 0.58 (*recover*) to 0.90 versus at least 0.95 for three-seed PPO margin inflation, attenuating a true 0.4 correlation to 0.30–0.38 in expectation (cross-split checkpoint rank agreement spans +0.580 on *grow* to +0.824, implying full-120-city Spearman–Brown reliability of 0.78 to 0.93). Adaptive CEM at $B = 10^3$ achieves +0.794 [+0.733, +0.850] against the PPO pass-rate gap (vs. +0.829 [+0.759, +0.884] for shooting on all 24 checkpoints). For task-agnostic fidelity metrics across all 24 checkpoints, continuous decision fidelity (mean Spearman correlation between 2-sample neural margins and ground-truth margins across the first 100 logged plans per core instance over 193 instances, with checkpoint range 0.574 to 0.721) correlates with expansion-split false acceptance at +0.044 pooled (+0.109 checkpoint-level; +0.104 in-split; within-family means across the six fidelity readings are -0.161, -0.115, -0.135, -0.137, +0.049, and +0.123 with within-family confidence widths 0.52–0.63; checkpoint-level held-out log-likelihood reads +0.075; simulated power at true checkpoint $|\rho| = 0.5$ is 0.00 pooled vs. 0.19 checkpoint-level at 0.5). Extending fidelity evaluation to month 24 across all 24 checkpoints yields pooled Spearman correlations within -0.154 to +0.090 and within-family correlations within -0.206 to +0.143 (-0.158 to +0.082 controlling model size and data scale).

**Table 12** | Monthly closed-loop replanning outcomes on realised simulator seeds (2,000 city-cluster bootstrap resamples). First panel validates the hysteresis-gated shrinking-horizon controller inside the exact simulator across 100 instances (replanned vs. open-loop commitment, plus independent fresh-seed control). Second panel pairs the primary 50M world model against both exact-simulator references across 100 benchmark instances (1,000 candidate proposals per month). Third panel reports mean months (out of 23) retaining the incumbent plan. Fourth panel gives the task-family breakdown of realised passes, excess false acceptance, and real-success shortfall. Bottom panel summarises aggregate real success and false acceptance on realised seeds alongside 32-seed majority open-loop verdicts.

| Family | Replanned passes | Open loop passes | Difference | Interval |
|---|---|---|---|---|
| grow | 11 of 20 | 8 | +0.150 | [-0.050, +0.400] |
| solvent | 8 of 20 | 4 | +0.200 | [+0.050, +0.450] |
| treasury | 20 of 20 | 9 | +0.550 | [+0.350, +0.800] |
| recover | 11 of 20 | 8 | +0.150 | [-0.050, +0.350] |
| power | 20 of 20 | 20 | 0.000 | [0.000, 0.000] |
| all | 70 of 100 | 49 | +0.210 | [+0.128, +0.292] |
| all, fresh seeds | 72 of 100 | 63 | +0.090 | [+0.020, +0.172] |

| | Primary exact-simulator reference | | | | Fresh-seed exact-simulator reference | | |
|---|---|---|---|---|---|---|---|
| Reading | Estimate | Interval | Width | Regime | Estimate | Interval | Regime |
| Excess, replanned | +0.130 | [+0.055, +0.205] | 0.150 | above zero | +0.160 | [+0.049, +0.245] | above zero |
| Change in excess | +0.170 | [+0.090, +0.257] | 0.167 | above zero | +0.080 | [+0.018, +0.154] | above zero |
| Shortfall, replanned | +0.300 | [+0.196, +0.423] | 0.227 | above zero | +0.320 | [+0.193, +0.436] | above zero |
| Change in shortfall | +0.200 | [+0.086, +0.337] | 0.251 | above zero | +0.080 | [-0.029, +0.222] | includes zero (seed-sensitive) |

| Family | World model | Exact simulator | Exact simulator, fresh seeds |
|---|---|---|---|
| grow | 21.8 | 21.6 | 21.9 |
| solvent | 21.2 | 21.3 | 21.8 |
| treasury | 21.2 | 21.9 | 22.4 |
| recover | 21.8 | 21.4 | 21.7 |
| power | 21.3 | 22.3 | 22.1 |

| | | World model | | Exact simulator | | Fresh seeds | | | |
|---|---|---|---|---|---|---|---|---|---|
| Family | $n$ | Replanned | Committed | Replanned | Committed | Replanned | Committed | Excess | Shortfall |
| grow | 20 | 8 | 7 | 11 | 8 | 15 | 12 | -0.150 [-0.400, +0.050] | +0.150 [-0.100, +0.400] |
| solvent | 20 | 6 | 3 | 8 | 4 | 9 | 6 | -0.050 [-0.299, 0.000] | +0.100 [-0.150, +0.382] |
| treasury | 20 | 10 | 6 | 20 | 9 | 19 | 15 | +0.200 [+0.050, +0.400] | +0.500 [+0.300, +0.700] |
| recover | 20 | 12 | 7 | 11 | 8 | 9 | 10 | 0.000 [-0.250, +0.250] | -0.050 [-0.250, +0.150] |
| power | 20 | 4 | 16 | 20 | 20 | 20 | 20 | +0.650 [+0.400, +0.850] | +0.800 [+0.600, +0.950] |

| | Real success | | | False acceptance | | |
|---|---|---|---|---|---|---|
| Model | Replanned | Open loop | Majority | Replanned | Open loop | Majority |
| World model (50M primary) | 0.400 | 0.390 | 0.340 | 0.230 | 0.160 | 0.150 |
| Exact simulator | 0.700 | 0.490 | 0.580 | 0.100 | 0.200 | 0.110 |

## B.8. Gallery Selection and Spatial Error Maps

Figures 2 and 11 display the highest-discrepancy exploited plans from random shooting at $B = 10^3$ on the 40-city core benchmark, where out of all 29 plans accepted by the primary model ($\geq$ one half of 8 rollouts) and rejected by the 32-seed oracle (1 *grow*, 2 *solvent*, 8 *treasury*, 12 *recover*, 6 *power*), we take the top four per task family ranked by model-minus-oracle pass share (breaking ties by most negative oracle margin, then instance ID), yielding 15 pairs. Each world-model map shows the first passing rollout among 8 fresh deterministic CPU draws (each plan produces at least one passing draw, with 2–8 of 8 passing) alongside the engine's primary oracle seed, cropped to the 40-by-40-tile window maximising structurally differing tiles (capturing a median of 80 per cent and at least 44 per cent of full-map divergences).

## B.9. Complete Benchmark Reference Tables

Tables 13–19 provide the complete numerical records underlying Sections 2–6, covering primary-model scaling across optimisers and budgets (Table 13), task-family breakdowns at $B = 10^3$ (Table 14), cross-ladder search curves across all 24 checkpoints (Table 15), short-horizon fidelity (Table 16), 24-month rollout drift and held-out ELBO (Table 17), threshold sensitivity on *recover* (Table 18), and per-checkpoint PPO results (Table 19).

**Table 13** | Primary 50M world model paired against the exact-simulator reference across all 193 core instances and the 148 well-posed subset. "False acceptance" denotes selections accepted by the model and rejected by the 32-seed oracle; "Exact ref." denotes the identical optimiser inside the exact-simulator reference, with the paired difference beside it; "Truly passing" gives ground-truth oracle pass share (model / exact reference); and "False rej." gives the share of truly passing selections rejected by the judge (model / exact reference). Best-of-$N$ runs on twenty cities (99 instances). Bottom rows evaluate initial trajectories from $10^6$-step PPO policies under the identical 8-sample verdict rule.

| | | All instances | | | | | | Well posed | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Search | $B$ | $n$ | False acceptance | Exact ref. | Difference | Truly passing | False rej. | $n$ | Difference | Truly passing |
| Random | $10^0$ | 193 | 0.083 [0.051, 0.123] | 0.021 | +0.062 [+0.026, +0.103] | 0.031/0.031 | 0.500/0.167 | 148 | +0.068 [+0.028, +0.116] | 0.027/0.027 |
| | $10^1$ | 193 | 0.124 [0.089, 0.167] | 0.031 | +0.093 [+0.048, +0.138] | 0.140/0.223 | 0.333/0.023 | 148 | +0.095 [+0.047, +0.146] | 0.176/0.284 |
| | $10^2$ | 193 | 0.130 [0.085, 0.173] | 0.057 | +0.073 [+0.005, +0.128] | 0.269/0.435 | 0.308/0.048 | 148 | +0.095 [+0.032, +0.149] | 0.318/0.527 |
| | $10^3$ | 193 | 0.150 [0.104, 0.195] | 0.083 | +0.067 [+0.010, +0.125] | 0.311/0.492 | 0.317/0.053 | 148 | +0.081 [+0.020, +0.147] | 0.378/0.574 |
| Adaptive | $10^1$ | 193 | 0.109 [0.072, 0.153] | 0.026 | +0.083 [+0.037, +0.130] | 0.067/0.130 | 0.462/0.040 | 148 | +0.088 [+0.049, +0.137] | 0.088/0.162 |
| | $10^2$ | 193 | 0.145 [0.104, 0.189] | 0.052 | +0.093 [+0.037, +0.144] | 0.192/0.347 | 0.324/0.090 | 148 | +0.101 [+0.048, +0.162] | 0.216/0.426 |
| | $10^3$ | 193 | 0.181 [0.130, 0.229] | 0.067 | +0.114 [+0.052, +0.173] | 0.275/0.466 | 0.208/0.100 | 148 | +0.108 [+0.048, +0.169] | 0.331/0.574 |
| Best-of-$N$ | $10^0$ | 99 | 0.071 [0.030, 0.117] | 0.000 | +0.071 [+0.030, +0.117] | 0.091/0.091 | 0.111/0.000 | 81 | +0.062 [+0.024, +0.120] | 0.111/0.111 |
| | $10^1$ | 99 | 0.141 [0.081, 0.212] | 0.040 | +0.101 [+0.031, +0.172] | 0.323/0.374 | 0.125/0.054 | 81 | +0.062 [0.000, +0.143] | 0.383/0.444 |
| | $10^2$ | 99 | 0.152 [0.081, 0.270] | 0.051 | +0.101 [+0.020, +0.227] | 0.404/0.404 | 0.225/0.025 | 81 | +0.099 [+0.012, +0.221] | 0.481/0.494 |
| | $10^3$ | 99 | 0.101 [0.051, 0.172] | 0.040 | +0.061 [0.000, +0.141] | 0.374/0.455 | 0.189/0.067 | 81 | +0.049 [-0.013, +0.133] | 0.444/0.543 |
| RL, main | $10^6$ m. | 579 | 0.178 [0.137, 0.225] | 0.056 | +0.122 [+0.076, +0.175] | 0.152/– | – | – | – | – |
| RL, ladder | $10^6$ m. | 13,896 | 0.233 [0.186, 0.284] | 0.064 | +0.170 [+0.124, +0.220] | 0.141/– | – | – | – | – |

**Table 14** | Task-family breakdown on the primary 50M checkpoint at $B = 10^3$ across all core instances and well-posed instances.

| | | All instances | | | | | | Well posed | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Task | Search | $n$ | False acceptance | Exact ref. | Difference | Accepted | Truly passing | $n$ | Difference | Truly passing |
| grow | random | 40 | 0.025 [0.000, 0.125] | 0.100 | -0.075 [-0.200, +0.025] | 0.050/0.375 | 0.225/0.275 | 36 | -0.056 [-0.167, +0.028] | 0.250/0.306 |
| | adaptive | 40 | 0.000 [0.000, 0.072] | 0.050 | -0.050 [-0.150, 0.000] | 0.050/0.300 | 0.150/0.275 | 36 | -0.056 [-0.167, 0.000] | 0.167/0.306 |
| solvent | random | 38 | 0.053 [0.000, 0.158] | 0.158 | -0.105 [-0.263, +0.026] | 0.053/0.421 | 0.026/0.289 | 20 | -0.050 [-0.228, +0.100] | 0.050/0.350 |
| | adaptive | 38 | 0.053 [0.000, 0.184] | 0.026 | +0.026 [-0.053, +0.132] | 0.079/0.184 | 0.053/0.211 | 20 | 0.000 [0.000, 0.000] | 0.100/0.350 |
| treasury | random | 40 | 0.200 [0.100, 0.350] | 0.025 | +0.175 [+0.050, +0.325] | 0.350/0.525 | 0.250/0.525 | 39 | +0.179 [+0.051, +0.333] | 0.256/0.538 |
| | adaptive | 40 | 0.275 [0.150, 0.425] | 0.075 | +0.200 [+0.025, +0.375] | 0.500/0.525 | 0.250/0.550 | 39 | +0.205 [+0.026, +0.385] | 0.256/0.564 |
| recover | random | 35 | 0.343 [0.200, 0.514] | 0.143 | +0.200 [0.000, +0.400] | 0.571/0.400 | 0.314/0.343 | 13 | +0.154 [0.000, +0.462] | 0.538/0.462 |
| | adaptive | 35 | 0.371 [0.229, 0.543] | 0.200 | +0.171 [-0.029, +0.343] | 0.571/0.400 | 0.257/0.257 | 13 | +0.077 [-0.154, +0.308] | 0.385/0.385 |
| power | random | 40 | 0.150 [0.050, 0.275] | 0.000 | +0.150 [+0.050, +0.275] | 0.800/1.000 | 0.725/1.000 | 40 | +0.150 [+0.050, +0.275] | 0.725/1.000 |
| | adaptive | 40 | 0.225 [0.125, 0.375] | 0.000 | +0.225 [+0.125, +0.375] | 0.800/1.000 | 0.650/1.000 | 40 | +0.225 [+0.125, +0.375] | 0.650/1.000 |

**Table 15** | Cross-model search audit across all 24 world-model checkpoints, reporting random shooting across all 120 cities ($B = 10^1$ to $10^3$) and adaptive CEM at $B = 10^3$ on the 40-city core benchmark across all 24 checkpoints, pooled across the five task families.

| | | | | Random shooting (120 cities) | | | | | | Adaptive CEM, $10^3$ (40-city core) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Objective | Size | Data | Stage | $10^1$ | $10^2$ | $10^3$ | Excess | Truly passing | Exact ref. | False acc. | Excess | Truly passing |
| one-step | 15M | full | half-trained | 0.103 | 0.132 | 0.120 | +0.051 | 0.375 | 0.070 | 0.192 | +0.124 | 0.233 |
| one-step | 15M | full | final | 0.120 | 0.131 | 0.153 | +0.084 | 0.362 | 0.070 | 0.192 | +0.124 | 0.275 |
| one-step | 15M | small | half-trained | 0.120 | 0.134 | 0.146 | +0.077 | 0.334 | 0.070 | 0.264 | +0.197 | 0.181 |
| one-step | 15M | small | final | 0.117 | 0.143 | 0.139 | +0.070 | 0.343 | 0.070 | 0.249 | +0.181 | 0.192 |
| one-step | 50M | full | half-trained | 0.125 | 0.122 | 0.108 | +0.038 | 0.366 | 0.070 | 0.202 | +0.135 | 0.181 |
| one-step | 50M | full | final | 0.138 | 0.138 | 0.162 | +0.092 | 0.336 | 0.070 | 0.181 | +0.114 | 0.275 |
| one-step | 50M | small | half-trained | 0.153 | 0.176 | 0.166 | +0.096 | 0.352 | 0.070 | 0.238 | +0.171 | 0.228 |
| one-step | 50M | small | final | 0.176 | 0.209 | 0.232 | +0.162 | 0.328 | 0.070 | 0.285 | +0.218 | 0.202 |
| one-step | 100M | full | half-trained | 0.129 | 0.145 | 0.145 | +0.075 | 0.355 | 0.070 | 0.223 | +0.155 | 0.238 |
| one-step | 100M | full | final | 0.096 | 0.092 | 0.110 | +0.040 | 0.345 | 0.070 | 0.140 | +0.073 | 0.269 |
| one-step | 100M | small | half-trained | 0.153 | 0.167 | 0.155 | +0.085 | 0.328 | 0.070 | 0.238 | +0.171 | 0.181 |
| one-step | 100M | small | final | 0.160 | 0.167 | 0.186 | +0.117 | 0.336 | 0.070 | 0.311 | +0.244 | 0.202 |
| pushforward | 15M | full | half-trained | 0.298 | 0.328 | 0.310 | +0.240 | 0.307 | 0.070 | 0.332 | +0.264 | 0.238 |
| pushforward | 15M | full | final | 0.232 | 0.265 | 0.272 | +0.202 | 0.291 | 0.070 | 0.275 | +0.207 | 0.254 |
| pushforward | 15M | small | half-trained | 0.166 | 0.174 | 0.153 | +0.084 | 0.343 | 0.070 | 0.269 | +0.202 | 0.181 |
| pushforward | 15M | small | final | 0.118 | 0.129 | 0.113 | +0.044 | 0.341 | 0.070 | 0.197 | +0.130 | 0.187 |
| pushforward | 50M | full | half-trained | 0.193 | 0.232 | 0.214 | +0.145 | 0.345 | 0.070 | 0.244 | +0.176 | 0.202 |
| pushforward | 50M | full | final | 0.183 | 0.188 | 0.188 | +0.118 | 0.321 | 0.070 | 0.249 | +0.181 | 0.181 |
| pushforward | 50M | small | half-trained | 0.179 | 0.204 | 0.186 | +0.117 | 0.315 | 0.070 | 0.259 | +0.192 | 0.181 |
| pushforward | 50M | small | final | 0.152 | 0.176 | 0.172 | +0.103 | 0.319 | 0.070 | 0.228 | +0.161 | 0.228 |
| pushforward | 100M | full | half-trained | 0.181 | 0.207 | 0.207 | +0.138 | 0.373 | 0.070 | 0.218 | +0.150 | 0.295 |
| pushforward | 100M | full | final | 0.155 | 0.192 | 0.190 | +0.120 | 0.336 | 0.070 | 0.192 | +0.124 | 0.275 |
| pushforward | 100M | small | half-trained | 0.120 | 0.152 | 0.134 | +0.064 | 0.310 | 0.070 | 0.218 | +0.150 | 0.192 |
| pushforward | 100M | small | final | 0.111 | 0.139 | 0.122 | +0.052 | 0.329 | 0.070 | 0.187 | +0.119 | 0.218 |

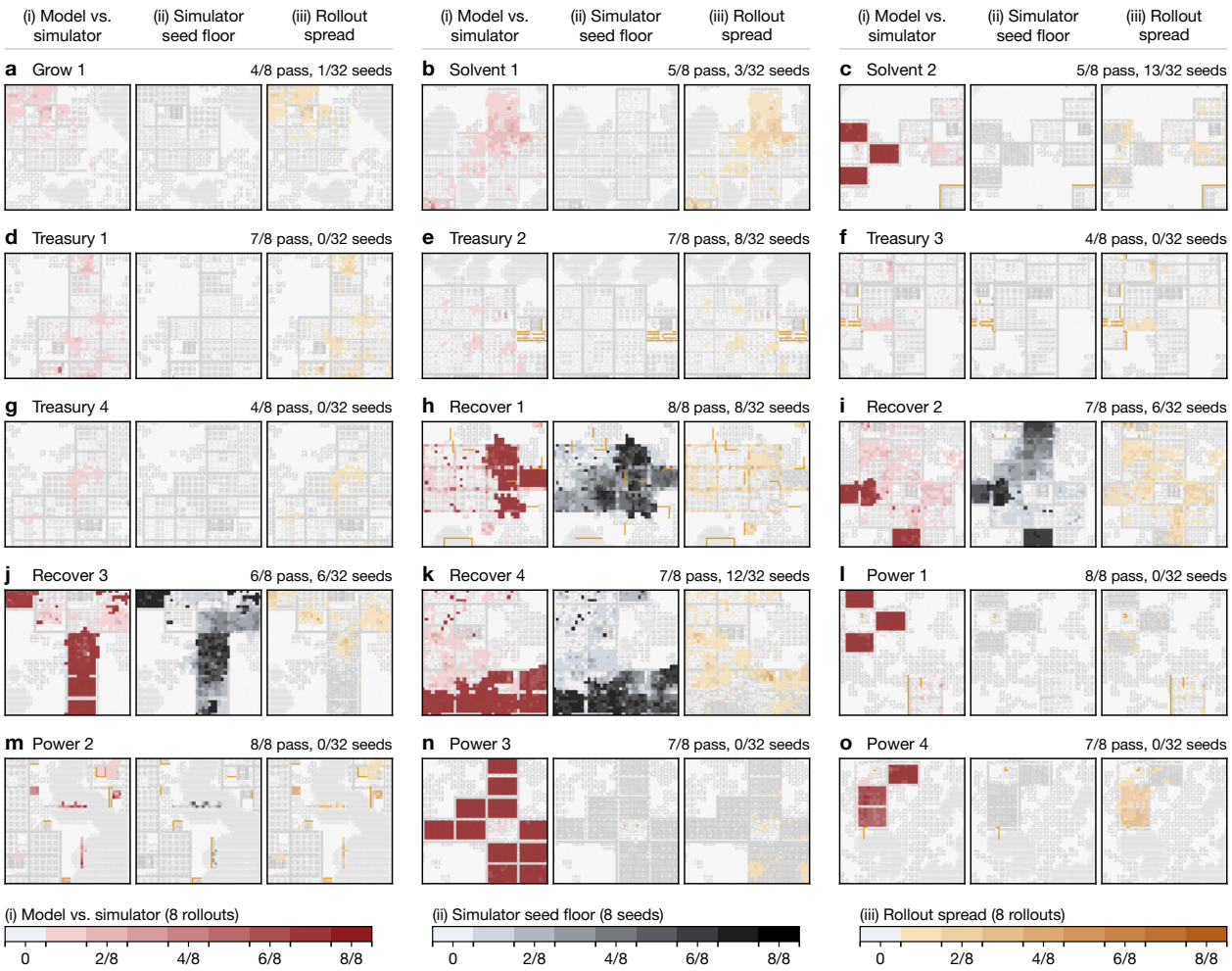


**Figure 11 | Spatial decomposition of world-model error, simulator stochasticity, and within-model rollout spread.** For each of the 15 exploited city plans in Figure 2 (matching panel letters and 40-by-40 spatial windows, overlaid on desaturated month-24 procedural city sprites with gold outlines marking the plan's direct spatial build/demolish footprint), three per-tile heat maps show (i) *Model vs. simulator* (share of 8 world-model rollouts diverging in tile class or power status from primary engine replay, crimson), (ii) *Simulator seed floor* (share of 8 independent engine seeds diverging from the primary seed in the exact simulator, charcoal), and (iii) *Rollout spread* (internal disagreement across the 8 world-model rollouts, amber).

### B.10. Additional Modeling, Search, and Controller Ablations

*Pushforward rollout fine-tuning versus one-step variational training.* Fine-tuning final one-step models on stop-gradient posterior-mode rollouts slightly lowers one-month structural changed-tile F1 in 8 of 8 parent-child pairs and 12-month changed-since-start F1 in 8 of 8 pairs (primary lineage shifts from 0.372 to 0.319), because in a stochastic cellular automaton whose own cross-seed changed-tile F1 is 0.206, pairing a rolled-out model state against a single archived engine state injects transition variance; both model families remain in our 24-checkpoint ladder.

*Unseeded column-independent CEM versus seeded elite-mutation CEM.* Standard unseeded CEM (de Boer et al., 2005), which resamples all 600 action columns independently from fitted marginals at each generation, disrupts multi-month spatial coordination, so even inside the exact-simulator reference at $B = 10^3$, unseeded CEM achieves only 0.062 ground-truth success versus 0.492 for random shooting (with the primary world model accepting 0.057 of its own unseeded selections), motivating our logged-plan-seeded elite-mutation CEM in Section 3.

*Shuffled-residual noise surrogates.* Replacing the world model with a synthetic surrogate that adds shuffled empirical model residuals to ground-truth margins (passing exact-simulator calibration across all 12 of 12 budget cells in volatility-matched form) yields 0.539 false acceptance under $10^3$ random shooting (0.715 unweighted), exceeding the actual neural model's 0.150 because smooth state-dependent model biases preserve local plan rankings better than i.i.d. residual draws.

**Table 16** | Short-horizon trajectory fidelity across all 16 completed training runs on the held-out test split (2,000 transitions, 700 trajectories), reporting one-month structural (and raw) changed-tile F1, 12-month changed-since-start structural F1, and median relative population and treasury errors. Non-zero seeds denote the additional 50M ensemble runs. Errors at the 24-month task horizon and held-out log-likelihoods appear in Table 17.

| Objective | Size | Data | Seed | F1, 1 m | Raw F1, 1 m | F1, 12 m | Pop., 12 m | Funds, 12 m |
|---|---|---|---|---|---|---|---|---|
| one-step | 15M | small | 0 | 0.170 | 0.287 | 0.369 | 0.057 | 0.046 |
| pushforward | 15M | small | 0 | 0.147 | 0.256 | 0.295 | 0.052 | 0.063 |
| one-step | 15M | full | 0 | 0.169 | 0.282 | 0.373 | 0.052 | 0.048 |
| pushforward | 15M | full | 0 | 0.153 | 0.263 | 0.308 | 0.053 | 0.060 |
| one-step | 50M | small | 0 | 0.166 | 0.287 | 0.372 | 0.045 | 0.038 |
| pushforward | 50M | small | 0 | 0.148 | 0.259 | 0.304 | 0.056 | 0.038 |
| one-step | 50M | full | 0 | 0.179 | 0.294 | 0.372 | 0.060 | 0.035 |
| pushforward | 50M | full | 0 | 0.168 | 0.282 | 0.319 | 0.050 | 0.034 |
| one-step | 100M | small | 0 | 0.172 | 0.297 | 0.376 | 0.052 | 0.034 |
| pushforward | 100M | small | 0 | 0.155 | 0.273 | 0.290 | 0.061 | 0.039 |
| one-step | 100M | full | 0 | 0.178 | 0.294 | 0.379 | 0.048 | 0.030 |
| pushforward | 100M | full | 0 | 0.168 | 0.280 | 0.338 | 0.044 | 0.027 |
| one-step | 50M | full | 1 | 0.182 | 0.294 | 0.382 | 0.048 | 0.034 |
| pushforward | 50M | full | 1 | 0.168 | 0.282 | 0.329 | 0.044 | 0.038 |
| one-step | 50M | full | 2 | 0.178 | 0.295 | 0.376 | 0.050 | 0.036 |
| pushforward | 50M | full | 2 | 0.172 | 0.286 | 0.323 | 0.046 | 0.056 |
| Engine under another seed | | | | 0.206 | 0.319 | 0.423 | 0.027 | 0.003 |
| Copy-last-state | | | | 0.000 | 0.000 | 0.000 | 0.112 | 0.400 |

**Table 17** | Long-horizon rollout drift at the 24-month task horizon and held-out variational log-likelihood (ELBO in nats/transition on the development split at mid-training and full budget) across all final checkpoints and the coverage-retrained model.

| Objective | Size | Data | Seed | Overlays ≥ copy | Pollution | Fire cov. | Powered zones | ELBO, half | ELBO, final | Final below half |
|---|---|---|---|---|---|---|---|---|---|---|
| one-step | 15M | small | 0 | 4 | 1.455 | 0.989 | 2.4 | -2948 | -3061 | yes |
| pushforward | 15M | small | 0 | 7 | 1.673 | 1.713 | 9.7 | -3000 | -3063 | yes |
| one-step | 15M | full | 0 | 5 | 1.530 | 1.017 | 2.8 | -2928 | -2779 | no |
| pushforward | 15M | full | 0 | 7 | 1.646 | 1.780 | 7.7 | -2843 | -2822 | no |
| one-step | 50M | small | 0 | 4 | 1.351 | 0.955 | 3.4 | -2952 | -3414 | yes |
| pushforward | 50M | small | 0 | 7 | 1.704 | 1.277 | 9.8 | -3159 | -3316 | yes |
| one-step | 50M | full | 0 | 6 | 1.520 | 1.127 | 6.1 | -2771 | -2664 | no |
| pushforward | 50M | full | 0 | 7 | 1.807 | 1.594 | 8.7 | -2719 | -2701 | no |
| one-step | 100M | small | 0 | 4 | 1.270 | 0.814 | 2.7 | -2905 | -3665 | yes |
| pushforward | 100M | small | 0 | 7 | 1.698 | 1.232 | 11.0 | -3280 | -3514 | yes |
| one-step | 100M | full | 0 | 6 | 1.546 | 1.180 | 7.9 | -2705 | -2574 | no |
| pushforward | 100M | full | 0 | 5 | 1.731 | 1.067 | 4.4 | -2626 | -2604 | no |
| one-step | 50M | full | 1 | 2 | 1.377 | 0.814 | 2.2 | -2780 | -2662 | no |
| pushforward | 50M | full | 1 | 7 | 1.745 | 3.591 | 8.9 | -2718 | -2700 | no |
| one-step | 50M | full | 2 | 6 | 1.577 | 1.088 | 6.6 | -2772 | -2666 | no |
| pushforward | 50M | full | 2 | 7 | 1.893 | 9.291 | 7.1 | -2725 | -2704 | no |
| one-step | 50M | full + cov. | 0 | 7 | 1.552 | 1.197 | 4.0 | -2786 | -2664 | no |
| Engine under another seed | | | | | 0.703 | 0.125 | 0.0 | | | |
| Copy-last-state | | | | | 1.000 | 1.000 | 0.5 | | | |

*Internal accounting drift as an unsupervised anomaly signal.* Checking whether imagined trajectories violate exact municipal budget conservation by more than 5 per cent of initial funds shows breach rates rising across random-shooting budgets (0.00, 0.03, 0.09, and 0.11 from $10^0$ to $10^3$ among accepted plans, vs. 0.13 on rejected plans and 0.09 on unoptimised logged plans), yet separates false acceptances (0.159) from true acceptances (0.082, difference +0.077 [-0.010, +0.171], rejected 0.126) too weakly to serve as a standalone filter.

*Behaviour-cloned warm-starting and hysteresis-free replanning.* Warm-starting PPO spatial heads from scripted builders improved engine-side returns on *power* (+0.297 [+0.175, +0.428]) and *solvent* (+0.109 [+0.036, +0.188]) with neutral effects elsewhere, so we report uniform from-scratch PPO throughout. Similarly, inside the exact simulator, naive monthly replanners without switching hysteresis (either redrawing fresh proposals each

**Table 18** | Target-threshold sensitivity on *recover* for random shooting at $B = 10^3$ on the primary checkpoint, re-evaluated exactly at thresholds set one quarter, one half, and three quarters of the way between the best constant no-build plan ($c$) and the engine-search ceiling ($e$).

| Threshold $\lambda$ (from $c$ to $e$) | $n$ | False acceptance | Exact ref. | Excess | Truly passing, model/ref. | World model accepts |
|---|---|---|---|---|---|---|
| 0.25 | 35 | 0.257 [0.143, 0.429] | 0.200 | +0.057 [-0.114, +0.229] | 0.514 / 0.486 | 0.657 |
| 0.50 (benchmark) | 35 | 0.343 [0.200, 0.514] | 0.143 | +0.200 [0.000, +0.400] | 0.314 / 0.343 | 0.571 |
| 0.75 | 35 | 0.400 [0.257, 0.571] | 0.171 | +0.229 [0.000, +0.429] | 0.143 / 0.143 | 0.486 |

**Table 19** | In-imagination PPO performance across all 24 world-model checkpoints after $10^6$ imagined steps (means over five task families and three seeds per cell, with engine-trained PPO reference optimism gap +0.077 [+0.059, +0.095]).

| Objective | Size | Data | Stage | Runs | Dream accepts | Oracle accepts | Gap | Closed loop | Margin inflation |
|---|---|---|---|---|---|---|---|---|---|
| one-step | 15M | full | half-trained | 15 | 0.282 | 0.181 | +0.101 | 0.255 | +0.007 |
| one-step | 15M | full | final | 15 | 0.285 | 0.167 | +0.118 | 0.233 | -0.017 |
| one-step | 15M | small | half-trained | 15 | 0.291 | 0.148 | +0.143 | 0.219 | +0.012 |
| one-step | 15M | small | final | 15 | 0.253 | 0.129 | +0.123 | 0.201 | -0.084 |
| one-step | 50M | full | half-trained | 15 | 0.327 | 0.159 | +0.168 | 0.223 | +0.097 |
| one-step | 50M | full | final | 15 | 0.271 | 0.138 | +0.133 | 0.202 | -0.016 |
| one-step | 50M | small | half-trained | 15 | 0.366 | 0.151 | +0.215 | 0.223 | +0.221 |
| one-step | 50M | small | final | 15 | 0.346 | 0.147 | +0.199 | 0.216 | +0.170 |
| one-step | 100M | full | half-trained | 15 | 0.330 | 0.167 | +0.162 | 0.242 | +0.111 |
| one-step | 100M | full | final | 15 | 0.211 | 0.141 | +0.070 | 0.192 | -0.177 |
| one-step | 100M | small | half-trained | 15 | 0.322 | 0.161 | +0.161 | 0.216 | +0.141 |
| one-step | 100M | small | final | 15 | 0.294 | 0.144 | +0.150 | 0.209 | +0.029 |
| pushforward | 15M | full | half-trained | 15 | 0.486 | 0.135 | +0.351 | 0.202 | +0.287 |
| pushforward | 15M | full | final | 15 | 0.433 | 0.128 | +0.305 | 0.187 | +0.190 |
| pushforward | 15M | small | half-trained | 15 | 0.342 | 0.128 | +0.214 | 0.181 | -0.010 |
| pushforward | 15M | small | final | 15 | 0.265 | 0.126 | +0.140 | 0.163 | -0.203 |
| pushforward | 50M | full | half-trained | 15 | 0.411 | 0.139 | +0.272 | 0.203 | +0.215 |
| pushforward | 50M | full | final | 15 | 0.380 | 0.119 | +0.261 | 0.162 | +0.251 |
| pushforward | 50M | small | half-trained | 15 | 0.316 | 0.128 | +0.188 | 0.196 | -0.023 |
| pushforward | 50M | small | final | 15 | 0.261 | 0.114 | +0.147 | 0.172 | -0.175 |
| pushforward | 100M | full | half-trained | 15 | 0.384 | 0.155 | +0.229 | 0.216 | +0.163 |
| pushforward | 100M | full | final | 15 | 0.337 | 0.142 | +0.195 | 0.205 | +0.104 |
| pushforward | 100M | small | half-trained | 15 | 0.246 | 0.121 | +0.125 | 0.168 | -0.273 |
| pushforward | 100M | small | final | 15 | 0.223 | 0.115 | +0.108 | 0.158 | -0.338 |

month without retaining the incumbent, passing 8 vs. 8 of 25 overall and 1 vs. 5 of 5 on *power*, or pooling the incumbent without a standard-error switching margin, passing 49 vs. 49 of 100 overall, 1 vs. 8 of 20 on *grow*, and 4 vs. 8 of 20 on *recover*) thrash between incompatible spatial build plans, motivating the two-standard-error hysteresis gate used in Section 3.5.

## C. Domain-General Audit Library and Continuous-Control Check

All search audits and statistical estimators are implemented in `hpr-dream`, a domain-agnostic Python library depending solely on `numpy` and `scipy`. It represents multi-step plans as integer arrays over a typed schema of categorical and bounded-integer columns, interfaces with any candidate world model via a batch scoring callback and with ground truth via an oracle margin callback, and logs standardised per-selection audit records (instance, city cluster, optimiser, budget, seed, plan hash, model pass share, and ground-truth verdict). Environment-specific adaptors handle rollout execution, with our Micropolis adaptor wrapping both neural checkpoints and the fresh-seed exact-simulator reference and our continuous-control adaptor snapshotting and restoring MuJoCo simulator states for TD-MPC2. From these callbacks, `hpr-dream` executes random shooting, best-of-$N$, standard CEM, and seeded adaptive CEM under exact call-budget accounting, and computes city-clustered percentile and BCa bootstrap intervals, exact zero-event upper bounds, false acceptance, conditional and false rejection, paired reference excesses, Brier and log-score decompositions, ROC AUC, clustered Spearman and Kendall rank correlations, shuffled-residual and binary-flip noise surrogates, and structured JSON audit cards.

To verify that `hpr-dream` operates cleanly out-of-the-box on external latent-space world models we did not train, we audited two publicly released single-task TD-MPC2 checkpoints (Hansen, 2024; Hansen et al., 2024) on DeepMind Control Suite tasks (Tunyasuvunakool et al., 2020). Here the world model comprises TD-MPC2's learned encoder, latent transition dynamics, and reward head, while the oracle replays continuous action sequences from exactly restored MuJoCo physics states (numerical state-restoration tolerance 0). Across 20 initial states per task, thresholds over 50-step open-loop trajectories are calibrated so 0.2–0.5 of correlated random behaviour plans pass; because MuJoCo replay is deterministic, the exact-simulator false-acceptance floor is identically zero. Because TD-MPC2 trains on short 3-step rollouts bootstrapped with terminal value estimates rather than open-loop 50-step sums, unrolling its latent reward head open loop over 50 steps induces a strong upward return bias of 1.9 to 2.9 logged-plan standard deviations, falsely accepting 0.844 (*cheetah-run*) and 0.894 (*walker-walk*) of failing behaviour plans. Consequently, open-loop shooting actually lowers unconditional false acceptance from 0.567 and 0.683 at $B = 10^0$ to 0.350 [0.233, 0.483] and 0.317 [0.167, 0.500] at $B = 10^3$ as genuinely higher-return trajectories are selected (running TD-MPC2's native MPPI planner open loop with 153,600 rollouts yields 0.017 and 0.117 false acceptance; *hopper-hop* was excluded because 18 of 20 starts yielded identically zero behaviour return). We treat this purely as an external software validation rather than a claim about model-predictive control when TD-MPC2 is run closed loop at its native 3-step horizon.

**Code, data, and licensing.** `hpr-dream` is released under the Apache-2.0 licence. The Micropolis C++ engine modifications, Python replay bindings, data generator, task harness, and evaluation scripts link against MicropolisCore and are released under GPL-3.0 with Electronic Arts' additional terms, with all modifications explicitly marked. Complete JSON/CSV result records, calibrated task definitions, figure scripts, and paper sources reproducing all numbers and graphics in this work are released under CC BY 4.0 via hyperprior.net. As noted in Appendix A.3, neural network checkpoints (documented by SHA-256 digest) and trained RL policies are not distributed.

## D. Extended Related Work

**Offline model-based optimisation and surrogate exploitation.** Optimising static designs against learned reward surrogates produces severe out-of-distribution overestimation in offline model-based optimisation (MBO), motivating conservative conditioning, autofocused retraining, and adversarial regularisation across Design-Bench (Brookes et al., 2019; Fannjiang and Listgarten, 2020; Kumar and Levine, 2020; Trabucco et al., 2021, 2022). Our setting extends static surrogate optimisation to autoregressive 24-step stochastic dynamics where spatial actions compound nonlinearly and exact simulator replay benchmarks excess error over regressional Goodhart selection noise (Bailey et al., 2017; Bastani et al., 2026; Eshwar Reddy M and Karmakar, 2026; Ma and Zhao, 2026; Manheim and Garrabrant, 2019) induced by structural dynamics errors (Jakobi et al., 1995; Koos et al., 2013; Pan et al., 2022; Skalse et al., 2022). In reward-model scaling for language and control agents, proxy-gold gaps grow predictably with KL divergence and optimisation budget (Frick et al., 2024; Gao et al., 2023; Kim et al., 2025; Wen et al., 2025).

**Uncertainty regularisation, rollout fidelity, and generative game engines.** Ensemble disagreement penalties (Chua et al., 2018; Kurutach et al., 2018; Yu et al., 2020), absorbing OOD halts (Kidambi et al., 2021), value conservatism (Rigter et al., 2022; Yu et al., 2021), and multi-step hallucinated or pushforward training (Brandstetter et al., 2022; Talvitie, 2014, 2017; Venkatraman et al., 2015) aim to bound rollout exploitation, though empirical correlations between ensemble variance and true dynamics error are often weak (Lu et al., 2022). Recent preprints and benchmarks document reward inflation and dream-reality gaps in robotic and game world models (AMD AI Group, 2026; Bhamidipaty et al., 2026a; Güzel et al., 2026; Hawke, 2026; Jiang et al., 2026; Khassanov, 2026; Kieto, 2026; Kong et al., 2026; Reka, 2026; Xu et al., 2026; Yang et al., 2026; Zhou et al., 2026) alongside off-policy world-model evaluation on fixed policy sets (Fu et al., 2021; Zhang et al., 2021) and interventional rollout debugging (Rahmani, 2026; Vakalis, 2026; Yao et al., 2026; Zhu et al., 2026). Finally, our testbed builds on Micropolis RL environments (Earle, 2020) and complements neural game and strategy simulators spanning Atari, Doom, platformers, StarCraft II, and turn-based strategy (Alonso et al., 2024; Bamford and Lucas, 2020; Bruce et al., 2024; Dyer et al., 2023; Egorov and Shpilman, 2022; Kaiser et al., 2020; Li et al., 2026; Micheli et al., 2023; Synnaeve et al., 2018; Valevski et al., 2025; Zhang et al., 2025, 2026).