TAPE: A CELLULAR AUTOMATA BENCHMARK FOR EVALUATING RULE-SHIFT GENERALIZATION IN REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution generalization in reinforcement learning is difficult to attribute when benchmark shifts jointly alter dynamics, observations, goals, and rewards. TAPE isolates one latent non-stationarity source—rule-shift in transition dynamics—while fixing the observation-action interface. The protocol combines deterministic splits, 20-seed replication, bootstrap uncertainty quantification, and continuous endpoints for sparse-success regimes. Across baseline families, experiments show a stable ID-to-OOD degradation pattern with strong heterogeneity across stable, periodic, and chaotic rules. The same ID–OOD gap persists in an intentionally simple 1D deterministic regime, indicating that current RL pipelines remain brittle under latent-law variation even after removing common perceptual confounds. We report a protocol-matched budgeted true-dynamics random-shooting reference ($p_{\text{oracle}} \approx 0.187$), define oracle-normalized scores $\text{ON}(p) = 100\,p/p_{\text{oracle}}$, and include a feasibility regime ($L{=}H{=}16$) where rule-wise solvability reaches 100%. Under this design, low absolute success in learned agents is interpreted as a stress signal for latent-mechanism adaptation, positioning TAPE as a mechanism-oriented benchmark for robustness under explicit distribution shift.

1 INTRODUCTION

A persistent challenge in reinforcement learning is the gap between strong in-distribution (ID) optimization and reliable out-of-distribution (OOD) control when the transition law shifts (Dulac-Arnold et al., 2021; Kirk et al., 2021; Packer et al., 2018; Cobbe et al., 2019; Farebrother et al., 2018; Zhang et al., 2018). In many benchmark suites, OOD degradation entangles shifts in perception, goals, dynamics, and rewards, making it difficult to attribute a failure to one mechanism (Cobbe et al., 2020; Kirk et al., 2021; Justesen et al., 2018; Agarwal et al., 2021; Henderson et al., 2018). Benchmark construction and evaluation diversity can also affect apparent algorithm rankings (Machado et al., 2018). This problem also connects to the view of RL as sequential decision making under a changing information state (Wang et al., 2017; Sutton & Barto, 2018). We study a minimal controlled system where an agent observes a 1D binary tape, applies local interventions, and interacts with deterministic dynamics governed by an unobserved rule z that maps local neighborhoods to next-state updates. During training, z varies across episodes; during evaluation, rules are drawn from a disjoint holdout set. The core objective is to test whether learned policies recover transferable structure or specialize to seen rule realizations.

Our protocol isolates *rule-shift* in latent transition dynamics. OOD evaluation uses held-out rules $z \in \mathcal{Z}_{\text{test}}$ with $\mathcal{Z}_{\text{test}} \cap \mathcal{Z}_{\text{train}} = \emptyset$, while the observation/action interface and goals stay fixed. By fixing the observation, action, and goal interfaces, the protocol isolates a hidden transition-law change while remaining comparable to contextual-MDP, system-identification, and meta-RL formulations (Hallak et al., 2015; Killian et al., 2017; Benjamins et al., 2022; Yu et al., 2017; Peng et al., 2018; Duan et al., 2017; Nagabandi et al., 2019; Yu et al., 2020; Zintgraf et al., 2020; Ni et al., 2022). We instantiate it with TAPE, derived from one-dimensional elementary cellular automata (CA) (Wolfram, 1983; Wolfram & Gad-el Hak, 2003), where each task is indexed by a latent CA update rule under identical interface semantics. Figure 1 visualizes a rollout under fixed latent rule ($z{=}30$, $L{=}8$), with each column showing one step (agent intervention then CA transition).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

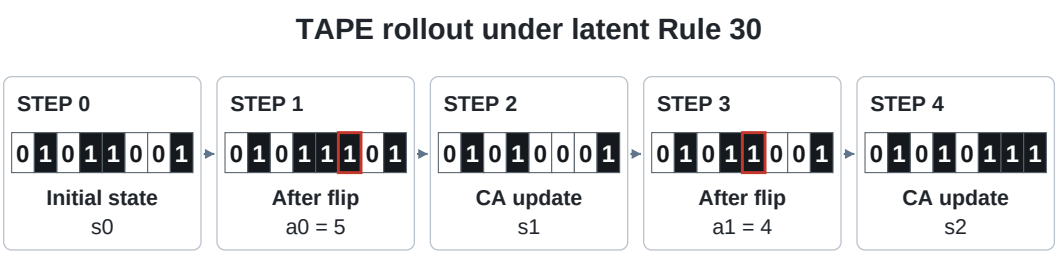

Figure 1: TAPE rollout (L=8, rule z=30): each step applies a single-cell flip, then a synchronous CA update. Only the latent rule varies across episodes.

TAPE contributes two linked elements: a controlled benchmark for generalization under latent rule shift and a reproducible protocol for diagnosing the resulting transfer difficulty. The benchmark is intended to expose where transfer fails, rather than to produce a state-of-the-art leaderboard.

Although the tape configuration is observed at each step, the transition rule z remains latent. Robust control therefore couples action selection with inference about the elementary CA law governing the dynamics, equivalent to a contextual MDP with a hidden discrete context. Representative RL families often optimize reactive behavior under fixed or weakly varying dynamics, yet existing evidence does not establish consistent recovery of transferable latent-law inference under strict rule holdout (Rakelly et al., 2019; Chua et al., 2018; Hafner et al., 2023). TAPE operationalizes this distinction: replacing training rules with disjoint held-out rules can preserve training competence while degrading control quality. We therefore use oracle-gap calibration (Section 6.1) to interpret residual headroom as evidence about latent-law inference fidelity, not a generic score deficit on one fixed MDP. Our contributions are organized around five elements:

- **Benchmark + protocol.** We construct TAPE, a CA-based RL benchmark that isolates latent rule-shift generalization with explicit holdout-rule and holdout-length regimes, and we release a reproducible pipeline that formalizes split generation, train-time rule sampling, and seed-level uncertainty reporting.
- **Calibrated evaluation stack.** We integrate strict-success reporting with oracle calibration (budgeted true-dynamics planner reference $p_{\text{oracle}} \approx 18.7\%$ plus smaller-scale feasibility checks), oracle-normalized scores in the main aggregate table, and continuous endpoints (final distance, AUC, soft success@ε) to stabilize interpretation in sparse-success regimes.
- **Comprehensive empirical diagnosis.** We benchmark model-free, augmentation-based, task-inference, and world-model families (20 seeds), then probe robustness across five data splits, horizon shift, and rule categories (stable/periodic/chaotic) to localize transfer degradation modes.
- **Non-learned Bayes diagnostic.** We evaluate a finite-rule belief-tracking baseline on the same split with 20 seeds (Section 6.2; full description in Appendix C).
- **Mechanism-oriented analysis.** We conduct a credibility analysis for DreamerV3-style world models (prediction-error growth and sensitivity trends) and provide formal IG identities with scope conditions and failure modes under rule shift (Appendix K).

2 The TAPE Benchmark

2.1 Environment: "tape physics" with local interventions

Each task is indexed by a latent rule $z \in \mathcal{Z}$ (e.g., $|\mathcal{Z}| = 256$ elementary CA rules). An episode is a length-H interaction with state $s_t \in \{0,1\}^L$ (binary tape), action $a_t \in \{1,\dots,L\}$ (single-cell flip at index a_t), dynamics $s_{t+1} = F_z(G(s_t, a_t))$ (intervention followed by CA update), and observation $o_t = [s_t,\ t/H] \in \mathbb{R}^{L+1}$.

Elementary CA rule F_z. At time t, the agent first applies the intervention $\tilde{s}_t = G(s_t, a_t)$; an elementary CA then synchronously updates each cell from a 3-bit neighborhood read on $\tilde{s}_t$ (not from the pre-flip tape s_t), parameterized by the 8-bit rule code z. The compact bitwise realization of this update, including neighborhood indexing and truth-table lookup, follows the standard elementary-CA construction (Wolfram & Gad-el Hak, 2003) and is specified for reproducibility in Appendix F. Figure 2 summarizes the benchmark protocol.

Reward and termination. We consider a goal tape $g \in \{0,1\}^L$ (e.g., all zeros). Let $\mathrm{dist}(s, g) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}[s_i \neq g_i]$. A strict success event is $\mathrm{dist}(s, g) = 0$. Rewards follow the released implementation: a *shaped negative distance* (a monotone transform of $-\mathrm{dist}$, used only as the training signal) plus an optional success bonus; all reported metrics use the normalized Hamming distance above unless a caption explicitly states raw units. An episode terminates upon success or when $t = H$. We discuss "floor effects" caused by strict success in Section 6.

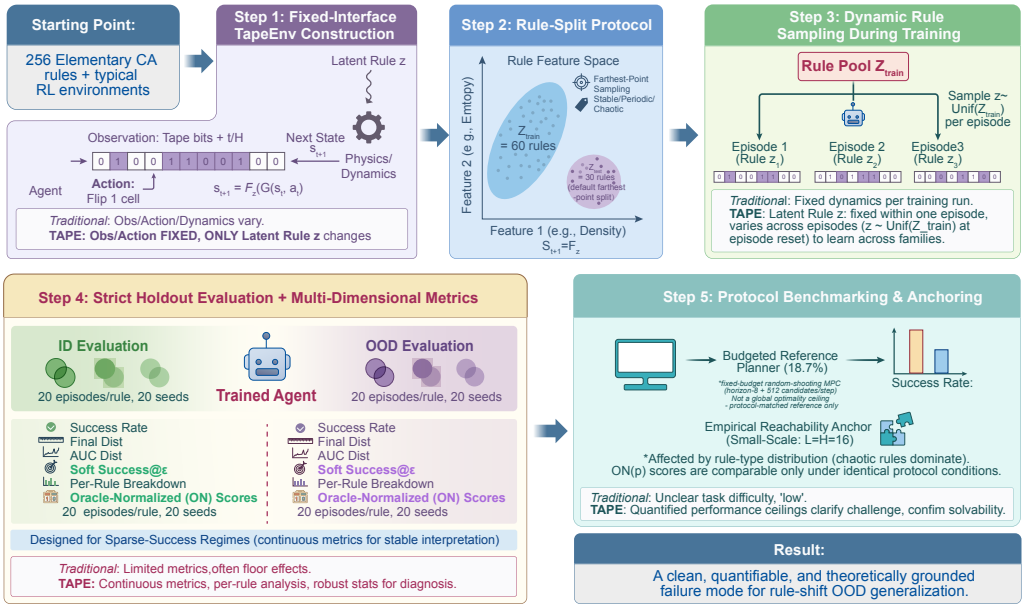

Figure 2: Benchmark pipeline for TAPE. **(1)** A fixed observation/action interface exposes the tape and normalized time, with single-cell flips as actions and a latent elementary CA rule. **(2)** Farthest-point sampling in rule-feature space creates disjoint train/test pools. **(3)** Each training reset samples $z \sim \mathrm{Unif}(\mathcal{Z}_{\mathrm{train}})$. **(4)** ID and holdout-rule OOD evaluation reports strict/soft success and final/AUC distance across 20 seeds. **(5)** Baseline comparisons and calibrated reporting support reproducible OOD assessment.

2.2 Rule-Split Protocols: Controlling What Changes

We partition rule identities into disjoint sets $\mathcal{Z}_{\mathrm{train}}$ and $\mathcal{Z}_{\mathrm{test}}$. Training resets sample only from $\mathcal{Z}_{\mathrm{train}}$, whereas OOD evaluation samples only from $\mathcal{Z}_{\mathrm{test}}$. The observation and action spaces remain identical, so the latent rule is the sole distributional change. We optionally increase horizon at test time (e.g., $H_{\mathrm{test}} > H_{\mathrm{train}}$) while keeping the same rule split. To construct diverse splits, we embed each rule in a feature vector of rollout statistics and apply farthest-point sampling to cover the resulting rule space. The procedure is deterministic; its implementation details and split artifacts are given in Appendix F. We assign each rule an operational type (stable / periodic / chaotic) using fixed thresholds on simulated rollout statistics; threshold values and implementation details are specified in the supplementary materials (Appendix F).

3 Evaluation Standards for High-Variance OOD RL

OOD comparisons can be underpowered when they use few seeds or small test sets, producing unstable rankings. Because latent-rule variation creates substantial task heterogeneity in TAPE, replication is part of the evaluation protocol rather than an optional robustness check. We report multi-seed estimates with uncertainty over training stochasticity; checkpointing, evaluation frequency, and bootstrap aggregation are specified in Appendix F, with formal definitions in Appendix L. For a fixed pipeline and split, the relevant stochasticity comes from initialization, exploration, and minibatch sampling, which seed-level bootstrap intervals quantify directly. The default training and evaluation protocol does not optimize information gain about z: agents optimize return-based objectives, with auxiliary contrastive or augmentation losses where applicable, and evaluation uses success and distance endpoints. Information-gain identities and their caveats under rule shift are provided in Appendix K.

4 Methods Under Evaluation

We evaluate representative RL families under matched training budgets and one shared split protocol. The suite includes model-free baselines (DQN (Mnih et al., 2015), PPO (Schulman et al., 2017)), augmentation-regularized variants (RAD-DQN with bit-flip/bit-shift transforms (Laskin et al., 2020b), and CURL-DQN with a contrastive auxiliary objective (Laskin et al., 2020a)), a task-inference baseline (PEARL-style DQN conditioned on context-inferred latent embeddings (Rakelly et al., 2019)), and a world-model baseline (DreamerV3-style latent dynamics with reconstruction/reward modeling and imagination-based policy learning (Hafner et al., 2023)). This coverage compares policy regularization, latent-task inference, and explicit dynamics modeling under the same rule-shift protocol. We also evaluate a **non-learned finite-rule Bayes filter** on the identical split and seed budget. Its formalization, objective, and complete result table appear in Appendix C; it is reported separately from Table 1, which compares trained policy and world-model families under the standard evaluation contract. Throughput-oriented implementation choices (vectorized rollouts; budgets counted in environment steps) are summarized in Appendix F.

5 Experimental Setup

Unless otherwise noted, we use tape length $L = 32$ and training horizon $H_{\text{train}} = 32$. The default goal is the all-zero tape. We use this goal to keep the primary protocol comparable; goal-conditioned difficulty need not be rule-invariant. For holdout-length evaluation we use $H_{\text{test}} = 64$. Training episodes sample a fresh latent rule $z \sim \text{Unif}(\mathcal{Z}_{\text{train}})$ at reset so the agent learns across a *family* of transition laws rather than memorizing one. Training interaction budgets, evaluation checkpoints, episodes-per-rule evaluation, metric definitions, and the rule for aggregating the last checkpoints into reported numbers are listed in Appendix F. Under the default farthest-point split used in our main runs, the held-out set is type-imbalanced (22/30 chaotic rules). The imbalance emerges from diversity-maximizing sampling in rule-feature space. We report it explicitly because it materially affects the global strict-success aggregate. Unless stated otherwise, headline aggregates are micro-averages over this realized composition. We therefore interpret them together with by-type results rather than as type-balanced macro estimates.

6 Results: ID vs Holdout-Rule OOD

Reader guide. Learned strict-success results are in Table 1 (ON). The Bayes filter headline is in Section 6.2, with details in Appendix C. The low strict-success regime is intentional. Under a fixed interface, low scores provide evidence of brittleness to latent-law shift, but do not by themselves establish infeasibility.

6.1 Upper Bounds and Task Feasibility

To calibrate the observed performance levels, we compute protocol-matched planning references and feasibility checks under known dynamics (see Table 2). At this scale, the objective is to diagnose transfer in a fixed, deliberately difficult regime, rather than to seek saturation-level absolute success.

Table 1: Strict success over 20 seeds. ID and OOD entries give mean success followed by ON in parentheses; brackets below give bootstrap 95% CIs for raw rates. Drop is ID−OOD. ON uses $p_{\text{oracle}} = 0.187$ and is not separately bootstrapped.

Agent	ID (ON)	OOD (ON)	Drop	OOD return
DQN	0.073 (39.04) [0.072, 0.075]	0.048 (25.67) [0.046, 0.050]	0.026 [0.024, 0.027]	0.05 [0.05, 0.05]
DQN + CURL	0.080 (42.78) [0.079, 0.081]	**0.056** (**29.95**) [0.054, 0.057]	**0.024** [0.023, 0.025]	**0.06** [0.05, 0.06]
DQN + RAD	**0.082** (**43.85**) [0.080, 0.084]	**0.056** (**29.95**) [0.055, 0.058]	0.026 [0.024, 0.028]	**0.06** [0.05, 0.06]
DreamerV3	0.062 (33.16) [0.060, 0.064]	0.036 (19.25) [0.035, 0.037]	0.026 [0.024, 0.027]	0.04 [0.04, 0.04]
PEARL-DQN	0.074 (39.57) [0.073, 0.076]	0.048 (25.67) [0.047, 0.049]	0.026 [0.025, 0.028]	0.05 [0.05, 0.05]
PPO	0.070 (37.43) [0.068, 0.072]	0.042 (22.46) [0.040, 0.044]	0.028 [0.026, 0.030]	0.04 [0.04, 0.04]
PPO + DR	0.069 (36.90) [0.068, 0.070]	0.040 (21.39) [0.038, 0.042]	0.029 [0.027, 0.031]	0.04 [0.04, 0.04]

True-dynamics MPC: With known CA dynamics, we instantiate random-shooting MPC to evaluate finite-budget sequence planning. At each control step, the planner samples candidate action sequences, rolls them forward with the true CA transition, and executes the first action of the best sequence (receding-horizon control). In the released protocol, the default budget uses horizon-8 shooting with 512 candidates per decision step, and all oracle aggregates share the same rule mix, initialization distribution, and horizon settings as learned agents. The reference is operational: it is a fixed-budget planner under matched evaluation conditions, rather than an asymptotic global solver. Under default (L=H=32) settings, the planner reaches 18.7% strict success and provides a calibration anchor for policy interpretation.

Because this planner uses a finite horizon and candidate count, it is neither a formal global-optimality ceiling nor a universal per-rule constant. At smaller scale (L=H=16), feasibility reaches 100% rule-wise in our sweep. Appendix G provides the full interpretation contract.

Table 2: Budgeted true-dynamics planning reference and task feasibility analysis.

Bound	Success Rate ↑	Final Dist ↓	AUC Dist ↓
MPC	0.187	0.376	0.414
Feasible (L=H=16)	**1.000**	–	–

Oracle-normalized score. Let p denote strict success rate under the default evaluation (same protocol as the MPC row). The budgeted true-dynamics planner achieves $p_{\text{oracle}} \approx 0.187$ under this protocol. For clarity, all oracle numbers in the main text use the same mixed rule distribution, horizon, and initialization protocol as learned agents; no per-rule reweighting or favorable initialization is applied. To map strict success onto a protocol-matched reference scale, we define

$$\text{ON}(p) \;=\; 100 \cdot \frac{p}{p_{\text{oracle}}}, \tag{1}$$

so that $\text{ON}(p_{\text{oracle}}) = 100$ when p is measured under identical conditions; values above 100 indicate performance above this budgeted planner reference, not a violation of reachability constraints. ON is therefore comparable only under matched rule composition, horizon, initialization, and planner budget; once these change, p_{oracle} must be re-estimated.

These bounds calibrate the meaning of low absolute success within the benchmark protocol. Strict success remains selective, while continuous endpoints and smaller-scale diagnostics (Section 6.5) provide complementary evidence about controllability. Main-text interpretation is conditional: at $L{=}H{=}32$, low strict success can mix latent-rule brittleness and finite-budget reachability limits, whereas $L{=}H{=}16$ feasibility rules out a blanket infeasibility explanation.

Metric interpretation contract. Strict success quantifies exact-goal controllability at episode end; final distance and AUC quantify trajectory-level proximity to the goal manifold; soft success@ε measures tolerance-based endpoint controllability. Accordingly, metric disagreement is expected in sparse-success regimes and should be interpreted as sensitivity to distinct operational targets, not as a contradiction in empirical evidence.

6.2 Headline: non-learned Bayes filter

We include a **non-learned** finite-rule Bayes filter on the same split and 20 seeds. Headline *strict* success is ID 0.2731 / OOD 0.2015 (ID–OOD gap 0.0716). Details—belief model, action objective, ablations, bootstrap table, and interpretation (vs. p_{oracle}, learned baselines, joint reading with Table 1)—are in Appendix C.

6.3 Holdout-Length Generalization

We evaluate longer-horizon transfer by testing trained agents on $H = 64$ after training on $H = 32$. Results are summarized in Table 3.

Longer-horizon evaluation reveals an additional cost of transfer. Relative to $H{=}32$, ID strict success at $H{=}64$ drops by about 4–6%, whereas OOD strict success drops by about 14–15%. The larger OOD change is consistent with horizon sensitivity. Because no protocol-matched planner is reported at $H{=}64$, we treat the pattern as diagnostic evidence rather than as a causal estimate. Among the evaluated baselines, DQN+RAD retains the strongest length transfer, while the world-model baseline degrades most under horizon mismatch.

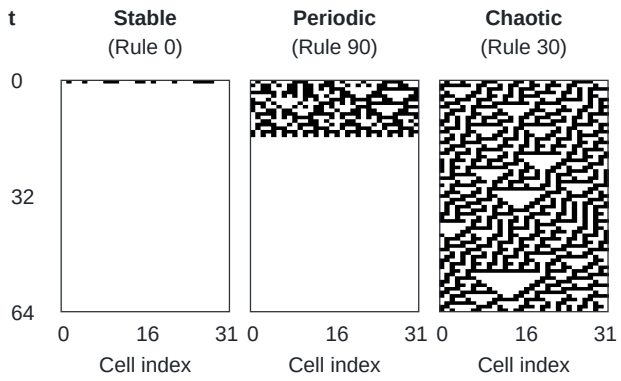

Figure 3: Tape evolution for stable (Rule 0), periodic (Rule 90), and chaotic (Rule 30) dynamics under a shared initial tape and action sequence ($L{=}32$, $H{=}64$).

Table 3: Holdout-length generalization. Entries report *success* / *final distance* (normalized Hamming). Bold denotes best and underlined denotes second best; gray columns denote holdout-rule OOD.

	ID				OOD			
	$H{=}32$		$H{=}64$		$H{=}32$		$H{=}64$	
Agent	Succ.↑	Dist.↓	Succ.↑	Dist.↓	Succ.↑	Dist.↓	Succ.↑	Dist.↓
DQN	0.073	0.927	0.070	0.930	0.048	0.952	0.041	0.959
DQN + CURL	0.080	0.920	0.076	0.924	**0.056**	**0.944**	**0.048**	**0.952**
DQN + RAD	**0.082**	**0.918**	**0.078**	**0.922**	**0.056**	**0.944**	**0.048**	**0.952**
DreamerV3	0.062	0.938	0.059	0.941	0.036	0.964	0.031	0.969
PEARL-DQN	0.074	0.926	0.070	0.930	0.048	0.952	0.041	0.959
PPO	0.070	0.930	0.067	0.933	0.042	0.958	0.036	0.964
PPO + DR	0.069	0.931	0.065	0.935	0.040	0.960	0.034	0.966

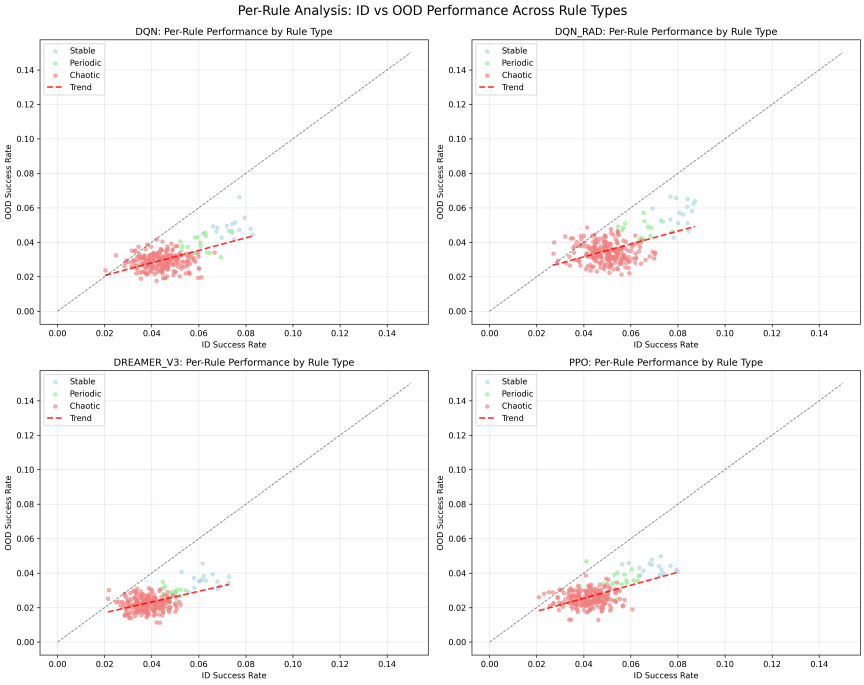

Figure 4: Per-rule performance analysis across rule types (stable/periodic/chaotic). Each point represents a CA rule; colors distinguish rule categories.

6.4 Per-Rule and Split Robustness Analysis

CA rules induce heterogeneous dynamics: stable rules approach fixed points, periodic rules oscillate, and chaotic rules produce irregular trajectories. We quantify performance across these operational categories and test split sensitivity with five partitions (three random and two farthest-point). Figure 3 visualizes the distinct rollout regimes; Figure 4 reports rule-wise performance, and Figure 5 shows OOD success across data partitions.

Table 1 reports strict success and oracle-normalized scores $\mathrm{ON}(p)$. Under this split, every learning-based family has 2.4–2.9 percentage points lower strict success on held-out rules than on training-distribution rules (Table 1). Among these methods, RAD and CURL attain the highest OOD strict success; CURL has the smallest ID$\rightarrow$OOD drop, while DreamerV3 has the lowest OOD strict success. The budgeted planner reference, $p_{\mathrm{oracle}} \approx 0.187$, provides a calibration point for these comparisons. Table 4 complements strict success with final distance, AUC distance, and soft success@0.1 at $H{=}32$. Augmentation methods attain the lowest OOD distances, consistent with better near-goal stabilization under rule shift. The world-model baseline remains more sensitive to latent-law mismatch in these measurements. Appendix G.1 records a short diagnostic for how the reference rate p_{oracle} changes with the planner's finite budget; the ON scaling used in the main text remains fixed.

6.5 Supplementary Diagnostics

Appendix B reports fixed-rule diagnostics and planner comparisons, including fixed-rule training and small-scale feasibility results (Tables 5 and 6). These results show that strict success can be limited by reachability even under known dynamics, while fixed-rule training remains learnable for some rules. Appendix H reports prediction-error and sensitivity diagnostics for the DreamerV3-style model; Appendix I discusses possible mechanisms across method families.

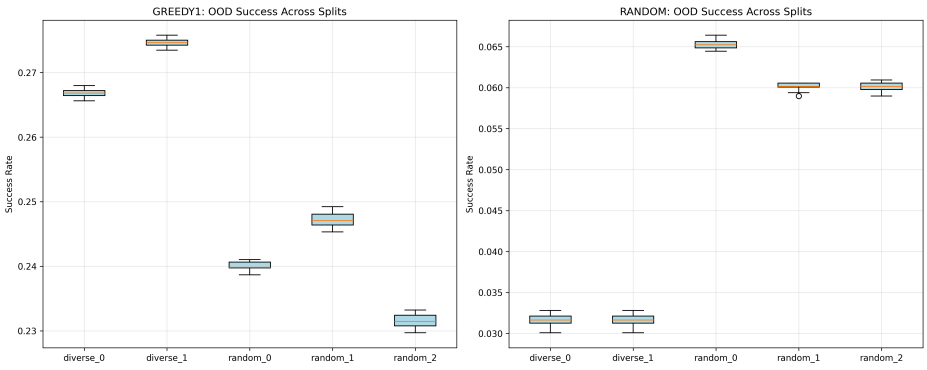

Figure 5: Split robustness across five data partitions (three random and two farthest-point splits). Box plots show each method's OOD success distribution.

Table 4: Continuous metrics at $H{=}32$, grouped by architecture family; shaded columns denote holdout-rule OOD.

	Final distance		AUC distance		Soft success@0.1	
Agent	**ID** ↓	**OOD** ↓	**ID** ↓	**OOD** ↓	**ID** ↑	**OOD** ↑
Value-based baselines						
DQN	0.927	0.952	0.934	0.957	0.132	0.072
DQN + CURL	0.920	**0.944**	0.928	0.950	0.144	**0.084**
DQN + RAD	**0.918**	**0.944**	**0.926**	**0.949**	**0.148**	**0.084**
PEARL-DQN	0.926	0.952	0.933	0.957	0.134	0.072
Actor–critic baselines						
PPO	0.930	0.958	0.937	0.962	0.126	0.063
PPO + DR	0.931	0.960	0.938	0.964	0.124	0.060
Model-based baselines						
DreamerV3	0.938	0.964	0.944	0.967	0.112	0.054

7 Discussion

The benchmark isolates a specific transfer question: what changes when the transition law is held out while the observation, action, and goal interfaces remain fixed? Under this protocol, every evaluated learning-based family loses 2.4–2.9 percentage points of strict success from ID to holdout-rule OOD (Table 1). The controlled setting makes this gap attributable to the evaluated transition-law shift, while the rule-family and horizon analyses show that its magnitude depends on the dynamical regime. These observations establish a reproducible transfer difficulty; they do not by themselves identify a unique failure mechanism.

The calibration results determine how that difficulty should be read. A budgeted true-dynamics planner reaches 18.7% strict success under the same evaluation distribution, and the finite-rule Bayes controller achieves higher absolute success while retaining an ID–OOD gap (Appendix C). The small-scale feasibility sweep reaches 100% rule-wise solvability under its own (L, H) setting. Together, these references separate three quantities that are otherwise easy to conflate: performance of a learned policy, performance of a finite-budget planner, and demonstrated reachability in a smaller regime. None is a global-optimality bound, and each comparison requires matched rule composition, horizon, initialization, and planning budget.

The continuous endpoints provide a second view of the same transfer difficulty. Augmentation-based value methods retain lower OOD distance and higher soft success, whereas the evaluated DreamerV3-

style model is more sensitive to horizon mismatch. PEARL-style context improves selected fixed-rule comparisons at smaller scale. These patterns motivate better representation transfer and latent-rule inference, but they are not module-isolated causal tests. Strict success, soft success, final distance, and AUC should therefore be reported together: they measure different operational targets when exact completion is sparse. The resulting role of TAPE is methodological: it supplies a common, calibrated protocol for testing robustness claims under a specified latent transition-law shift.

8 LIMITATIONS

Ecological scope. The deterministic 1D environment isolates latent transition-law variation but omits long-range dependencies, stochastic disturbances, and perceptual complexity found in embodied control. The evidence therefore concerns mechanism-level transfer in a controlled system, not deployment performance.

Metric semantics. Exact success is sparse at $L{=}H{=}32$. Planner references, feasibility checks, and continuous endpoints improve interpretation, but they do not characterize the complete reachable set at the default scale. A low success rate can therefore combine adaptation difficulty with unreachable targets.

Baseline coverage. The main suite spans value-based, actor–critic, task-inference, and one world-model family, together with a finite-rule Bayes diagnostic. It does not provide a systematic comparison with recurrent policies, full Bayes-adaptive planning, or large offline sequence models. The supplementary history intervention and recurrent pilots address narrower questions and do not establish dominance over memory-explicit alternatives.

Protocol sensitivity. Oracle-normalized scores depend on planner budget and evaluation composition. The main comparisons use one operating point; Appendix G.1 gives a compact sensitivity grid. Broader sweeps over planning budgets, rule-typing thresholds, and split construction remain open.

Goals and initialization. The main results use the all-zero goal and one initialization contract. Additional goals and initial-state distributions are needed to characterize target-dependent success sparsity.

Compute requirements. The 20-seed, multi-split, multi-horizon protocol is costly. Reduced-budget tracks could broaden access if they report uncertainty and identify the protocol changes needed for comparison.

9 CONCLUSION

TAPE turns rule-shift generalization into a controlled measurement problem. Disjoint cellular-automaton rules change the transition law while the interface remains fixed, allowing the observed ID–OOD gap to be studied separately from perceptual, goal, and reward shifts. Across the evaluated families, the gap is accompanied by rule-type heterogeneity and sensitivity to horizon changes. Strict and continuous endpoints, a budgeted true-dynamics reference, and a smaller feasibility regime place these observations on a calibrated scale that accounts for reachability and control difficulty.

This design also provides a practical standard for future comparisons. A robustness claim should state the held-out rule composition, report exact and near-goal outcomes together, and recalibrate its planning reference whenever the protocol changes. These requirements support comparisons of recurrent policies, belief-state controllers, and world models without treating low success as a sufficient explanation of failure. Extending the protocol to stochastic and higher-dimensional dynamics is a direct test of whether the present mechanism-level conclusions persist beyond the controlled setting.

AI USE STATEMENT

In this work, generative AI tools were utilized for grammatical polishing, sentence structure optimization, and academic expression refinement of the manuscript text, and the experimental code of this study was generated with the assistance of Cursor AI. No generative AI tools were employed for the design of research schemes, conduction of core experiments, analysis of original experimental data, and formulation of research conclusions. The original figures were prepared by the authors using PowerPoint. AI-assisted scripts were subsequently used to refine figure typography and layout; these revisions did not generate new experimental data or alter the reported numerical results. All other undisclosed AI-assisted tasks are not applicable to this study. Furthermore, large language models (LLMs) were adopted to assist in sorting out literature frameworks and optimizing the logical fluency of manuscript paragraphs. All AI-generated and AI-assisted contents in this work, including the Cursor AI-generated experimental code and polished manuscript text, have been fully reviewed, verified, and revised by multiple authors. Specifically, the authors manually checked all AI-polished text contents to eliminate logical errors, semantic deviations and potential plagiarism risks; the AI-generated experimental code has been jointly examined, tested and confirmed to be completely correct by multiple authors to ensure the accuracy and reproducibility of experimental results; all AI-optimized manuscript frameworks and narrative structures were repeatedly verified and adjusted according to the actual research content and experimental results of this study. The authors take full and ultimate responsibility for all contents of this manuscript, including all texts, research claims, experimental descriptions, code implementations, academic conclusions and visual figures produced or optimized with the assistance of generative AI tools or manually created by the authors.

ETHICS STATEMENT

This work uses synthetic cellular-automaton environments and does not involve human subjects, personal data, or interventions in deployed systems. The main research-integrity risk is over-interpreting low benchmark scores; we address it with matched planner references, feasibility checks, uncertainty reporting, and explicit limitations. The paper does not claim that benchmark scores predict real-world robustness.

REPRODUCIBILITY STATEMENT

The experimental code, configuration files, deterministic split artifacts, and evaluation scripts are provided in the supplementary materials. They support rerunning the training and evaluation pipelines, regenerating the reported tables and figures, and checking the data-processing and aggregation steps used for the results. The supplementary materials document the environment definition, budgets, random seeds, and metric conventions. Together, these artifacts identify the materials needed to reproduce the reported experiments and assess the presented claims.

REFERENCES

M. Aach, J. H. Goebbert, and J. Jitsev. Generalization over different cellular automata rules learned by a deep feed-forward neural network. *arXiv preprint arXiv:2103.14886*, 2021.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Marcin Andrychowicz, Misha Denil, Sergio G"omez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Carolin Benjamins, Theresa Eimer, and Marius Lindauer. Carl: A benchmark for contextual reinforcement learning. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL `https://arxiv.org/abs/2110.03484`.

Carolin Benjamins, Theresa Eimer, Frederik Schubert, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. Contextualize me – the case for context in reinforcement learning. *arXiv preprint arXiv:2202.04500*, 2022.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Karl Cobbe, Oleg Klimov, Christopher Hesse, Tae Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1282–1289, 2019.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. In *International Conference on Learning Representations*, 2017. arXiv:1611.02779 (2016).

Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

William Farebrother, Marlos C. Machado, and Michael Bowling. Generalization and regularization in DQN. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 2096–2103, 2018.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. URL https://arxiv.org/abs/2010.02193.

Danijar Hafner, J"urgen Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *arXiv preprint arXiv:2007.04309*, 2020.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 3207–3214, 2018.

Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

Niels Justesen, Philip Bontrager, Julian Togelius, and Sebastian Risi. Deep learning for video game playing. In *IEEE Transactions on Games*, volume 13, pp. 1–20, 2018.

Taylor W. Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter Markov decision processes. In *Advances in Neural Information Processing Systems*, 2017.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 1:16, 2021.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pp. 5639–5650. PMLR, 2020a.

Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33: 19884–19895, 2020b.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. Genrl: Multimodal-foundation world models for generalization in embodied agents. *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.18043.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing neural cellular automata. *Distill*, 5(2):e23, 2020.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free RL can be a strong baseline for many POMDPs. In *International Conference on Machine Learning*, 2022.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Seohong Park et al. Ogbench: Benchmarking offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2410.20092*, 2024.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. *Advances in neural information processing systems*, 30, 2017.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340. PMLR, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. In *arXiv preprint arXiv:1801.00690*, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Mark Towers et al. Minari: A dataset api for offline reinforcement learning. Farama Foundation, 2023. https://minari.farama.org/.

Da Wang, Lin Li, Wei Wei, Qixian Yu, Jianye Hao, and Jiye Liang. Improving generalization in offline reinforcement learning via latent distribution representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21053–21061, 2025. doi: 10.1609/aaai.v39i20.35402. URL https://ojs.aaai.org/index.php/AAAI/article/view/35402.

Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumalasetti, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, Greg Wayne, and Matthew Botvinick. Learning to reinforcement learn. In *Proceedings of the 3rd International Conference on Learning Representations*, 2017.

Jane X. Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charles Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, Gavin Buttimore, David Reichert, Neil Rabinowitz, Hubert Soyer, Russell Everett, Oriol Vinyals, Demis Hassabis, Andrew Lerchner, and Matthew Botvinick. Alchemy: A structured task distribution for meta-reinforcement learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks*, 2021. arXiv:2102.02926.

Junjie Wang, Qichao Zhang, Yao Mu, Dong Li, Dongbin Zhao, Yuzheng Zhuang, Ping Luo, Bin Wang, and Jianye Hao. Prototypical context-aware dynamics for generalization in visual control with model-based reinforcement learning. *IEEE Transactions on Industrial Informatics*, 20(11): 14578–14589, 2024. doi: 10.1109/TII.2024.3396525. URL https://doi.org/10.1109/TII.2024.3396525.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 1714–1721. IEEE, 2017.

Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of modern physics*, 55(3):601, 1983.

Stephen Wolfram and M Gad-el Hak. A new kind of science. *Appl. Mech. Rev.*, 56(2):B18–B19, 2003.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020.

Wenhao Yu, Jie Tan, C. Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. In *Robotics: Science and Systems*, 2017.

Chiyuan Zhang, Oriol Vinyals, Rémi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for Bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020.

APPENDIX ROADMAP

Supplementary sections and their starting pages.

Section	Topic	Page
A	Related work	15
B	Fixed-rule diagnostics and oracle comparisons	15
C	Finite-rule Bayes filter	16
D	Environment specification	18
E	Agents and objectives	18
F	Implementation and reproducibility	19
G	Oracle and feasibility interpretation	20
G.1	Oracle sensitivity settings	20
H	World-model credibility	21
I	Mechanistic interpretation	21
J	MPC by rule type	21
K	Information gain identities	21
L	Statistical reporting	22
M	Additional metrics and scope	23
N	Supplementary experimental analyses	24
O	Benchmark positioning	24
P	Floor-relief evaluation	25
Q	Type-balanced evaluation	26
R	Planner reference protocols	26
S	History-retention ablation	27
T	Finite-horizon rule-family labels	29
U	Environment throughput	29
V	Supplementary results and scope	30

A Related Work

OOD generalization benchmarks and protocol design. Prior work has documented generalization gaps and evaluation pitfalls in RL (Packer et al., 2018; Kirk et al., 2021; Dulac-Arnold et al., 2021), and procedural-generation benchmarks show that strong training returns do not guarantee robust transfer (Cobbe et al., 2020). Contextual-benchmark lines such as CARL (Benjamins et al., 2021) further emphasize controlled context variation with standardized evaluation protocols. Offline resources such as D4RL (Fu et al., 2020), Minari (Towers et al., 2023), and OGbench (Park et al., 2024) provide standardized datasets and evaluation protocols for coverage shift and goal-conditioned OOD, including single-task settings that remove multi-goal nonstationarity. TAPE is complementary: it keeps the interface (o, a) fixed while shifting the latent transition law z, thereby isolating rule-shift OOD in dynamics rather than data-coverage or goal-sampling effects. Visual-generalization and contextual benchmarks primarily stress observation or context variation, whereas TAPE stresses latent transition-law variation under a fixed interface. We view these axes as complementary. Earlier CA+RL formulations in synthetic domains are conceptually adjacent. Our distinction is the emphasis on protocol-calibrated OOD diagnosis—split construction, uncertainty reporting, and planner-referenced interpretation—rather than task-specific control performance alone.

Dynamics modeling and task inference. World-model methods can be sample-efficient in-distribution but remain sensitive to model error under shift (Chua et al., 2018; Hafner et al., 2019; 2023), motivating credibility-oriented analysis (Hafner et al., 2020). Latent-task inference methods such as PEARL formalize context-conditioned adaptation under hidden task variables (Rakelly et al., 2019).

Robust and Bayes-adaptive control context. Beyond benchmark construction, robust-control and Bayes-adaptive RL motivate treating latent transition uncertainty explicitly. Risk-sensitive and uncertainty-aware control perspectives emphasize policy performance under distributional mismatch, while BAMDP/PSRL-style formulations cast hidden dynamics as posterior-updated latent variables. TAPE does not instantiate full Bayes-adaptive planning in the main suite; instead, it provides a controlled setting where the gap between reactive learning and explicit latent-rule belief tracking can be measured under a shared protocol.

Evaluation calibration, CA priors, and uncertainty. Interpretable OOD claims require calibrated references and dense endpoints: feasibility analysis separates optimization failure from unreachable targets (Bertsekas, 2012), and MPC-style planning under known dynamics provides a protocol-matched planning reference (Hoi et al., 2021; Williams et al., 2017). When strict success is sparse, continuous and soft metrics retain additional discrimination (Rajeswaran et al., 2017; Andrychowicz et al., 2016). Cellular automata provide a compact but behaviorally diverse rule space for controlled generalization analysis (Aach et al., 2021; Mordvintsev et al., 2020; Chollet, 2019). Information-gain identities and uncertainty caveats are reported as reference material in Appendix K. Augmentation regularization (RAD/CURL) and domain randomization provide complementary robustness baselines in this regime (Laskin et al., 2020b;a; Tobin et al., 2017).

Transferable dynamics in continuous and offline MBRL. Complementary to TAPE's discrete, fully observed tape *states* with a latent rule z, recent work studies transferable dynamics and representation learning in continuous-control and offline regimes: multimodal foundation representations tied to generative world models (Mazzaglia et al., 2024), prototypical context-aware dynamics for visual MBRL (Wang et al., 2024), and offline RL with latent distributional structure (Wang et al., 2025). These lines inform inductive-bias expectations for world models, but they do not substitute for the explicit finite rule space and exact holdout splits used here.

B Fixed-z Diagnostics and Oracle-vs-RL Comparisons

Motivation and scope. Strict success under rule shift reflects both adaptation to unknown dynamics and rule-dependent reachability. Under irreversible CA dynamics, some goals may be unreachable even when the transition rule is known; this limits strict success but does not imply that learning without observing z is impossible when goals are reachable. We therefore examine three complementary settings: **(i)** planner-referenced outcomes where target reachability varies by rule; **(ii)** fixed-rule

training, which measures learnability without cross-episode rule variation; and **(iii)** a smaller regime with rule-wise feasibility, where learned methods can be compared with a true-dynamics planner and PEARL-style task inference.

Fixed-rule training ($L{=}H{=}32$, all-zero goal). We train DQN+RAD and DQN+CURL with a *single* rule per run (no train-time rule randomization) on six representative rules, for five independent seeds. Table 5 shows high success on stable and some periodic rules, while performance remains low on the selected chaotic rules. Thus, learnability under fixed dynamics varies across rule families even without cross-episode rule shift.

Table 5: Fixed-z training at $L{=}H{=}32$ (five seeds; point estimates; bootstrap 95% CIs are tight and omitted for space). "Strict" / "Soft@0.1" are episode success rates; "Dist" is mean Hamming distance at episode end.

Rule	Type	Agent	Strict	Soft@0.1	Dist
0	stable	CURL	1.00	1.00	0.00
0	stable	RAD	1.00	1.00	0.00
4	periodic	CURL	1.00	1.00	0.00
4	periodic	RAD	0.99	1.00	$<$0.01
108	periodic	CURL	0.00	0.02	0.23
108	periodic	RAD	0.00	0.02	0.23
204	periodic	CURL	0.00	$<$0.01	0.38
204	periodic	RAD	0.00	0.00	0.41
30	chaotic	CURL	0.00	0.00	0.50
30	chaotic	RAD	0.00	0.00	0.49
110	chaotic	CURL	0.00	0.00	0.53
110	chaotic	RAD	0.00	0.00	0.48

Small-scale feasibility and method comparisons ($L{=}H{=}16$). At this scale, all rules are feasible in principle. Table 6 reports multi-rule training with $z \sim \mathrm{Unif}(\mathcal{Z}_{\mathrm{train}})$ at each reset, fixed-z runs for the same six rules with DQN+RAD and PEARL-DQN, and true-dynamics random-shooting results. PEARL-DQN substantially improves several difficult rules (e.g., 110, 30, 204) relative to RAD under fixed z, supporting that *context-based* objectives help when z is identifiable within an episode even without explicit z given to the policy.

Takeaway. Together, these results distinguish three questions: **(A)** whether chaotic or unreachable targets remain difficult with full information; **(B)** whether nontrivial policies can be learned without exposing z when the rule is fixed; and **(C)** how oracle, model-free, and task-inference methods compare in a controlled small-scale regime. Five seeds provide directional evidence with bootstrap intervals; the twenty-seed main protocol narrows those intervals without changing the qualitative conclusions.

C Finite-Rule Bayes Filter: Formalization, Configuration, and Results

C.1 Finite-hypothesis model and role

In TAPE, a natural reference uses hypothesis space $\mathcal{Z}$ (e.g., the 256 elementary rules) with prior $p_0(z)$. Under known deterministic dynamics, the transition "likelihood" is the indicator

$$p(s_{t+1} \mid s_t, a_t, z) = \mathbf{1}[s_{t+1} = F_z(G(s_t, a_t))]$$

(stochastic relaxations replace this with a noise-smoothed likelihood; the same update structure applies). After observing a transition (s_t, a_t, s_{t+1}), Bayes' rule gives

$$p_{t+1}(z) \propto p_t(z)\, p(s_{t+1} \mid s_t, a_t, z),$$

Table 6: Small-scale feasibility and method comparisons at $L{=}H{=}16$ (five seeds; bootstrap CIs). Planner entries use budgeted true-dynamics random shooting; under multi-rule training, z is resampled each episode.

Setting	Rule / mode	Agent	Strict success
Oracle (true random shooting; multi-rule rollouts, 40 episodes)			
	multi-rule	–	0.48 [0.33, 0.63]
Oracle (true dynamics; representative rules, 20 episodes each)			
	0 (stable)	–	1.00
	4 (periodic)	–	1.00
	108 (periodic)	–	1.00
	204 (periodic)	–	1.00
	30 (chaotic)	–	0.10
	110 (chaotic)	–	0.05
Multi-rule training (monitor eval; last-K checkpoints)			
	multi	RAD	0.28 [0.25, 0.30]
	multi	PEARL	0.35 [0.33, 0.37]
Fixed z (per rule)			
	0	RAD / PEARL	1.00 / 1.00
	4	RAD / PEARL	1.00 / 1.00
	108	RAD / PEARL	0.06 / 1.00
	204	RAD / PEARL	0.00 / 1.00
	30	RAD / PEARL	0.01 / 0.95
	110	RAD / PEARL	0.02 / 0.94

with a small mismatch floor under the indicator for numerical stability. This defines a *belief-state controller*: a diagnostic midpoint between model-free RL and planning under known z, aimed at *rule-identification* fidelity at the TAPE interface. Headline strict-success numbers are summarized in Section 6.2 for quick cross-check with the main tables.

C.2 Action selection, hyperparameters, and ablations

The controller starts with a uniform belief over a finite candidate rule set and updates it after each observed transition. It scores each candidate action a by combining expected controllability with information gain about z:

$$-\mathbb{E}_{b_t}[\mathrm{dist}(s', g)] + \beta\, \mathrm{IG}(a), \qquad \beta = 0.25$$

in the default run. The first term prefers immediate goal-distance reduction under the current belief; the second contracts belief support over rules. We also evaluate *train-rule* vs. *full-rule* candidate sets to stress-test sensitivity when the support for inference is mismatched to evaluation-time rules.

C.3 How to read results vs. the planner and learned baselines

The budgeted true-dynamics reference $p_{\mathrm{oracle}} \approx 0.187$ comes from a finite-budget *random-shooting* planner, not an exhaustive solver. OOD strict success of this Bayes filter can therefore *exceed* p_{oracle} without contradicting environment reachability: the filter measures belief-guided closed-loop control with correct per-z physics, whereas p_{oracle} measures one finite-budget MPC procedure. Relative to PEARL-style task inference and DreamerV3-style world models in the main suite, the explicit finite-rule filter indicates that targeted inference about z can raise *absolute* OOD control, while an ID→OOD gap under held-out z remains. The filter uses explicit belief tracking over a finite rule set and is evaluated with its own protocol; the main table reports end-to-end learned baselines under the benchmark's standard protocol. These results therefore provide complementary evidence about rule inference and control, rather than a directly matched ranking of all methods.

Table 7: Finite-rule Bayes filter over 20 seeds. CIs use seed-level bootstrap resampling. Distances are normalized endpoint Hamming distances; returns use the benchmark's shaped reward.

Endpoint	Split	Mean	95% CI
Strict success	ID	0.2731	[0.2722, 0.2740]
	OOD	0.2015	[0.1991, 0.2039]
	ID−OOD	0.0716	[0.0692, 0.0739]
Soft success@0.1	ID	0.3896	[0.3871, 0.3918]
	OOD	0.2373	[0.2363, 0.2383]
Final distance	ID	0.3216	[0.3207, 0.3226]
	OOD	0.3626	[0.3613, 0.3639]
Return	ID	−11.3221	[−11.3423, −11.3036]
	OOD	−12.6518	[−12.6645, −12.6395]

C.4 Multi-seed aggregate (20 seeds, bootstrap 95% CIs)

Table 7 reports aggregate means and bootstrap CIs at the *seed* level, using the same resampling contract as the main benchmark (Appendix L). The strict success and ID–OOD gap rows match the headline in Section 6.2; the remaining rows add soft success@0.1, mean normalized endpoint distance, and mean undiscounted return under the same evaluation contract. At $p_{\text{oracle}} = 0.187$, the implied oracle-normalized strict scores are $\text{ON} \approx 146.0$ (ID) and $\text{ON} \approx 107.8$ (OOD). Per-seed values are provided in the supplementary experimental release; we omit the full $20\times$ grid here for readability.

D Minimal Environment Specification

This appendix gives an abstract specification of the environment; the exact observation, action, and reward definitions follow the experimental protocol.

D.1 State, latent rule, action, transition

State. A tape state is a binary vector of length L:

$$s_t \in \{0,1\}^L.$$

Latent rule. A latent rule $z \in \mathcal{Z}$ determines a CA update operator F_z.

Action. An action $a_t \in \{1,\dots,L\}$ applies an intervention operator G, then CA update:

$$\tilde{s}_t = G(s_t, a_t), \qquad s_{t+1} = F_z(\tilde{s}_t).$$

In the default instantiation, G flips the selected bit: $\tilde{s}_{t,a_t} = 1 - s_{t,a_t}$ and $\tilde{s}_{t,i} = s_{t,i}$ for $i \neq a_t$.

Observation. We expose $o_t = [s_t, t/H] \in \mathbb{R}^{L+1}$.

D.2 Reward and termination

Let g be the goal tape. Define distance $\text{dist}(s, g)$ as normalized Hamming distance. The environment provides shaped reward that is monotone in $-\text{dist}$ plus an optional success bonus at $\text{dist} = 0$. Episodes end upon success or after H steps.

E Agents and Objectives

The benchmark instantiates four reference families: model-free RL, augmentation-based RL, task inference (meta-RL), and world-model RL.

Augmentation baselines. RAD-style augmentation applies simple transforms to the observation (e.g., bit shifts and bit flips) during training. CURL-style representation learning adds a contrastive objective to encourage stable features across augmentations.

Task inference baseline. PEARL-style methods infer a latent embedding of the current rule z from a short context window of transitions, and condition the policy/Q-function on this embedding.

World-model baseline. Dreamer-style methods learn a latent dynamics model and a reward model, then optimize the policy using imagined rollouts in latent space. Under rule shift, the relevant question is whether the learned model captures structure that transfers across the held-out operators F_z.

F Implementation and Reproducibility Details

This section records the implementation details needed to reproduce the benchmark and interpret its principal empirical claims.

F.1 Elementary CA update (bitwise realization)

At time t, the agent first applies the intervention $\tilde{s}_t = G(s_t, a_t)$. An elementary CA then updates each cell using a 3-bit neighborhood read from $\tilde{s}_t$ (not from the pre-flip tape s_t). Let $\tilde{x}_{t,i}$ denote the ith cell of $\tilde{s}_t$ (wrap-around indexing), and let

$$\eta_{t,i} \;=\; (\tilde{x}_{t,i-1}, \tilde{x}_{t,i}, \tilde{x}_{t,i+1}) \in \{0,1\}^3.$$

Define $\mathrm{idx}(\eta_{t,i}) \in \{0, \dots, 7\}$ as the integer whose binary expansion *is* the 3-bit pattern (equivalently: index into the 8-bit truth table z and read off that bit). Let $x_{t+1,i}$ be the ith cell of s_{t+1}; a standard bitwise realization (indexed-bit lookup in rule code z) is

$$x_{t+1,i} \;=\; \big(z \gg \mathrm{idx}(\eta_{t,i})\big) \;\&\; 1,$$

with $\eta_{t,i}$ computed from $\tilde{s}_t$, and s_{t+1} obtained by synchronous updates over all i.

F.2 Split generation and typing (released code)

Train/test splits are generated deterministically by embedding each rule into a feature vector of density, entropy, and activity statistics from short rollouts, then applying farthest-point sampling to cover the rule space. The resulting split artifacts are retained with the supplementary materials. For operational taxonomy, we simulate short CA rollouts from random initial tapes and compute two scalars averaged over time and trials: **activity** act (mean fraction of cells that change in one CA step) and **entropy** ent of the Bernoulli marginal of the bit distribution. We classify a rule as **stable** if $\mathrm{act} < 0.06$ and $\mathrm{ent} < 0.25$; **chaotic** if $\mathrm{act} > 0.22$ and $\mathrm{ent} > 0.55$; and **periodic** otherwise. These thresholds are fixed throughout the study and are kept consistent across plots and tables.

F.3 Training budget, checkpoints, and evaluation frequency

Unless otherwise noted, each run trains for 200,000 *environment steps*. We evaluate every 10,000 steps. Each evaluation averages metrics over 20 episodes per rule on the full ID rule set and the full heldout OOD rule set. Reported baselines use $n = 20$ independent training seeds. Final scalar summaries for each seed aggregate the last $K{=}3$ evaluation checkpoints (i.e., the last three logged evaluations under the schedule above), then we bootstrap over seeds with 2,000 resamples to form 95% confidence intervals; paired-bootstrap intervals are used for ID–OOD drop. Welch-style tests are included as diagnostics, but effect sizes and confidence intervals are treated as primary evidence (see also Appendix L).

F.4 Reported metrics (definitions)

Reported metrics include success rate (fraction of episodes with $\mathrm{dist}(s, g) = 0$), soft success@ε (final distance $\leq \varepsilon$ for $\varepsilon \in \{0.03125, 0.0625, 0.1\}$), final distance (mean normalized Hamming distance at episode end), AUC distance (trajectory-averaged distance), return (mean undiscounted episodic return under the shaped reward), and ID–OOD drop (per-seed success difference aggregated with paired bootstrap). Unit conventions for distances are summarized in Appendix M.

Table 8: Diagnostic random-shooting settings for a compact oracle-sensitivity check (same $L{=}H{=}32$ mixed-rule protocol as the main text). The central column reports the headline reference used for $\mathrm{ON}(p)$; side rows are one-step budget ablations to run locally, not an alternate main-table reference.

Setting	H_{shoot}	N_{cand}	Role
primary (ON scaling)	8	512	$p_{\mathrm{oracle}} \approx 0.187$ in Table 2
ablation (a)	8	256	lower candidate cap
ablation (b)	4	512	shorter receding-horizon look-ahead

F.5 Sampling throughput and step accounting

To reduce wall-clock time, training can use multi-process vectorized environment sampling (especially helpful for world-model rollouts). Training budgets are counted in environment steps so that parallel sampling does not change the total interaction budget.

F.6 Auxiliary evaluation summaries

The supplementary materials provide the evaluation summaries used to construct the by-type MPC results in Appendix J.

F.7 Hardware and wall-clock runtime

All experiments were conducted on a single NVIDIA RTX 5090 GPU with a 25-core CPU and 90 GB RAM (system disk 30 GB, data disk 50 GB). The main benchmark runs (20 seeds) took approximately 5 days, the additional rule-type diagnostics took 2.5 days, and the Bayes filter baseline took about 5 hours.

G Oracle and Feasibility Interpretation Details

For strict success (an exact goal match), rates below 100% can arise under irreversible or chaotic dynamics because some target tapes are unreachable from random initial states within horizon H, even when z is known. The MPC entry in Table 2 is computed with a finite-budget random-shooting planner and should be interpreted as a budgeted planning reference under the evaluation distribution, not a formal global-optimality ceiling. The default $p_{\mathrm{oracle}} \approx 18.7\%$ is therefore a mixed-distribution empirical reference that aggregates rules with different reachable-set geometry and planner-search difficulty. At smaller scale ($L{=}H{=}16$), rule-wise feasibility reaches 100% in our sweep, indicating that the environment family is not intrinsically unsolvable; rather, solvability depends on the joint regime (state space size, horizon, evaluation distribution) and planning budget. By rule type, true-dynamics planning is typically strongest on stable rules, intermediate on periodic rules, and weakest on chaotic rules, consistent with trajectory divergence and horizon-limited controllability.

G.1 Budgeted reference: diagnostic sensitivity to random-shooting parameters

The headline $p_{\mathrm{oracle}} \approx 0.187$ in Table 2 uses random-shooting with planning horizon 8 and 512 candidate action sequences per control decision, evaluated under the same mixed-rule protocol as learned agents. Because this reference is itself finite-budget, a minimal robustness check is to vary one budget component at a time. Table 8 names two standard ablations: halving the candidate count (e.g., 512→256 at fixed look-ahead) or halving the planning horizon (e.g., 8→4 at fixed candidate count). Such reductions are expected to lower the finite-budget success rate and therefore to lower any $\mathrm{ON}(p)$ that uses the ablated p_{oracle}. We do *not* retabulate a full sweep here: the main paper fixes one primary operating point for cross-method comparability, while the supplementary materials support reproducing the short grid on a fixed test split to audit sensitivity to search budget.

Table 9: True-dynamics MPC strict success by operational rule type (test split). n denotes the number of episode rollouts in each bucket.

Type	**Strict success**	**Mean return**	n **episodes**
Stable	0.50	-15.19	2
Periodic	0.17	-15.08	6
Chaotic	0.00	-16.37	22
All (test)	0.067	-16.03	30

H DreamerV3 Credibility Details

Our DreamerV3-style baseline is intended as a representative latent-dynamics model, not an exhaustive sweep across world-model families. The credibility checks target three axes. First, error propagation: one-step prediction error is higher on OOD rules (0.15 vs 0.12 on ID), and rollout error increases sharply with planning depth (about 4× by H=16). Second, sensitivity: in a focused sweep, increasing imagination horizon beyond 10 yields limited gains under rule shift, indicating diminishing utility when model misspecification dominates planner depth. Third, calibration against external reference: the persistent gap to true-dynamics MPC (about 3.6% vs 18.7% strict success) across seeds supports a representation-transfer bottleneck rather than a single-run artifact. These diagnostics do not establish impossibility for world models. They instead constrain interpretation: under this protocol and model family, the observed degradation is consistent with compounding model bias under latent-law shift.

I Mechanistic Interpretation Details

The results support a three-part mechanistic interpretation of the observed transfer gap. Model-free baselines primarily aggregate response regularities over training-era laws; augmentation regularizes invariance and can reduce brittle dependence on local patterns. Task-inference baselines explicitly infer a context-conditioned latent embedding and condition value estimation on that inferred task identity. World-model baselines learn a latent transition operator and optimize through imagined trajectories, which can amplify representation error as rollout depth increases. Within this lens, the observed ID→OOD degradation is interpreted as a mismatch between policy optimization strength and latent-law identification fidelity. Improving on this benchmark therefore requires both stronger action optimization and better-calibrated inference of the latent mechanism under distributional shift.

J True-Dynamics MPC: Strict Success by Operational Rule Type

Table 9 reports true-dynamics MPC (ensemble planner with known z) on the **holdout test** split, with L=H=32 and one episode per rule for fast replication. The headline MPC rate $\approx 18.7\%$ in Table 2 uses the project's full evaluation budget and may pool over additional episodes; this table isolates *by-type* variability on the same protocol. Under our split, test rules are predominantly chaotic (22/30), so the *aggregate* strict success is dominated by the chaotic bucket.

K Information Gain: Core Identities

Scope. This section records compact identities used to interpret exploration and rule inference; these identities are not part of the default training or testing pipeline for the benchmark metrics (see Sec. 3).

Let z be a latent rule with posterior $p(z \mid \mathcal{D}_t)$ after history $\mathcal{D}_t$, and let S' be the next state after (s, a).

Definition K.1 (Information Gain)**.**

$$\mathsf{IG}(s,a) = \mathsf{H}(z \mid \mathcal{D}_t) - \mathrm{E}_{S' \sim p(\cdot \mid s,a,\mathcal{D}_t)}\big[\mathsf{H}(z \mid \mathcal{D}_t \cup (s,a,S'))\big].$$

Theorem K.2 (IG equals conditional mutual information)**.** $\mathsf{IG}(s,a) = \mathsf{I}(z; S' \mid s, a, \mathcal{D}_t)$.

Proof. Fix $(s, a, \mathcal{D}_t)$ and take expectation with respect to $S' \sim p(\cdot \mid s, a, \mathcal{D}_t)$. By the entropy form of conditional mutual information,

$$\begin{aligned}\mathsf{I}(z; S' \mid s, a, \mathcal{D}_t) &= \mathsf{H}(z \mid s, a, \mathcal{D}_t) - \mathsf{H}(z \mid S', s, a, \mathcal{D}_t) \\ &= \mathsf{H}(z \mid \mathcal{D}_t) - \mathrm{E}_{S' \sim p(\cdot \mid s, a, \mathcal{D}_t)}\big[\mathsf{H}(z \mid S', s, a, \mathcal{D}_t)\big],\end{aligned}$$

where the second line uses that (s, a) is conditioned/fixed for the decision query. Now apply Bayes' rule for posterior updating after observing one transition:

$$p(z \mid S', s, a, \mathcal{D}_t) = p(z \mid \mathcal{D}_t \cup (s, a, S')).$$

Hence

$$\mathsf{H}(z \mid S', s, a, \mathcal{D}_t) = \mathsf{H}\big(z \mid \mathcal{D}_t \cup (s, a, S')\big),$$

so

$$\mathsf{I}(z; S' \mid s, a, \mathcal{D}_t) = \mathsf{H}(z \mid \mathcal{D}_t) - \mathrm{E}_{S'}\big[\mathsf{H}(z \mid \mathcal{D}_t \cup (s, a, S'))\big] = \mathsf{IG}(s, a),$$

which is exactly Theorem K.1. □

Theorem K.3 (IG equals expected posterior KL)**.**

$$\mathsf{IG}(s, a) = \mathrm{E}_{S' \sim p(\cdot \mid s, a, \mathcal{D}_t)}\big[\mathrm{KL}(p(z \mid \mathcal{D}_t \cup (s, a, S')) \,\|\, p(z \mid \mathcal{D}_t))\big].$$

Proof. From Theorem K.2, $\mathsf{IG}(s, a) = \mathsf{I}(z; S' \mid s, a, \mathcal{D}_t)$. Using the KL form of conditional mutual information with $(U, V, W) = (z, S', (s, a, \mathcal{D}_t))$,

$$\mathsf{I}(z; S' \mid s, a, \mathcal{D}_t) = \mathrm{E}_{S' \sim p(\cdot \mid s, a, \mathcal{D}_t)}[\mathrm{KL}(p(z \mid S', s, a, \mathcal{D}_t) \,\|\, p(z \mid s, a, \mathcal{D}_t))]\,.$$

Again, for fixed (s, a) we have $p(z \mid s, a, \mathcal{D}_t) = p(z \mid \mathcal{D}_t)$, and after one observed transition, $p(z \mid S', s, a, \mathcal{D}_t) = p(z \mid \mathcal{D}_t \cup (s, a, S'))$. Substituting gives

$$\mathsf{IG}(s, a) = \mathrm{E}_{S' \sim p(\cdot \mid s, a, \mathcal{D}_t)}\big[\mathrm{KL}(p(z \mid \mathcal{D}_t \cup (s, a, S')) \,\|\, p(z \mid \mathcal{D}_t))\big].$$

□

K.1 What IG does not imply under rule shift

The identities above are *exact* under a well-specified Bayesian model, but they do *not* imply that maximizing IG improves reward or OOD transfer. High IG only means S' is informative about z; in TAPE, goal-reaching actions need not maximize IG. Under learned models, "IG" is computed from an approximate posterior and can be miscalibrated under holdout rules; under nonstationary z, exploration may not align with long-horizon return. For full derivations and additional examples, see standard references on Bayesian experimental design and mutual information; we omit lengthy textbook material here.

L Statistical Reporting Details

L.1 Bootstrap confidence intervals over seeds

Let $x_1, \ldots, x_n$ be final performance values per seed (e.g., OOD success averaged over the last K checkpoints). A bootstrap CI is obtained by resampling $\{x_i\}_{i=1}^n$ with replacement B times, computing the mean each time, and taking the 2.5% and 97.5% quantiles of these bootstrap means.

L.2 Paired bootstrap for ID–OOD drop

For each seed i, compute $d_i = x_i^{\mathrm{ID}} - x_i^{\mathrm{OOD}}$. Bootstrap resample the paired tuples (equivalently, resample the d_i directly) and compute CIs on the mean drop. This preserves correlation between ID and OOD within a seed.

L.3 Hypothesis tests (diagnostic only)

We include Welch-style tests comparing OOD seed distributions between methods as a diagnostic. Given multiple comparisons, a conservative option is Holm correction; in this paper we emphasize effect sizes and CIs as primary evidence.

M Additional Metrics, Scope, and Future Directions

M.1 Continuous metric definitions and unit consistency

We distinguish normalized and raw distances explicitly:

$$d_{\text{norm}}(s,g) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{1}[s_i \neq g_i] \in [0,1], \qquad d_{\text{raw}}(s,g) = \sum_{i=1}^{L} \mathbf{1}[s_i \neq g_i] \in \{0,\dots,L\}.$$

They satisfy $d_{\text{raw}} = L \cdot d_{\text{norm}}$. In this paper, strict success is always $\{d_{\text{norm}} = 0\}$ (equivalently $\{d_{\text{raw}} = 0\}$), and soft success@ε uses the normalized threshold $d_{\text{norm}} \leq \varepsilon$. Reported "final distance" and "AUC distance" are normalized unless a caption explicitly states raw units.

If readers observe near-equality between "final distance" and $1 -$ success in some settings, that indicates a near-binary endpoint distribution rather than a metric definition bug. Specifically, with D as endpoint normalized distance and $S = \mathbf{1}[D = 0]$, we have

$$\mathbb{E}[D] = (1 - \mathbb{E}[S]) \cdot \mathbb{E}[D \mid D > 0].$$

When failures cluster near maximal distance, $\mathbb{E}[D \mid D > 0] \approx 1$, so $\mathbb{E}[D] \approx 1 - \mathbb{E}[S]$. This is a property of outcome geometry under the current regime, not an identity enforced by the metric definition.

M.2 Oracle-normalized score stability and scope

The oracle-normalized score $\text{ON}(p) = 100\, p/p_{\text{oracle}}$ is conditional on the evaluation protocol used to estimate p_{oracle} (rule mix, horizon, initial-state distribution, and planner budget). Therefore, ON values should be compared within matched protocols; across protocol shifts (e.g., different rule-type composition or $H = 64$), the correct reference is a re-estimated p_{oracle} under that same condition. Values above 100 indicate outperforming the specific budgeted planner reference under the matched protocol. We avoid claims of cross-protocol invariance for ON without this recalibration.

M.3 Unexplored baseline families

The current evaluation does not include recurrent policy/value backbones (e.g., PPO–LSTM or DRQN), belief-explicit or Bayes-adaptive RL beyond the finite-rule filter in Appendix C, or additional world-model inductive biases such as discrete latents and automata encoders. Broader task-inference methods, including attention-based context and hierarchical codes, are also outside the present suite. These comparisons would extend the evidence beyond the method families evaluated here.

Varying the number of training rules and the length or informativeness of context would help characterize rule-coverage sample complexity and identifiability of z from transition histories. The present results do not resolve these questions.

M.4 Interpreting cross-table comparisons

By-type MPC and the global planner reference use different evaluation distributions and episode budgets; their rates are not directly interchangeable without protocol matching. Final and AUC distances use normalized Hamming units unless a caption states otherwise. At $L{=}32$, a normalized-distance difference of 0.02–0.03 corresponds to about 0.64–0.96 cells at episode end and should be read alongside uncertainty and calibrated references. In sparse-success regimes, the main strict-success and continuous-metric tables provide complementary views. Split-robustness plots describe variation in OOD performance across partitions; they do not establish a fixed split-wise ID–OOD drop for every partition.

M.5 Compute reporting and accessibility track

The default training and evaluation schedule is summarized in Appendix F. For accessibility, a low-cost track can be pre-registered without new method claims: reduced seed count, reduced checkpoint frequency, fixed heldout rule subset, and mandatory reporting of wall-clock time alongside interaction steps. This preserves comparability while lowering entry cost for compute-constrained labs.

M.6 Reproducibility materials

The supplementary materials provide the experimental code, configurations, deterministic split artifacts, random seeds, and environment specification needed to reproduce the reported evaluations. The evaluation pipeline supports regenerating the main tables and figures, including planner and rule-stratified summaries. Appendix F defines the environment, training budgets, metrics, and aggregation procedures used by these materials.

N Supplementary Experimental Analyses

This appendix reports additional protocol details and experiments that complement the main evaluation. The sections are organized by experimental role. Section O records the intended use of TAPE and a shift-type comparison to neighboring benchmarks. Section P reports a floor-relief evaluation at $L{=}H{=}16$. Section Q reports a type-balanced split and rule-family stratification. Section R maps every reported planner reference to its exact protocol. Section S is the primary addition: a matched Full-history vs. Reset-history ablation under a unified 20-seed protocol, with DQN and Stage-A references. Section T clarifies finite-horizon Stable / Periodic / Chaotic-like labels. Section U reports environment throughput. Section V summarizes the evidence boundaries of the supplementary results.

Unless a caption states otherwise, "success" denotes strict success (normalized Hamming distance zero at episode end). Five-seed pilots are directional; the 20-seed history suite is the primary statistical evidence.

O Intended Use and Benchmark Positioning

Diagnostic questions. TAPE is a controlled mechanism-level diagnostic for methods evaluated under latent transition changes with a fixed observation, action, and reward interface. It is designed to answer:

- Does retaining interaction history improve adaptation to a latent transition rule?
- Is low OOD performance caused by inherent control difficulty or by failure to adapt to held-out rules?
- Is failure concentrated in Stable, Periodic, or Chaotic-like rule families?
- Does improved control coincide with improved inference of the latent rule?

TAPE complements realistic visual and continuous-control benchmarks. Scores on TAPE characterize the specified rule-shift protocol and are not interpreted as predictions of real-world robustness. Its value is internal diagnostic precision: the (o, a, r) interface is fixed, latent transition rules are discrete and characterizable, and performance can be separated by rule family.

Table 10: How representative OOD / contextual benchmarks distribute shift types. TAPE holds visual, layout/goal, and reward interfaces fixed while varying latent dynamics.

Benchmark	Visual	Layout/ goal	Reward	Dynamics only
ProcGen (Cobbe et al., 2020)	✓	✓	✓	—
Meta-World (Yu et al., 2020)	—	✓	✓	—
Alchemy (Wang et al., 2021)	✓	✓	✓	—
DMC visual gen. (Tassa et al., 2018; Hansen et al., 2020)	✓	—	—	—
TAPE (this work)	fixed	fixed	fixed	✓

Relation to neighboring lines of work. TAPE is related to contextual and hidden-parameter MDP formulations, which model environment variation through observed or latent context variables (Hallak et al., 2015; Killian et al., 2017). It studies a restricted diagnostic setting in which the observation, action, and reward interfaces remain fixed, while a discrete latent transition rule varies across

episodes and selected rules are held out. Contextual RL benchmarks such as CARL (Benjamins et al., 2021; 2022) systematically vary environment context across broader parameterizations; our goal is complementary—discrete, exactly characterizable transition rules with held-out-rule and rule-family diagnostics, rather than broad coverage of realistic continuous parameters. A closely related line couples universal policies with online system identification (Yu et al., 2017), and dynamics-randomization methods train across distributions of physical parameters for sim-to-real robustness (Tobin et al., 2017; Peng et al., 2018). Relative to these approaches, TAPE focuses on controlled generalization to held-out transition rules and on localizing which rule families benefit from interaction history. The Full-history variant connects to recurrent Meta-RL (Duan et al., 2017), latent-context methods such as PEARL (Rakelly et al., 2019) and VariBAD (Zintgraf et al., 2020), and recurrent model-free POMDP baselines (Ni et al., 2022; Hausknecht & Stone, 2015). Because the inferencer is trained with training-rule bit labels, we characterize it as supervised open-support rule inference with history-dependent control; it is evaluated alongside, rather than identified with, label-free Meta-RL. Closest conceptual relatives among multi-factor benchmarks include Alchemy (Wang et al., 2021) (latent-structure inference with simultaneous visual/reward variation) and ProcGen (Cobbe et al., 2020) / Meta-World (Yu et al., 2020) (entangled layout, goal, and appearance shifts). Together, these design choices make TAPE a controlled complement to broader contextual benchmarks: it fixes the (o, a, r) interface, uses discrete and characterizable latent transition rules, and evaluates held-out rules by rule family.

P FLOOR-RELIEF EVALUATION AT $L{=}H{=}16$

P.1 MOTIVATION

At $L{=}H{=}32$, strict success for several learned methods clusters near 3–6%. Headline ID→OOD drops can then look like gaps between two near-zero numbers. We therefore re-evaluate a matched algorithm subset in a less sparse regime.

P.2 PROTOCOL

Table 11: Protocol for the floor-relief pilot.

Item	Setting
Rule split	Main-suite rule split (same identities as the main suite)
Tape / horizon	$L{=}H{=}16$
Algorithms	DQN (Mnih et al., 2015), PPO (Schulman et al., 2017), RAD-DQN (Laskin et al., 2020b), PEARL-DQN (Rakelly et al., 2019)
Training seeds	5 (directional pilot)
Evaluation	ID and OOD rules; strict success and soft success@0.1
Planner reference	true-dynamics random shooting (Williams et al., 2017; Hoi et al., 2021), plan horizon 8, 512 candidates
Matched p_{oracle}	0.343 under this mixture and scale

P.3 RESULTS

Table 12: Floor-relief results ($L{=}H{=}16$, 5 seeds). Drop = ID−OOD strict success.

Method	ID	OOD	Drop	Soft drop
DQN	0.176	0.124	+0.052	+0.059
PPO	0.161	0.123	+0.038	+0.051
RAD-DQN	0.106	0.084	+0.022	+0.030
PEARL-DQN	0.177	0.120	+0.058	+0.064

DQN's ID success of 0.176 corresponds to an oracle-normalized score of about 51% of the matched planner, confirming that agents learn non-trivial policies once lifted from the near-zero floor. Directional ID>OOD persists for all four methods in Table 12. We treat $L{=}16$ as a complementary diagnostic regime alongside the harder $L{=}32$ stress setting; the 20-seed history suite below also uses $L{=}H{=}16$.

Q Type-Balanced Split and Rule-Family Stratification

Q.1 Motivation

The default farthest-point split places $22/30$ Chaotic-like rules in the test set. A micro-average over that split conflates "held-out" with "intrinsically harder."

Q.2 Protocol

Table 13: Protocol for the type-balanced pilot.

Item	Setting
Type mix	Stable : Periodic : Chaotic-like = 1:1:1
Test rules	8 per type (24 total)
Train rules	16 per type (48 total)
Scale	$L{=}H{=}32$
Algorithm	DQN (primary); PPO / RAD / PEARL also run under the same split
Seeds	5 (directional)
Matched planner	$p_{\text{oracle}}{=}0.375$ under this mixture

Q.3 Results

Table 14: Type-stratified DQN under the balanced split ($L{=}32$, 5 seeds).

Rule type	ID	OOD	Drop	Oracle MPC
Stable	0.437	0.623	-0.186	0.625
Periodic	0.158	0.058	$+0.100$	0.375
Chaotic-like	0.010	0.041	≈ 0	0.125

Three regimes appear. Stable rules transfer (OOD can be easier). Periodic rules show a large positive ID$\rightarrow$OOD drop where learning is still possible. Chaotic-like rules remain near floor on both splits. Equal-weight macro-averaging of this pilot can reverse the aggregate gap because Stable OOD dominates; the scientifically meaningful readout is therefore per-type (or equal-type macro as a sensitivity check), not a single chaotic-dominated headline.

R Planner Reference Protocols

R.1 Motivation

Reported planner rates include 0.187, 0.067, and values under composition ablations. They must not be conflated: each is tied to a budget, rule mixture, and episode count.

Table 15: Mapping of reported planner values to protocols. ON denominators use only the protocol-matched budgeted reference.

Value	Protocol	Role
≈ 0.187	Full-budget random shooting matched to the main RL evaluation	ON denominator
≈ 0.067	Test-split-only export, 1 episode/rule	Diagnostic only; **not** ON denominator
0.260	Imbalanced test-side, 20 ep/rule, plan 8×512	Composition sensitivity
0.375	Type-balanced test-side, 20 ep/rule, plan 8×512	Composition sensitivity
0.343	Imbalanced $L{=}H{=}16$, plan 8×512	Floor-relief reference
0.339 / 0.348	Imbalanced $L{=}16$ ID/OOD under the history-suite protocol	History-suite MPC reference

R.2 Protocol map

R.3 Interpretation

p_{oracle} is a finite-budget true-dynamics random-shooting reference, not a theoretical success ceiling. Exceeding a budgeted planner under a matched protocol is not a contradiction. "Feasibility $= 100\%$" means each rule has at least one solvable initial configuration in the small-scale regime; it is not a success rate under the default evaluation budget. ON scores are comparable only within a fixed (budget, rule mixture, L, H) configuration.

S History-Retention Ablation

S.1 Motivation and POMDP view

The latent rule z is episode-fixed and unobserved, while the tape state is observed. TAPE is therefore a contextual POMDP in the sense of hidden-context MDPs (Hallak et al., 2015; Killian et al., 2017): competent agents should use interaction history to infer z while balancing information gain and control. Feedforward baselines alone cannot isolate the effect of history retention. Following recurrent Meta-RL and POMDP practice (Duan et al., 2017; Rakelly et al., 2019; Zintgraf et al., 2020; Ni et al., 2022; Hausknecht & Stone, 2015), we construct a matched Full-history vs. Reset-history intervention and report a standard DQN (Mnih et al., 2015) and a true-rule Stage-A conditional-Q reference under the same seeds.

S.2 Agent design

- **Stage-A conditional Q.** A conditional DQN $Q(o, z, a)$ is trained with access to the true 8-bit rule vector on the training-rule split, then frozen.
- **Open-support inferencer.** A GRU reads (o_t, a_{t-1}, r_{t-1}) and outputs a factorized posterior over the 8 rule bits (support over all 256 ECA rules).
- **Decision mode.** At evaluation, the agent scores actions by a QMDP-style average of conditional advantages under the inferred posterior.
- **Full vs. Reset.** Both variants share Stage-A Q, inferencer architecture, budgets, training-rule bit supervision, and evaluation rules. Full retains GRU state and belief across timesteps; Reset clears them every timestep.

No OOD-rule trajectories are used for training, hyperparameter selection, early stopping, or checkpoint selection. We characterize the method as supervised open-support rule inference with history-dependent control; its training uses rule-bit supervision.

Table 16: Unified protocol for the 20-seed history suite.

Item	Setting
Rule split	Main-suite rule split
Environment	$L{=}H{=}16$
ID / OOD rules	60 / 30
Episodes per rule	20
Training seeds	$0, \ldots, 19$ (complete set; no exclusions)
Stage-A / DQN steps	$5{\times}10^5$ environment steps
Inferencer steps	$3{\times}10^5$ (controller frozen)
Primary decision mode	QMDP-style posterior averaging
Uncertainty	paired-seed bootstrap, $B{=}10{,}000$
Comparison unit	independently trained seed (episodes averaged first)

S.3 Unified protocol

S.4 Aggregate results

Table 17: Twenty-seed aggregate under the protocol of Table 16. Stage-A is a true-rule-conditioned reference, not an optimal-control upper bound. MPC is a finite-budget planner reference.

Method	ID success	OOD success
DQN	0.1477 $[0.1103, 0.1864]$	0.1157 $[0.0796, 0.1558]$
Reset-history QMDP	0.3185 $[0.3149, 0.3220]$	0.2665 $[0.2588, 0.2740]$
Full-history QMDP	$\mathbf{0.3823}$ $[0.3745, 0.3893]$	$\mathbf{0.3218}$ $[0.3122, 0.3308]$
True-rule Stage-A reference	0.4411 $[0.4388, 0.4436]$	0.2757 $[0.2687, 0.2826]$
Finite-budget MPC (ID/OOD)	0.3392	0.3483

Table 18: Paired Full$-$Reset differences. All 20 seeds have positive ΔID and positive ΔOOD. Seed-level ΔID ranges from $+0.0275$ to $+0.0925$; ΔOOD from $+0.0133$ to $+0.0783$.

Comparison	ID	OOD
Full $-$ Reset (point)	$+0.0638$	$+0.0553$
95% paired-bootstrap CI	$[0.0565, 0.0708]$	$[0.0471, 0.0627]$
Macro-type Full $-$ Reset	$+0.0556$ $[0.0486, 0.0620]$	$+0.0252$ $[0.0211, 0.0291]$
ID$-$OOD gap (Full / Reset)	0.0604 / 0.0520	—

S.5 Type-level readout

The OOD history benefit is concentrated in Chaotic-like rules. For Periodic OOD, $14/20$ seeds have nonzero Full$-$Reset differences (approximately -0.033 to $+0.017$) that cancel in the mean, so there is no systematic average history benefit for that family. Equal-type macro-averaging reduces the OOD lift from $+0.0553$ to $+0.0252$ while preserving sign.

S.6 Paired comparisons to DQN

Full and Reset outperform DQN on ID for all 20 seeds. Full outperforms DQN on OOD for all 20 seeds. Reset outperforms DQN on OOD for $19/20$ seeds (seed 5 is -0.025). Full vs. Reset remains the primary history intervention; DQN provides broader baseline context.

Table 19: Type-stratified Full/Reset means (20 seeds). OOD rule counts: Stable 2, Periodic 6, Chaotic-like 22. Stable OOD equals 0.500 for both methods because the two OOD Stable rules evaluate to $\{0, 1\}$.

Type	ID/OOD n	Full ID	Reset ID	Full OOD	Reset OOD
Stable	8 / 2	0.5531	0.5241	0.5000	0.5000
Periodic	20 / 6	0.4975	0.4296	0.3217	0.3217
Chaotic-like	32 / 22	0.2675	0.1976	0.3057	0.2302

Table 20: Paired bootstrap comparisons with DQN (align by seed, then bootstrap the 20 differences).

Paired difference	ID	OOD
Full − DQN	$+0.2345$ $[0.1927, 0.2732]$	$+0.2062$ $[0.1662, 0.2428]$
Reset − DQN	$+0.1707$ $[0.1298, 0.2088]$	$+0.1508$ $[0.1116, 0.1871]$

S.7 Auxiliary inference metrics

Full−Reset mean rule-bit accuracy differences are approximately $+0.125$ (ID) and $+0.098$ (OOD). We treat these as end-to-end mechanism evidence rather than strict causal isolation of inference quality, because the two controllers may induce different state-visitation distributions.

S.8 Finite-rule Bayes filter (context)

As an additional reference, the non-learned finite-rule Bayes filter at $L{=}32$ (20 seeds) reaches ID 0.273 / OOD 0.202 (gap $+0.072$), substantially above feedforward DQN absolute rates while retaining a held-out gap. This separates "memory/inference helps ID" from "held-out rules remain hard even with explicit belief tracking over the training library."

S.9 Recurrent policy pilots

Under $L{=}H{=}16$ (5 seeds), DRQN-style GRU agents (Hausknecht & Stone, 2015) and RL^2-style PPO (Duan et al., 2017) did not dominate a matched DQN under the budgets tested. This five-seed pilot does not constitute a comprehensive sequence-model comparison. Carefully tuned recurrent agents (Ni et al., 2022) and modern sequence models remain outside the evaluated suite.

T Finite-Horizon Rule-Family Labels

Because each instance has a finite deterministic state space, every trajectory eventually enters a periodic orbit. Strict infinite-time aperiodicity is therefore impossible. Our labels describe behavior inside a finite observation window:

- **Stable:** reaches a fixed point within the window;
- **Periodic:** a non-unit global recurrence is detected within the window;
- **Chaotic-like / long-transient:** no such stabilization or recurrence is observed before the horizon.

A trajectory labeled Chaotic-like may still become periodic after the episode ends. The taxonomy is not evidence of mathematical chaos in the dynamical-systems sense; it remains behaviorally relevant because attractors reached only after episode termination cannot be exploited during the episode.

U Environment Throughput

Environment transition throughput (CPU, without learner overhead):

Table 21: Environment step throughput.

Tape length	Throughput	Wall-clock per 10^6 steps
L=16	5.02×10^4 steps/s	≈ 20 s
L=32	5.21×10^4 steps/s	≈ 19 s

V Supplementary Results: Interpretation and Scope

The history-retention comparison improves absolute ID and OOD control under a matched Stage-A Q across all 20 seeds (Tables 17–18). Its OOD benefit varies by rule family and is concentrated in Chaotic-like rules; equal-type macro-averaging reduces the OOD lift from 0.0553 to 0.0252 while preserving its sign. Type-stratified results distinguish rule-family difficulty from aggregate holdout performance, and the L=16 regime provides a less sparse complementary evaluation. The planner references remain budgeted protocol-matched baselines, not theoretical ceilings. TAPE is designed as a controlled diagnostic that complements broader benchmarks.

These experiments do not establish that history reduces the ID–OOD gap: the Full-history gap is 0.0604, compared with 0.0520 for Reset-history. They do not establish uniform OOD gains across rule families, optimality of Stage-A or MPC, label-free Meta-RL, or prediction of real-world robustness from benchmark scores. These conclusions are specific to the methods and protocols evaluated here; feedforward comparisons alone do not establish how all RL methods behave without history-aware controls.