Reset-Window Limits for Latent World-Model Interfaces

Anonymous authors
Paper under double-blind review

Abstract

We study the data interface that underlies a class of latent world models: a homogeneous two-state controlled process observed through binary emissions. Given only N independent reset triples, how much post-training control loss is unavoidable over a fresh H-step episode? For binary actions and observations, a known uniform reset, and reward $(1 + X)/2$, we characterize the boundary-uniform loss for $H \geq 8, N \geq 1$ as $\min\{H, H^{3/2}N^{-1/4}, H^3N^{-1/2}\}$ up to absolute constants. The lower permits adaptive training and useful deployment-time inference; the upper uses thirteen intervention moments, a one-dimensional legal-model fit, and a finite belief controller. A disclosed two-model construction shows how the reset window changes the information available at equal observation count. The result is an interface theorem for controlled latent processes, not a neural representation-learning benchmark.

1 Introduction

World models are useful only insofar as predictions support decisions under actions that were not observed in the data. This creates a data-interface question that can be hidden behind a representation: how much long-horizon control information is available when every training trajectory is forced to stop after a short reset window? We isolate this question in a small, fully specified controlled latent process. The result is about the statistical interface of a world model, not about whether a learned latent representation is better than an observable recurrent controller.

Training consists of N independent episodes, each with three observations and two effective actions. The fresh deployment episode has H observations and permits a policy to use its entire public history. We study minimax expected deployment policy loss uniformly over a class in which the observation signal and transition drift may approach zero. This targets control loss rather than latent-coordinate recovery: parameter identification can be ill-conditioned while all policies have nearly identical value.

Our contributions are:

1. A boundary-uniform joint characterization at the rate $F(N, H) = \min\{H, H^{3/2}N^{-1/4}, H^3N^{-1/2}\}$,

$$\frac{F(N,H)}{176160768} \leq R_{\mathrm{sym}}(N,H) \leq R_{\mathrm{gen}}(N,H) \leq CF(N,H), \qquad H \geq 8,\ N \geq 1. \tag{1}$$

   The lower is on the centered subclass and the upper on the full class; the same observable-history policy quantifier is used in both risks.

2. A constructive legal-quotient learner based on thirteen intervention moments and a finite stored-pointer belief controller, together with a disclosed window-separation construction that explains why total observation count is not a sufficient interface summary.

The formal object is deliberately smaller than a modern neural world model: the latent process is the environment, the moment fit is the predictive model, and the stored-pointer policy is the planner. Representation learning, high-dimensional observations, and continuing control are outside the claim.

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.

The reset is known, uniform, independent, and fresh; rewards and kernels are fixed; and parameters may depend on N, H in lower instances. We do not prove an unknown-reset, continuing-control, arbitrary-rank, fixed-floor, neural, or population result. The accompanying artifact record preserves failed branches and these boundaries instead of hiding them in final prose.

2 Controlled latent-process interface

Let $S_t, X_t \in \{-1, +1\}$ be hidden state and observation and let $A_t \in \{0, 1\}$. A legal homogeneous model has

$$\mathbb{P}(X_t = x \mid S_t = s) = \frac{1 + x(\mu + \rho s)}{2}, \qquad \rho \geq 0, \quad |\mu| + \rho \leq 1, \tag{2}$$

$$\mathbb{E}[S_{t+1} \mid S_t = s, A_t = a] = b_a + \beta_a s, \qquad |b_a| + |\beta_a| \leq 1. \tag{3}$$

Every episode starts from an independent uniform hidden state. The first observation is rewarded before the first effective action; all H observations contribute reward $(1+X_t)/2$, so deployment has $H-1$ effective actions. The lower bound permits adaptive training actions within and across episodes, but the constructive upper learner uses independent uniform action logging (known propensity $1/2$) and no reset is available inside an episode.

For learner $\mathcal{A}$, let V_M^* be the best expected deployment return among total observable-history policies and let $V_M(\pi_{\mathcal{A}})$ be the return of its output policy. Define

$$R_{\mathcal{C}}(N, H) = \inf_{\mathcal{A}} \sup_{M \in \mathcal{C}} \mathbb{E}\left[V_M^* - V_M(\pi_{\mathcal{A}})\right]. \tag{4}$$

The outer expectation includes training and learner randomness; the value uses an independent deployment episode. Let $\mathcal{C}_{\text{gen}}$ be the full legal class and let $\mathcal{C}_{\text{sym}}$ fix only $\mu = 0$. The policy class is identical in the two risks, including histories impossible under a fitted model. Thus the theorem is neither a hidden-state oracle comparison nor a parameter-estimation statement. We use the abbreviations $R_{\text{gen}}(N, H) = R_{\mathcal{C}_{\text{gen}}}(N, H)$ and $R_{\text{sym}}(N, H) = R_{\mathcal{C}_{\text{sym}}}(N, H)$ below.

2.1 The interface question

The object being learned is not an abstract latent vector. It is a legal predictive interface: after a training transcript, the learner must output a public-history policy whose action-conditioned observation law is useful on a new reset episode. This separates three quantities often conflated in world-model discussions:

1. representation error: whether a hidden coordinate is recovered;
2. prediction error: whether a specified action-observation test is predicted under deployment; and
3. control loss: the value gap after optimizing a policy in the fitted model.

Our theorem concerns only the third quantity. Two legal models can have different parameterizations while inducing the same policy values, whereas a small error in a weak action-conditioned moment can be amplified over H steps. The learner therefore needs a stable quotient of the model class, not an arbitrary canonical latent coordinate system.

The interface in Figure 1 has two deliberate asymmetries. The reset is available between training episodes but not within an episode, and deployment can use the entire public history even when training does not identify the hidden sign. These asymmetries produce the intermediate fourth-root regime and make a generic supervised prediction guarantee insufficient.

2.2 Observable moments and the legal quotient

The thirteen moments in the upper-bound construction are a sufficient reset-accessible collection for the action-conditioned drift terms needed by the controller. We do not claim

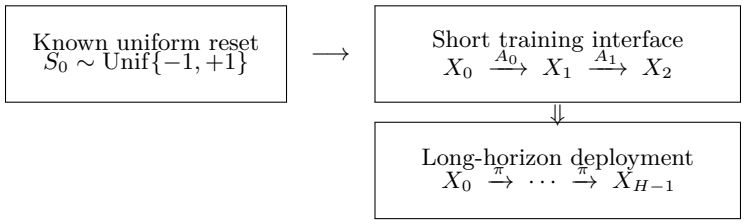

Figure 1: The reset-triple interface. Training reveals short action-conditioned tests; deployment evaluates a policy on a fresh long episode. The theorem measures the value gap, not latent-coordinate recovery.

Table 1: Reset-accessible moment families used by the constructive learner.

Family	Observable statistic	Controlled quantity
Offset	$\mu = \mathbb{E}X_0$	emission offset
One-step mean	$P_a = \mathbb{E}[X_1 \mid do(a)]$	$C_a = \rho b_a$
One-step product	$Q_a = \mathbb{E}[X_0X_1 \mid do(a)]$	$M_a = \rho^2\beta_a$
Two-step mean	$R_{ab} = \mathbb{E}[X_2 \mid do(a,b)]$	$D_{ab} = C_b + \beta_b C_a$
Two-step product	$S_{ab} = \mathbb{E}[X_0X_2 \mid do(a,b)]$	$N_{ab} = \beta_b M_a$

that thirteen is information-theoretically minimal. The following table records their roles, for $a, b \in \{0, 1\}$.

The products are important at the boundary: division by ρ is illegal when the emission contrast vanishes, and division by ρ^2 is unstable when it is merely small. The fitter therefore searches a grid of legal contrasts and compares raw moments after projection onto the transition diamond. This is a finite-dimensional quotient construction: it preserves the moments relevant to control while allowing hidden-coordinate non-identifiability.

3 Joint minimax law

Theorem 1 (reset-triple control law). For integers $H \geq 8$ and $N \geq 1$, the risks in (4) obey (1), where

$$C = 1 + \max\{\sqrt{22C_{\mathrm{fast}}}, 3C_{\mathrm{fast}}\}, \qquad C_{\mathrm{fast}} = 29 + 407\sqrt{32\log 26}. \tag{5}$$

Consequently both risks have order $F(N, H)$ by class inclusion: the lower is proved on $\mathcal{C}_{\mathrm{sym}}$ and the upper on $\mathcal{C}_{\mathrm{gen}}$. The rate joins are $N = H^2$ and $N = H^6$; H^4 is only a join used by one lower construction.

The lower uses three controlled families. For $N \leq H^2$, an absorbing action-label pair gives loss $H/8192$. For $H^2 \leq N \leq H^4$, weak-flip label uncertainty gives $H^{3/2}N^{-1/4}/176160768$; the argument permits useful hidden-sign inference during deployment. A clipped weak-signal pair and budget monotonicity cover $H^4 \leq N \leq H^6$, and the same pair gives $H^3N^{-1/2}/22020096$ for $N \geq H^6$. These likelihood arguments allow adaptive training and do not impose a blind deployment policy.

3.1 Reading the rate

The theorem has three regimes, each with a different statistical mechanism. The phase diagram is more informative than the single minimum notation:

The joins follow directly from equating adjacent terms: $H = H^{3/2}N^{-1/4}$ at $N = H^2$, and $H^{3/2}N^{-1/4} = H^3N^{-1/2}$ at $N = H^6$. The H^4 threshold appearing in the proof is not a fourth statistical regime; it is where one explicit lower construction is replaced by a clipped continuation. This distinction matters when interpreting finite-budget experiments.

Table 2: Rate regimes for the reset-triple interface.

Training budget	Dominant scale	Interpretation
$1 \le N \le H^2$	H	The action label remains uncertain; one wrong long-lived choice costs a constant fraction of the horizon.
$H^2 \le N \le H^6$	$H^{3/2}N^{-1/4}$	Weak action-conditioned drift is hard to estimate, but deployment observations can still reveal useful hidden-sign information.
$N \ge H^6$	$H^3N^{-1/2}$	Ordinary moment error is propagated through a length-H controller; the cubic horizon factor is the planning sensitivity.

3.2 Proof roadmap and lower-bound intuition

The upper proof has three interfaces. First, under the independent-uniform logging contract, concentration controls the raw moment vector with bounded inverse-propensity weights. (Adaptive logging is permitted only in the lower bound's likelihood argument.) Second, the legal projection transfers raw moment error to a model whose transition coefficients satisfy the same diamond constraints as the truth. Third, a stored-pointer belief controller transfers model error to value error. The first two steps produce $w \lesssim N^{-1/2}$; the final step has both a boundary-insensitive $H^3 w$ bound and a signal-sensitive pair of bounds. The minimum of those transfers is what produces the stated rate.

The lower proof mirrors these interfaces. The low-budget construction hides an action label in an almost absorbing pair. The middle construction makes the label statistically close while preserving a weak but decision-relevant drift; the deployment policy is allowed to use its observations, so the argument is not an open-loop reduction. The high-budget construction clips the same weak pair and uses a two-point change-of-measure bound. Adaptive action choices factor out of the likelihood calculation, which is why the lower applies to the full training policy class rather than only to uniform logging.

3.3 Constructive upper

The learner uses independent uniform actions while training. Thirteen intervention moments are sufficient:

$$\mu, \quad P_a = \mathbb{E}[X_1 \mid do(a)], \quad Q_a = \mathbb{E}[X_0 X_1 \mid do(a)],$$
$$R_{ab} = \mathbb{E}[X_2 \mid do(a,b)], \quad S_{ab} = \mathbb{E}[X_0 X_2 \mid do(a,b)].$$

Writing $C_a = P_a - \mu$, $M_a = Q_a - \mu P_a$, $D_{ab} = R_{ab} - \mu$ and $N_{ab} = S_{ab} - \mu R_{ab}$ gives

$$C_a = \rho b_a, \qquad M_a = \rho^2 \beta_a, \qquad D_{ab} = C_b + \beta_b C_a, \qquad N_{ab} = \beta_b M_a.$$

Weighted empirical versions have coordinates bounded by four, so their maximum error E_N satisfies $\mathbb{E}E_N \le \sqrt{32\log(26)/N}$.

The fit searches $m+1$ emission contrasts with $m = \lceil 13\sqrt{N} \rceil$, clips the offset to its legal interval, projects each transition onto the legal diamond, and selects the candidate with smallest maximum residual over all thirteen raw moments. The zero-contrast branch is separate, so no unstable division is hidden at $\rho = 0$. For a grid point within $1/m$ of the true contrast, projection inequalities give

$$w \le 377/m + 407E_N, \qquad \mathbb{E}w \le C_{\text{fast}} N^{-1/2},$$

where w is the maximum raw-moment error. This is a legal-model approximation, not latent-state identification.

For an exact fitted optimizer, three transfer bounds cover the boundary:

$$\text{Regret} \le 2(H-1)s, \qquad \text{Regret} \le 11H^2 w/s, \qquad \text{Regret} \le K_H w,$$

where $s = \max_a(|C_a|, |M_a|)/\rho$ when nonzero and $K_H = H(2H-1)(3H-1)/2 \le 3H^3$. If the denominator vanishes, all true policies have equal value. Balancing the first two bounds yields the fourth-root branch; the third yields the large-budget branch. A finite

belief-grid controller with stored successor pointers adds at most $1/m$. The arithmetic and bit complexity are polynomial for genuine empirical inputs, although constants and dense planning cost are not claimed practical.

The executable learner has four fixed stages: inverse-propensity estimates of the thirteen moments, a legal contrast mesh with a zero-contrast branch, residual-based model selection after transition projection, and a stored-pointer belief controller on the selected model. All tie-breaking and total-policy conventions are fixed before evaluation; the full tabular contract is in the supplement. This keeps the theorem constructive without spending main-text space on implementation pseudocode.

4 Why the observation window matters

Equal observation counts do not imply equal reset or computation budgets. We make this precise only on a disclosed pair, not on the full class. For each $H \geq 8$, let $\alpha = 1/8192$, $\rho_H = \alpha/\sqrt{H}$, common drift $b_a = 1/H$, and let an unknown label θ select which action has $\beta = 1 - 2/H$ rather than $1 - 3/H$. The learner knows this pair and only the label is unknown. Compare $n_3 = H^4$ reset triples with $n_H = 3H^3$ full-H episodes; both use $B_H = 3H^4$ observations.

Proposition 2 (disclosed pair). For the separate known-pair risks,

$$R_3^{\mathrm{pair}}(H) \geq \frac{\sqrt{H}}{22020096}, \qquad R_H^{\mathrm{pair}}(H) \leq \alpha\sqrt{H}\exp\left(-\frac{3\alpha^2 H^2}{131072}\right). \tag{6}$$

Thus their ratio tends to zero. The short-window lower is an adaptive training-law overlap argument. For the long-window upper, one training action and a terminal-observation mean test separate the labels: a recurrence gives terminal hidden-mean gap at least $1/128$, and Hoeffding gives the exponent in (6).

This is an asymptotic interface construction with conservative constants; for the displayed $\alpha = 1/8192$, the exponential separation is not a practical sample-size forecast. The learners have different reset counts, effective action counts, computation, and model knowledge; transitions and signal depend on H. It does not order $R_{\mathcal{C}}$ for the full unknown class and is not a practical benchmark.

Table 3: Resource ledger for Proposition 2. Only total observations are matched; the other rows are deliberately not matched.

Resource	Triples	Full-window episodes
Total observations	$3H^4$	$3H^4$
Reset episodes	H^4	$3H^3$
Effective actions	$2H^4$	$3H^3(H-1)$
Pair knowledge	disclosed	disclosed
Full-class minimax ordering	not claimed	

The proposition should therefore be read as a warning against treating total observation count as a sufficient description of a world-model data interface. It is not a ranking of two practical training pipelines.

5 Boundary and related work

The theorem is deliberately a controlled-interface result, not a claim about arbitrary latent rank. A disclosed homogeneous three-state pair has transition entries at least $1/12$ and emission probabilities at least $1/8$ while its complete adaptive triple laws coincide for every finite N; at $H = 4$ its known-pair loss is $67/6144$. This boundary example rules out silently promoting the two-state rate to a general-rank statement.

Our closest theory neighbors are belief-state and predictive-state planning, spectral HMM/PSR identification, finite-sample POMDP learning, and history-dependent off-policy

evaluation (Shani et al., 2005; Pineau et al., 2006; Kurniawati et al., 2008; Kaelbling et al., 1998; Littman et al., 2001; Boots et al., 2011; Hsu et al., 2012; Grinberg et al., 2018; Jiang et al., 2017; Jin et al., 2020; Efroni et al., 2022; Uehara et al., 2022; Tennenholtz et al., 2020; Nair and Jiang, 2021; Zhang and Jiang, 2025; Jordan and Kamgarpour, 2026). Those works use different trajectory, coverage, or feedback contracts; none is the reset-triple post-training value problem studied here. The latent-MDP and controlled-world-model literature studies representation or transition identifiability under richer excitation, while the neural lineage (World Models, PlaNet, Dreamer, MuZero, TD-MPC, DreamerV3 and successors) evaluates learned representations with sampled interaction (Kwon et al., 2021; Zhang et al., 2026; Ha and Schmidhuber, 2018; Hafner et al., 2019; 2020; Schrittwieser et al., 2020; Hansen et al., 2022; 2024; Hafner et al., 2023). We cite these systems to define the connection, not as empirical baselines or evidence of latent superiority.

Table 4: Comparison by data interface rather than model branding.

Line of work	Data/objective	Difference here
POMDP / PSR	Belief or future-test sufficiency	No reset-triple minimax law
Spectral HMM / PSR	Rank or coverage assumptions	Different estimator and data law
Online / offline RL	Long trajectories or behavior coverage	Different feedback and target
Neural world models	Pixel prediction and sampled control	No neural or population claim

6 Deterministic implementation study

The theorem is accompanied by a closed-world float64 implementation audit. Every row uses a supplied record or exact population moment; no episode is sampled and no row is interpreted as a population estimate. We compare the complete thirteen-moment fit with endpoint-product and product-free masks using the same stored-pointer planner and exact public-history evaluator. The audit contains 15 legal worlds, 32 ordered $N = 1$ records per world, an exact mechanism fixture, an exhaustive 528-pair follow-up, and one-axis ladder and parameter-grid stress tests; full rows and receipts are in the supplement.

Table 5: Deterministic supplied-record implementation screen. Each world value averages 32 ordered records and the displayed mean averages 15 fixed legal worlds (480 executions per arm). These are exact finite-fixture diagnostics.

Screen	Coverage	full13	no_endpoint	no_products
Fixed-world replay	15 worlds, 32 supplied records/world, $H = 8$	0.51564	0.50966	0.50966
Relative outcome	same cells	wins 2; losses 6; ties 7	–	–
Mechanism fixture	exact moments, $H = 8$	0	7305/131072	7/16
Receipt parity	normal vs optimized replay	byte-identical; exact occupancy checks pass		

The mixed replay is a deliberate negative result: the complete interface is not a universal finite-cell winner, and the two ablations can share a controller. On the exact fixture, however, removing endpoint products produces regret 7305/131072 and removing all products produces 7/16, while the true-model planner has zero regret. The 528-pair screen finds distinct fitted hashes on 36 pairs but tied exact regrets on 459, separating parameter identifiability from control relevance. The supplement also reports 84 ladder rows, 45-cell transport grids, 675 logging-support rows, and the dense 132-cell audit; these are stress tests, not learning curves.

6.1 Monte Carlo calibration of the reset-triple learner

The deterministic screens above verify algebra and pointer execution, but they do not show how the released learner behaves on independent training streams. We therefore ran a separate, local Monte Carlo calibration using the same three arms and exact public-history evaluator. The runner hardcodes four fixed legal mechanisms (interior, weak emission, weak transition, and near boundary), $N \in \{1, 4, 16, 64, 256\}$, 32 independent replicates per cell,

and $H = 8$. Within each replicate all arms receive the same sampled reset triples. The unit of analysis is the replicate, not the individual episode. The exact evaluator computes the public-history oracle once per mechanism; fitted controllers are then scored by an independent exact occupancy pass.

Table 6: Local Monte Carlo calibration at $H = 8$. Entries are mean public-history regret over 4 fixed mechanisms and 32 paired replicates per mechanism; parentheses show the descriptive replicate SD after pooling the fixed cells. This is an exploratory local result, not a population or minimax estimate.

N	full13	no_endpoint	no_products
1	0.761 (0.199)	0.742 (0.194)	0.742 (0.194)
4	0.512 (0.286)	0.547 (0.311)	0.698 (0.299)
16	0.276 (0.219)	0.365 (0.298)	0.620 (0.359)
64	0.152 (0.056)	0.149 (0.058)	0.553 (0.392)
256	0.113 (0.066)	0.113 (0.062)	0.475 (0.299)

The calibration is consistent with a finite-sample improvement in the complete/product-aware interface on the near-boundary and weak-transition cells, while the weak-emission cell has uniformly small gaps because the observation channel carries little action-relevant signal. The complete arm is not uniformly best: no_endpoint is marginally lower on the interior cell at $N = 64$, and the fixed mechanisms do not identify a single winner at every budget. Figure 2 retains the per-mechanism curves and descriptive 95% replicate intervals. This grid does not validate the theorem's minimax constants: at fixed $H = 8$, the second asymptotic join is at $N = H^6 = 262{,}144$, far beyond the largest evaluated budget.

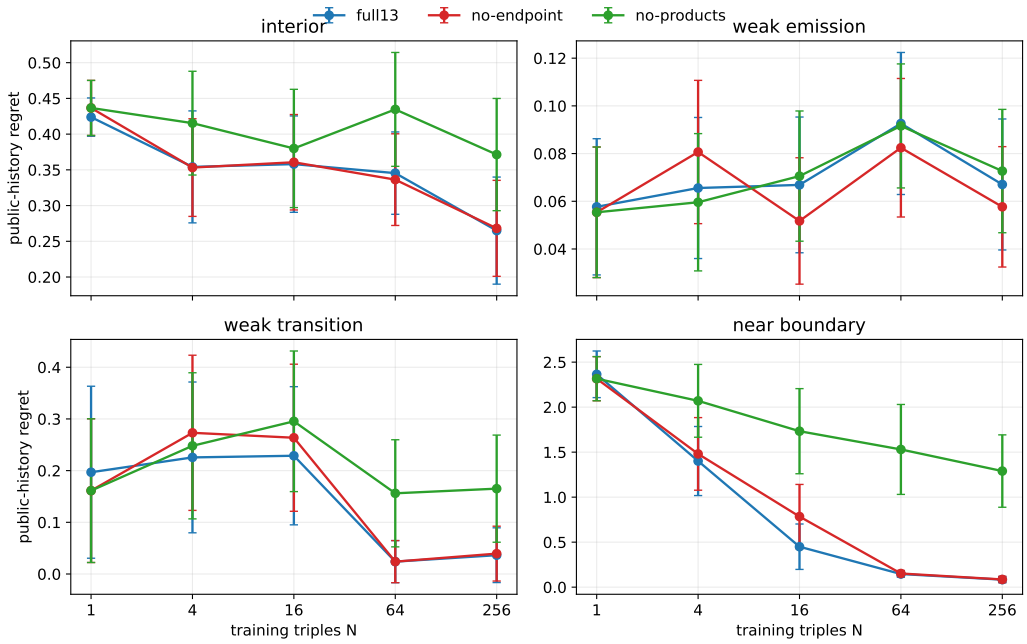

Figure 2: Paired Monte Carlo calibration of the exact reset-triple learner. Points are means over 32 independent training replicates and error bars are descriptive normal-approximation intervals. All arms share a replicate's training stream. The four panels are fixed legal mechanisms, not population samples; no asymptotic slope is fitted.

The complete receipt contains 1,920 arm-level rows (4 mechanisms $\times$5 budgets $\times$32 replicates $\times$3 arms), all training-stream hashes, model/controller hashes, exact regrets, and the independent occupancy check. It is labeled EXPLORATORY_LOCAL_COMPLETED_NO_ADMISSION: the stdlib random source has no external signer or native D10 admission. The result therefore supplies corresponding implementation evidence for the learner/planner construction, but it is not evidence

for latent superiority, population generalization, a neural benchmark, or minimax-rate validation.

6.2 Exploratory visual world-model audit

As a separate implementation check, we trained a convolutional encoder, action-conditioned latent transition, decoder, and reward head on a 16×16 raster environment. The earlier in-distribution screen has 243 closed-loop rows, 216 rollout rows, and 54 training histories over three worlds, three seeds, $N \in \{32, 128, 512\}$, and planning depths $1, 3, 5$. It is explicitly EXPLORATORY_NOT_ADMITTED: the model-free diagnostic used privileged expert labels and the reference is a heuristic. The public supplement keeps the training, rollout, and control curves; the held-out follow-up below is the primary visual result.

6.3 Paired held-out-world transfer audit

The earlier visual screen is a useful smoke test but leaves two confounds: its worlds are all used for fitting, and its model-free comparator uses privileged expert labels. We therefore ran a frozen paired follow-up with a separate evaluation protocol. Each of the three training worlds is fit at $N \in \{64, 256, 1024\}$ with three seeds; each fitted checkpoint is evaluated on its own world and on three fixed parameter-shifted worlds. Every arm sees the same 24 initial states and observation-noise tape within a cell. The primary model-free comparator, reward_greedy, predicts immediate rewards from the random replay buffer and receives no expert actions or hidden states. The reported quantity is the signed return gap to the matched one-step reference; positive values mean lower return, and the reference is a heuristic rather than an optimal-control certificate.

Table 7: Paired visual transfer audit. Entries are mean signed return gap ± sample SD over fixed train-world/evaluation-world/seed cells at planning depth 3 for the two latent planners and depth 0 for the non-planning controls. The same-world block has 9 cells per arm and the held-out block has 27 cells; episode rows are not treated as independent seeds. Lower is better and zero is the matched heuristic by construction.

Evaluation split and N	latent8	wide16	reward_greedy	random
Same world, 64	0.772 ± 0.525	0.772 ± 0.525	0.134 ± 0.087	4.180 ± 0.751
Same world, 256	0.987 ± 0.551	0.772 ± 0.525	0.100 ± 0.069	4.180 ± 0.751
Same world, 1024	0.772 ± 0.525	0.772 ± 0.525	0.117 ± 0.082	4.180 ± 0.751
Held out, 64	0.881 ± 0.135	0.881 ± 0.135	0.744 ± 0.230	4.020 ± 0.455
Held out, 256	1.015 ± 0.287	0.881 ± 0.135	0.744 ± 0.222	4.020 ± 0.455
Held out, 1024	0.881 ± 0.135	0.881 ± 0.135	0.756 ± 0.232	4.020 ± 0.455

The completed run produces 972 cell-level control rows, 864 endpoint-rollout rows, 81 serialized fits, and 729 logged update rows. The result is a negative but informative boundary: both latent planners are worse than the reward-only comparator on held-out worlds, and the latent8 curve is not monotone in N. The two latent widths are nearly indistinguishable in control despite lower validation frame error for wide16; prediction loss and control value therefore separate under this budget. These rows are not a claim of world-model failure in general: the environment is a small synthetic raster fixture and the reference is heuristic. They do rule out presenting the current implementation as evidence for latent superiority.

We added a post-hoc checkpoint action audit to explain the repeated curves. It reloaded all 81 saved fits and replayed 324 learned-policy cells (7,776 episodes) with exact action-trace parity. Every wide16 held-out cell and every latent8 held-out cell at $N \in \{64, 1024\}$ executed $+1$ at all 12 steps; at $N = 256$, latent8 did so on 21/27 held-out cells. A fixed $+1$ control therefore matches the latent arms on many cells, while the reward-only arm remains non-constant. This is a finite-fixture policy-behavior diagnostic, not proof of encoder collapse, and it further rules out interpreting the current curves as learned latent control evidence.

7 Limitations and process boundary

The scientific claim remains narrow: known uniform resets, fixed binary alphabets, a two-state class, parameter-dependent lower instances, no continuing control, no unknown reward,

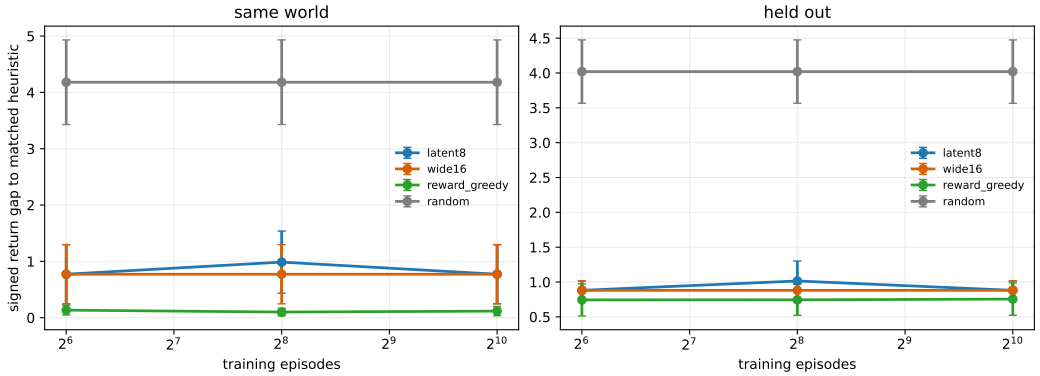

Figure 3: Paired transfer control curves from the completed v3 campaign. Error bars are sample SD over fixed cells. The held-out panel evaluates three worlds whose target, inertia, or observation noise was not used for fitting. The reward-only pixel baseline is below both latent planners on the held-out split; the random control is shown to expose metric scale.

and no fixed-floor theorem. The visual neural runs are exploratory and not admitted; there is no trusted sampled population campaign, proof-assistant certificate, or external novelty clearance. ARA and DAG views document provenance and intermediate research state; they are not latent-state evidence or theorem certificates. The strongest conclusion is therefore an interface theorem with audited deterministic implementation diagnostics, not a general claim about learned world-model advantage.

8 Conclusion

We characterize the conditioning-uniform cost of long-horizon control from initial reset triples in a two-state controlled latent process. The result shows how a short reset window creates a fourth-root intermediate regime even when deployment may infer useful hidden information, and why matching total observations alone does not erase the cost of the window. The disclosed three-state boundary and recorded negative routes define where this conclusion stops. A neural or general latent-representation claim requires a separate admitted experiment and is outside the present evidence.

AI use statement

Generative AI assisted idea exploration, derivations, code checks, literature retrieval, drafting, and process-archive preparation. The author remains responsible for mathematical arguments, citations, interpretation, and final submission review; local checks are not independent peer review or certification.

Ethics and reproducibility

The study uses finite-state models and synthetic supplied-input fixtures, with no human subjects, personal data, or consequential deployment. The model class, access contract, estimator, controller, constants, exact replay receipts, and scope limitations are released with the source and supplement. The exploratory pixel campaign is explicitly non-admitted and does not establish neural superiority, population generalization, or external novelty.

References

B. Boots, S. M. Siddiqi, and G. J. Gordon. Closing the learning-planning loop with predictive state representations. The International Journal of Robotics Research, 30(7):954–966, 2011.

Y. Grinberg, H. Aboutalebi, M. Lyman-Abramovitch, B. Balle, and D. Precup. Learning predictive state representations from non-uniform sampling. In Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018.

W. Hamilton, M. M. Fard, and J. Pineau. Efficient learning and planning with compressed predictive states. Journal of Machine Learning Research, 15(1):3395–3439, 2014.

D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. Journal of Computer and System Sciences, 78(5):1460–1480, 2012.

D. Ha and J. Schmidhuber. World models. arXiv:1803.10122, 2018.
T. Kipf, E. van der Pol, and M. Welling. Contrastive learning of structured world models. arXiv:1911.12247, 2019.
N. Hansen, H. Su, and X. Wang. Temporal difference learning for model predictive control. In Proceedings of the 39th International Conference on Machine Learning, volume 162, pages 8380–8393. PMLR, 2022.
N. Hansen, H. Su, and X. Wang. TD-MPC2: Scalable, robust world models for continuous control. In International Conference on Learning Representations, 2024.
D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Learning latent dynamics for planning from pixels. arXiv:1811.04551, 2019.
D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. Dream to control: Learning behaviors by latent imagination. In International Conference on Learning Representations, 2020.
R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. In Proceedings of the 37th International Conference on Machine Learning, volume 119. PMLR, 2020.
D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Mastering Atari with discrete world models. In International Conference on Learning Representations, 2021.
V. Micheli, E. Alonso, and F. Fleuret. Transformers are sample-efficient world models. In International Conference on Learning Representations, 2023.
H. Zhou, C. Pan, Y. LeCun, and L. Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. arXiv:2411.04983v2, 2024.
Y. Zhu, M. Simchowitz, V. Gadipudi, and A. Gupta. RePo: Resilient model-based reinforcement learning by regularizing posterior predictability. arXiv:2309.00082v2, 2023.
L. Kaiser et al. Model-based reinforcement learning for Atari. In Advances in Neural Information Processing Systems, 32, 2019.
C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 2309–2318. PMLR, 2019.
N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Proceedings of the 34th International Conference on Machine Learning, volume 70, pages 1704–1713. PMLR, 2017.
C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. In Advances in Neural Information Processing Systems, 33:18530–18539, 2020.
L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 101(1–2): 99–134, 1998.
J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor. RL for latent MDPs: Regret guarantees and a lower bound. In Advances in Neural Information Processing Systems, 34, 2021.
Q. Liu, A. Chung, C. Szepesvari, and C. Jin. When is partially observable reinforcement learning not scary? arXiv:2204.08967, 2022.
M. L. Littman, R. S. Sutton, and S. Singh. Predictive representations of state. In Advances in Neural Information Processing Systems, 14, 2001.
Y. Nair and N. Jiang. A spectral approach to off-policy evaluation for POMDPs. arXiv:2109.10502, 2021.
S. Singh, M. R. James, and M. R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pages 512–519. AUAI Press, 2004.
G. Tennenholtz, S. Mannor, and U. Shalit. Off-policy evaluation in partially observable environments. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2020.
M. Uehara, A. Sekhari, J. D. Lee, N. Kallus, and W. Sun. Provably efficient reinforcement learning in partially observable dynamical systems. arXiv:2206.12020, 2022.
W. Zhan, M. Uehara, W. Sun, and J. D. Lee. PAC reinforcement learning for predictive state representations. arXiv:2207.05738, 2022.
Y. Efroni, C. Jin, A. Krishnamurthy, and S. Miryoosefi. Provable reinforcement learning with a short-term memory. arXiv:2202.03983, 2022.
K. Abraham, Z. Naulet, and E. Gassiat. Fundamental limits for learning hidden Markov model parameters. arXiv:2106.12936v3, 2022.
K. Abraham, E. Gassiat, and Z. Naulet. Frontiers to the learning of nonparametric hidden Markov models. arXiv:2306.16293v2, 2023.
F. Chen, Y. Bai, and S. Mei. Partially observable RL with B-stability: Unified structural condition and sharp sample-efficient algorithms. arXiv:2209.14990v2, 2022.
F. Chen, H. Wang, C. Xiong, S. Mei, and Y. Bai. Lower bounds for learning in revealing POMDPs. arXiv:2302.01333v1, 2023.
R. Huang, Y. Liang, and J. Yang. Provably efficient UCB-type algorithms for learning predictive state representations. arXiv:2307.00405v3, 2024.
D. Hafner, J. Pasukonis, J. Ba, and T. Hafner. Mastering diverse domains through world models. arXiv:2301.04104, 2023.
P. Jajoo. Exponential hardness of off-policy evaluation under history-dependent logging. arXiv:2609.19135v1, 2026.
P. Jordan and M. Kamgarpour. Model-based learning of near-optimal finite-window policies in POMDPs. arXiv:2604.01024v2, 2026.
S. M. Kakade, A. Krishnamurthy, G. Mahajan, and C. Zhang. Learning hidden Markov models using conditional samples. arXiv:2302.14753v2, 2024.
J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature, 588:604–609, 2020.
W. Ye et al. Mastering Atari games with limited data. In Advances in Neural Information Processing Systems, 34, 2021.
M. Baltieri, F. Torresan, Y. Zhang, A. Boyd, and F. E. Rosas. World models of environment, agent and joint agent-environment systems. arXiv:2608.20401v1, 2026. Scope comparison only.
X. Zhang, Y. Guan, B. Zhang, H. Li, Y.-Q. Zhang, and S. E. Li. On the identifiability of controlled world models. arXiv:2607.22430v2, 2026.
Y. Zhang and N. Jiang. Statistical tractability of off-policy evaluation of history-dependent policies in POMDPs. ICLR 2025; arXiv:2503.01134v1.
G. Shani, J. Pineau, and R. Kaplow. A survey of point-based POMDP solvers. Autonomous Agents and Multi-Agent Systems, 12:1–37, 2005.

J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In Proceedings of IJCAI, pages 1025–1032, 2003.
H. Kurniawati, D. Hsu, and W. S. Lee. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In Robotics: Science and Systems IV, 2008.
M. Karl, M. Soelch, J. Bayer, and P. van der Smagt. Deep variational Bayes filters: Unsupervised learning of state space models from raw data. In ICLR, 2017.
R. Kidambi, R. Rajeswaran, P. Netrapalli, and T. Joachims. MOReL: Model-based offline reinforcement learning. In NeurIPS, 33, 2020.
T. Yu, A. Kumar, Y. Rafailov, A. Rajeswaran, S. Levine, and C. Finn. MOPO: Model-based offline policy optimization. In NeurIPS, 33, 2020.
T. Yu, A. Kumar, R. Rafailov, J. Rajeswaran, S. Levine, and C. Finn. COMBO: Conservative offline model-based policy optimization. In NeurIPS, 34, 2021.
L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.
O. Cappé, E. Moulines, and T. Rydén. Inference in Hidden Markov Models. Springer Series in Statistics. Springer, 2005.
A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. Journal of Machine Learning Research, 15:2773–2832, 2014.
G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. Biometrika, 103(2):423–434, 2016.
E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space nonparametric hidden Markov models and applications. Statistics and Computing, 26:61–71, 2016.
K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of POMDPs using spectral methods. In Proceedings of the 29th Conference on Learning Theory, volume 49. PMLR, 2016.
M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. Machine Learning, 49:209–232, 2002.
S. S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudík, and J. Langford. Provably efficient reinforcement learning with rich observations. In Advances in Neural Information Processing Systems, 32, 2019.
M. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In Advances in Neural Information Processing Systems, 28, 2015.
N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In Proceedings of the 33rd International Conference on Machine Learning, volume 48, pages 652–661. PMLR, 2016.
N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. Journal of Machine Learning Research, 21(167):1–63, 2020.
M. Janner, Q. Li, and S. Levine. Offline reinforcement learning as one big sequence modeling problem. In Advances in Neural Information Processing Systems, 34, 2021.
L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision Transformer: Reinforcement learning via sequence modeling. In Advances in Neural Information Processing Systems, 34, 2021.

Supplementary Material: Latent World-Model Interfaces

Anonymous authors
Paper under double-blind review

A Model and policy contract

This supplement uses the exact contract in the main paper. Hidden state and observation are $S_t, X_t \in \{-1, +1\}$ and action is $A_t \in \{0, 1\}$. The emission law and homogeneous transition law are

$$\mathbb{P}(X_t = x \mid S_t = s) = [1 + x(\mu + \rho s)]/2,$$
$$\mathbb{E}[S_{t+1} \mid S_t = s, A_t = a] = b_a + \beta_a s,$$

with $\rho \geq 0$, $|\mu| + \rho \leq 1$, and $|b_a| + |\beta_a| \leq 1$. Every episode begins from an independent uniform state. The policy sees only public observations and its own actions, and is total on every public history. The deployment value counts all H rewards, while the first reward arrives before the first effective action. The reset-triple learner observes three observations and takes two effective actions per training episode.

For a class $\mathcal{C}$, the risk is

$$R_{\mathcal{C}}(N, H) = \inf_{\mathcal{A}} \sup_{M \in \mathcal{C}} \mathbb{E}[V_M^* - V_M(\pi_{\mathcal{A}})].$$

The lower class fixes $\mu = 0$ and the upper class is the full legal class. No part of the proof assumes a hidden-state policy oracle.

B Upper transfer and fitted model

Under independent uniform training actions, the required interventional moments are

$$\mu, \quad P_a = \mathbb{E}[X_1 \mid do(a)], \quad Q_a = \mathbb{E}[X_0 X_1 \mid do(a)],$$
$$R_{ab} = \mathbb{E}[X_2 \mid do(a, b)], \quad S_{ab} = \mathbb{E}[X_0 X_2 \mid do(a, b)].$$

With $C_a = P_a - \mu$, $M_a = Q_a - \mu P_a$, $D_{ab} = R_{ab} - \mu$, and $N_{ab} = S_{ab} - \mu R_{ab}$,

$$C_a = \rho b_a, \quad M_a = \rho^2 \beta_a, \quad D_{ab} = C_b + \beta_b C_a, \quad N_{ab} = \beta_b M_a.$$

The logger uses bounded weights $1, 2, 4$. A union bound over the thirteen coordinates gives $\mathbb{E}E_N \leq \sqrt{32 \log(26)/N}$.

Set $m = \lceil 13\sqrt{N} \rceil$ and enumerate $r = j/m$, $0 \leq j \leq m$. For each r, clip the empirical offset to $[-1+r, 1-r]$, center all moments using that same offset, initialize the transition coefficients using M_a/r^2 when $K \leq r^2 + 10^{-15}$ (otherwise use the anchored two-action difference with anchor 0 on ties; the no-products ablation always uses this anchor), and project each (C_a, β_a) onto $|C_a|/r + |\beta_a| \leq 1$. The $r = 0$ branch sets transitions to zero and performs no division. Select the candidate minimizing the maximum residual over all thirteen raw moments.

For any legal comparison model at the selected grid point with raw discrepancy δ, centering gives errors at most δ for C, D and 2δ for M, N. The initializer has $K|\widehat{\beta} - \beta^*| \leq 3\delta$ with $K = \max\{r^2, |\widehat{C}_0|, |\widehat{C}_1|\}$. The weighted projection then gives

$$|\widehat{C} - C^*| \leq 4\delta, \quad K|\widehat{\beta} - \beta^*| \leq 6\delta,$$

and raw discrepancies bounded by $4, 10, 16, 28$ times δ for the corresponding moment families. Candidate selection yields

$$w \leq 377/m + 407E_N, \qquad \mathbb{E}w \leq C_{\text{fast}} N^{-1/2}, \quad C_{\text{fast}} = 29 + 407\sqrt{32 \log 26}.$$

The fit is a legal-model approximation and is not a hidden-coordinate identification theorem.

For the exact fitted optimizer, let $q = \max_a(|C_a|, |M_a|)$ and $s = q/\rho$ when both are nonzero. Signed insertion bounds along true prefixes and legal fitted suffixes give

$$\text{Regret} \le 2(H-1)s, \quad \text{Regret} \le 11H^2 w/s,$$
$$\text{Regret} \le K_H w, \qquad K_H = H(2H-1)(3H-1)/2.$$

If $\rho = 0$ or $q = 0$, all policies have the same value. Balancing the first two inequalities and using $K_H \le 3H^3$ gives the three branches in Theorem 1. The finite stored-pointer belief controller uses $M = \max\{1, H(H-1)m/2\}$ intervals (the $H = 1$ case keeps one inert interval) and adds at most $1/m$ fitted-model loss. Its arithmetic cost is $O(N+\sqrt{N}+H^3\sqrt{N})$ for genuine empirical inputs; this is not a claim of practical planning cost.

C Adaptive lower mechanisms

The lower constructions use complete adaptive training likelihoods. A causal training record factors into action probabilities and the interventional triple law. Summing the action weights over action words gives one, so the total-variation/KL bounds remain valid for within-episode and cross-episode adaptation. Deployment policies may use their observations to infer the hidden sign; the lower does not impose open-loop deployment.

For $N \le H^2$, an absorbing action-label pair leaves an action choice uncertain and incurs $H/8192$. For $H^2 \le N \le H^4$, a weak-flip pair sets the signal and action-label separation to the scale that makes the training laws overlap while the deployment value gap remains of order $H^{3/2}N^{-1/4}$. The direct middle construction is stated for $H \ge 64$ and $N \le H^4/4096$; clipping and monotonicity bridge the remaining range. For $N \ge H^4$, the clipped pair gives the $H^3N^{-1/2}$ branch. The constants in the main statement are retained exactly from the audited lower packet; no fixed transition floor is used.

D Proof of the window proposition

For $H \ge 8$, let $\alpha = 1/8192$, $\rho_H = \alpha/\sqrt{H}$, common drift $b_a = 1/H$, and let θ choose which action has $\beta = 1 - 2/H$ instead of $1 - 3/H$. The pair is disclosed to the learner. Use H^4 triples or $3H^3$ full-H episodes, for the common observation budget $B_H = 3H^4$.

At the triple interface, the clipped lower construction gives

$$R_3^{\text{pair}}(H) \ge \sqrt{H}/22020096.$$

This is a pair-level result and is not inferred from a larger class lower. Its proof bounds the adaptive training total variation by $2(\alpha + 5\alpha^2) < 1/4$, bounds the deployment proxy error by α, and retains a paired action-value gap of at least $H\rho_H/2688$.

At the full-window interface, take action zero throughout training. If $m_{0,t}$ and $m_{1,t}$ are the hidden means under the two labels, then

$$m_{0,t} = (1 - \beta_s^t)/2, \qquad m_{1,t} = (1 - \beta_f^t)/3,$$

and the difference satisfies

$$d_{t+1} = \beta_f d_t + m_{0,t}/H, \qquad d_0 = 0.$$

Keeping the last $\lfloor H/4 \rfloor$ terms in the convolution at $t = H - 1$ gives $d_{H-1} \ge 1/128$. The terminal observations from the $3H^3$ episodes are independent and have means separated by at least $\rho_H/128$. The midpoint test and Hoeffding's inequality therefore give

$$\mathbb{P}(\text{wrong label}) \le \exp\left(-\frac{3\alpha^2 H^2}{131072}\right).$$

A correct label permits an optimal public-history policy for the disclosed model; a wrong label costs at most $H\rho_H$. Hence

$$R_H^{\text{pair}}(H) \le \alpha\sqrt{H}\exp\left(-\frac{3\alpha^2 H^2}{131072}\right).$$

Reset count, effective actions, computation, and model knowledge are not matched by this proposition.

E Three-state boundary

A separate homogeneous three-state pair has all transition entries at least $1/12$ and all emission probabilities at least $1/8$, while its complete adaptive triple-training laws are identical for every finite N. At $H = 4$ its exact known-pair minimax loss is $67/6144$. This is an explicit boundary example only; it is not an arbitrary-rank rate or an arbitrary-horizon impossibility theorem.

F Reproduction and process map

The finite exact checks are supplied-input diagnostics. The retained packets include the fitted-controller replay, public-history policy enumeration, window recurrence checker, normal/optimized parity receipts, and explicit claim-scope audits. They do not constitute an external proof certificate or neural experiment. A separate process artifact records chronology and claim-to-evidence bindings; its provenance views are not latent-state or scientific-authority oracles.

F.1 Experimental questions and deterministic coverage

The implementation evidence is organized around nine falsifiable questions, rather than a single aggregate score. Table 1 records the input set, control, and observable that answer each question. This separation matters because a fitted-model hash can change without changing a deployment policy, while a policy can change without being better on every legal world. All nine protocols are closed-world exact checks: the records, moments, horizons, and evaluator are fixed before execution, and no episode is sampled during the study.

Reporting convention. Unless a caption states otherwise, a "receipt row" is one supplied-input execution, "mean" is an unweighted mean over the disclosed legal cells, "min/max" are finite-set extrema, and a "strict winner" is a unique minimum-regret arm. All implementation tables below are deterministic diagnostics with no standard error or population interpretation; sampled neural pilots are outside this evidence set and are not pooled into these tables.

The broad replay tests interface compliance and exposes a negative result: the complete moment family is not a universal winner on the supplied cells. The fixture then changes only one factor at a time to separate identifiability, model-to-value transfer, and planner capacity. The mesh ladder is a resource stress test, not a learning curve. Finally, the pair screen replaces two hand-picked controls by an exhaustive finite list, and the ladder study checks whether behavior is stable under declared sample, horizon, mesh, and mechanism changes. This ordering prevents a favorable fixture from being mistaken for an empirical ranking.

The implementation layer now includes a separate exhaustive float64 screen. It enumerates the same fifteen fixed legal worlds and all thirty-two public $H = 3$ triples. Each triple is replayed separately as a supplied $N = 1$ record, then weighted by its exact world-conditional law for an independent exact $H = 8$ deployment evaluation. The three arms are the complete thirteen-moment learner, an arm without endpoint products, and an arm using only the first-order moments. The complete arm is strictly better than both ablations in $2/15$ worlds, strictly worse than at least one ablation in $6/15$, and tied across all arms in $7/15$. This mixed result is useful as an implementation and falsification check, but it is not an IID estimate, confidence interval, population result, neural baseline, compute-matched comparison, or minimax claim; the separate admitted IID campaign remains blocked by the experiment-integrity gate.

For scale, the unweighted mean of the exact world-conditional regrets across the fifteen fixed cells is 0.5156427701 for full13 and 0.5096643663 for each ablation; the fixed-cell ranges are $[0, 2.625]$ and $[0, 2.5625]$, respectively. The two ablations are cellwise identical on this screen. These numbers are included to expose the mixed outcome, not to define a population mean or a statistical estimate.

Table 1: Deterministic experiment protocols used by the implementation study. For every row, the input records, legal-world set, horizon, evaluator, and aggregation unit are frozen before execution. Each row is a supplied-input diagnostic, not a statistical estimate or an admitted population result.

Protocol	Supplied input	Control or ablation	Primary observable (aggregation)
Broad replay	15 legal worlds, 32 public triples per world, $H = 8$	full thirteen moments vs. endpoint-product and product-free masks	exact true-world regret; wins/losses/ties
Mechanism fixture	one legal world, exact population moments, $H = 8$	true-model planner and three fitting masks	fitted parameters, surrogate value, true regret
Mesh ladder	same fixture moments; seven declared N values from 1 to 2048	fixed moments; only mesh and pointer budget vary	stored nodes, pointer bytes, exact-evaluator refusal
Pair screen	all 528 unordered pairs with replacement from the 32 triples	same fixture and public-history oracle	model-hash equality, regret distribution, strict-best counts
Ladder transfer	fixed alternating records 0/5; $N = 1, \ldots, 256$, $H = 1, \ldots, 8$, six meshes, five worlds	no sampling; vary one declared axis at a time	exact regret, fitted contrast, resource count, transfer gap
Parameter grid	45 labelled cells (33 unique mechanisms), fixed fit and exact-moment fit	four one-dimensional axes plus a 5×5 Cartesian slice	270 exact deployment rows; transfer gap and strict-winner counts
Observable exact grid	15 worlds $\times$ 32 ordered $H = 3$ records, $N = 1$	latent, two suffix arms, and three observable skeletons	2,880 exact fits; regret counts versus latent (no compute ranking)
Fine mechanism grid	132 labelled cells (131 unique mechanisms), fixed $N = 32$, $H = 8$	11×11 strength/persistence slice plus 11 transition-bias cells	792 exact deployment rows; dense failure-boundary screen
Horizon transport grid	45 labelled cells $\times$ $H = 1, \ldots, 8$, fixed $N = 32$	three moment masks; world and horizon crossed	1,080 exact rows; horizon amplification and strict-winner counts

Table 2: Finite exhaustive learner screen. Each entry is the exact world-conditional deployment regret, averaged equally over the 15 fixed legal worlds after averaging the 32 ordered supplied records within each world (480 executions per arm). The relative-outcome column compares the three arms cellwise; no sampled-world interpretation is intended. Deterministic supplied-input diagnostic.

Arm	Fixed-cell mean	Min	Max	Relative outcome
full13	0.5156427701	0	2.6250	wins 2; losses 6; ties 7
no_endpoint_products	0.5096643663	0	2.5625	same as other ablation
no_products	0.5096643663	0	2.5625	same as other ablation

F.2 Observable exact-grid screen at $N = 1$, $H = 3$

An independent exact-grid packet evaluates the latent and observable alternatives on the short-reset contract. It contains 15 predeclared legal worlds (five emission settings crossed with three transition settings) and all 32 ordered public records $(x_0, a_0, x_1, a_1, x_2)$ in each world. Every learner receives one supplied record only; deployment is then evaluated exactly at $H = 3$. The six named arms are latent, suffix_k1, suffix_k2, and three observable skeletons: observable_time_aware, observable_full_history, and observable_recurrent_2bit. Thus the receipt contains $15 \times 32 \times 6 = 2{,}880$ fit/evaluation rows before world-conditional weighting. An independent forward-occupancy calculation agrees with the exact evaluator, and normal/optimized receipts are byte-identical.

Table 3 reports the mean, minimum, and maximum regret across the 15 disclosed world-conditional values. The last column is a diagnostic comparison with the latent arm; it is not a compute-matched ranking, because the packet explicitly denies common runtime or memory accounting and the observable skeletons are fixed finite alternatives rather than a full observable-policy class.

Table 3: Observable exact-grid screen at $N = 1$, training and deployment $H = 3$. Each world value exactly averages all 32 ordered public records. The last column counts cells lower/equal/higher than the latent arm.

Arm	Mean regret	Min	Max	Lower/equal/higher
latent	0.1208333333	0	0.7500	baseline
suffix_k1	0.1239583333	0	0.7500	0/12/3
suffix_k2	0.1246673584	0	0.7500	2/9/4
observable_time_aware	0.1273518880	0	0.7500	1/10/4
observable_full_history	0.1268936157	0	0.7500	1/9/5
observable_recurrent_2bit	0.1611501058	0	1.0625	0/7/8

This is a deterministic short-reset comparison, not evidence that a latent representation is necessary or that any arm generalizes beyond the disclosed grid. The packet's explicit resource-comparability denial is retained in its public verification receipt.

F.3 Mechanism fixture for the product moments

The preceding screen is intentionally broad but uses one supplied triple at a time. To test whether its two ablations are merely duplicate code paths, we also use one exact legal model as a mechanism fixture: $\mu = 0$, $\rho = 1/2$, $b = (0, 0)$, and $\beta = (1/2, -1/2)$. Feeding its exact population moments as a supplied $N = 4$ vector gives three different fits. The full interface recovers $\rho = 1/2$ and has exact $H = 8$ regret zero; removing the endpoint products gives $\rho = 5/13$ and regret 7305/131072; removing all product moments gives $\rho = 0$ and regret 7/16. A true-model stored-pointer planner on the same grid reaches the public-history oracle, so this fixture isolates fitted-model transfer rather than a planner incapacity. Two supplied $N = 2$ controls are retained: records 0 and 5 separate all three masks, while records 0 and 6 leave the full and endpoint-ablated fits equal but change the no-products fit. These are exact diagnostics, not sampled risk estimates or population claims.

The same fixture also exposes the resource boundary of the stored-pointer planner. We hold the exact population moments fixed and vary only the declared mesh budget; the resulting ladder is therefore a planning diagnostic rather than an estimator learning curve. The retained exact evaluator refuses value evaluation once its 50,000-node safety limit is exceeded.

The same fixed-moment fixture gives a small horizon transfer check before the evaluator reaches its supported horizon limit. With the declared sample size held at $N = 4$ (so the mesh and fitted model are unchanged), the complete mask has zero regret at all three horizons, while the two ablations accumulate different value gaps. This is a deterministic sensitivity check for the model-to-value transfer, not evidence for the asymptotic H^6 join.

Table 4: Horizon sensitivity on the exact mechanism fixture with fixed moments and fixed declared $N = 4$. The exact evaluator supports $H \leq 8$ here.

Horizon H	Stored nodes	full13 regret	no_endpoint regret	no_products regret
4	471	0	3/256	3/16
6	1,955	0	123/4096	5/16
8	5,103	0	7305/131072	7/16

Two supplied-record controls check that this separation is not an artifact of the population-moment fixture. Pair A (canonical records $0, 5$) makes all three fitted model hashes differ, whereas Pair B (records $0, 6$) keeps the complete and endpoint-ablated fits identical while changing the product-free fit. Their exact true-world values are shown in Table 6; the controls use the same $H = 8$ public-history evaluator as the main fixture.

To replace the two hand-picked N=2 controls by a finite mechanism screen, we enumerate every unordered pair with replacement from the 32 canonical public triples. This gives 528 supplied pairs in the same legal fixture and keeps the exact H=8 public-history oracle fixed at 71/16. The three fitted model hashes are all distinct on 36 pairs, while all three exact deployment regrets are equal on 459 pairs. Pairwise fitted-model equality counts are

Table 5: Fixed-moment mesh/planner ladder on the mechanism fixture. "NE" means not evaluated by the exact evaluator after its 50,000-node limit; model fits and resource counts are still retained.

Declared N	Mesh m	Stored nodes	full13	no_endpoint	no_products
1	13	2,555	0	7305/131072	7/16
4	26	5,103	0	7305/131072	7/16
16	52	10,199	0	21381/262144	7/16
64	104	20,391	0	11217/131072	7/16
256	208	40,775	0	11985/131072	7/16
1024	416	81,543	NE	NE	NE
2048	589	115,451	NE	NE	NE

Table 6: Supplied $N = 2$ negative and positive controls on the mechanism fixture. Values are exact deployment returns, not fitted-model surrogates.

Pair	Hash relation	full13 value	no_endpoint value	no_products value
A: $(0, 5)$	all three differ	4	4	4173/1024
B: $(0, 6)$	full = endpoint; product-free differs	4	4	71/16

406/528 for full versus no-endpoint, 356/528 for full versus no-products, and 418/528 for no-endpoint versus no-products. The finite-set deployment summaries are shown below. These are descriptive summaries of a fixed pair list, not a population risk or confidence interval.

The screen reinforces the mechanism interpretation without turning it into an empirical ranking: most record pairs have the same finite controller value, but a small, explicitly retained subset exposes transfer differences. Pair indices $(0, 5)$ and $(0, 6)$ recover the earlier positive and negative controls. The receipt and source-aware verifier are released separately; all inputs are supplied records and the admitted IID campaign remains outside this result.

F.4 Deterministic ladder study: fixed records, multiple axes

The preceding mechanism fixture uses exact population moments. To check that its conclusions are not tied to that one input representation, we ran a separate, fully disclosed ladder on the frozen public-record grid. The schedule alternates canonical records 0 and 5, so increasing N repeats a supplied sequence rather than drawing new episodes. The learner, mesh, and exact evaluator are unchanged; only one declared axis is varied at a time. The resulting receipt contains exactly $(9 + 8 + 6 + 5) \times 3 = 84$ rows: four one-axis slices (nine N values, eight horizons, six mesh values, and five legal transfer worlds), each evaluated for all three arms. These slices are not a $9 \times 8 \times 6 \times 5$ Cartesian product.

Table 8 shows a stable deployment regret for the two larger moment masks under this contrast schedule, while the product-free arm changes after the first pair. This is a controller behavior on a fixed schedule, not a learning curve. The exact regrets are reported alongside the fitted contrast so that a parameter movement without a policy movement is visible.

The horizon ladder uses the same supplied $N = 32$ prefix; its complete-mask, endpoint-ablation, and product-free values are listed in Table 9. The non-monotone product-free entries are a policy effect on this one disclosed mechanism, not a horizon law. Holding the same fitted moments and varying the mesh over $26, 39, 52, 74, 104, 148$ moves ρ by up to 0.0192 but leaves each arm's $H = 8$ regret unchanged, exposing the distinction between parameter and controller sensitivity.

The two axis-wise slices are tabulated explicitly below. They make the negative result visible: changing the declared budget or mesh can move a fitted parameter without moving the exact deployment policy, while the product-free controller can change at a few horizons under the same fitted moments. No row should be read as an estimate over a task distribution.

The biased-transition row is a useful failure-boundary example: a fit that is acceptable on the training fixture is not a universally safe controller across even five disclosed legal mechanisms. All ladder rows passed exact evaluator identity checks; the normal and optimized receipts are byte-identical. The study remains a deterministic supplied-input diagnostic and does not

Table 7: Exhaustive supplied $N = 2$ pair mechanism screen in the same fixture. The denominator is 528 unordered pairs with replacement; mean and population SD use equal weight 1/528, and max is the finite-set maximum. This is a deterministic supplied-input diagnostic.

Arm	Mean regret	Population SD (528)	Max regret	Strict best pairs
full13	0.4406669031	0.1127915215	0.875	11
no_endpoint	0.4407441183	0.1520005469	0.875	10
no_products	0.4437335513	0.1565993635	0.875	20

Table 8: Fixed-record N ladder at $H = 8$. Rows alternate supplied records 0 and 5; no record is sampled or shuffled. Entries are fitted ρ followed by exact true-world regret.

N	full13 ρ/regret	no_endpoint ρ/regret	no_products ρ/regret
1	0.0000/0.4375	0.0000/0.4375	0.0000/0.4375
2	0.7368/0.4375	0.6842/0.4375	1.0000/0.3623
4	0.7308/0.4375	0.6538/0.4375	1.0000/0.3623
8	0.7568/0.4375	0.6757/0.4375	1.0000/0.3623
16	0.7500/0.4375	0.6731/0.4375	1.0000/0.3623
32	0.7432/0.4375	0.6757/0.4375	1.0000/0.3623
64	0.7500/0.4375	0.6731/0.4375	1.0000/0.3623
128	0.7500/0.4375	0.6689/0.4375	1.0000/0.3623
256	0.7500/0.4375	0.6683/0.4375	1.0000/0.3623

Table 9: Horizon ladder on the fixed supplied $N = 32$ prefix. Entries are exact true-world regrets for the disclosed fixture; the non-monotone product-free column is a controller effect, not a horizon-rate claim.

H	full13	no_endpoint	no_products
1	0	0	0
2	0.0625	0.0625	0
3	0.1250	0.1250	0.1250
4	0.1875	0.1875	0.109375
5	0.2500	0.2500	0.2500
6	0.3125	0.3125	0.23828125
7	0.3750	0.3750	0.3750
8	0.4375	0.4375	0.3623046875

Table 10: Mesh sensitivity at fixed supplied $N = 32$ and $H = 8$. Each entry is fitted ρ/exact true-world regret (the slash means fitted-ρ/regret). The unchanged regret columns show that parameter movement need not induce a controller change on this fixture.

Mesh	full13 ρ/regret	no_endpoint ρ/regret	no_products ρ/regret
26	0.7308/0.4375	0.6538/0.4375	1.0000/0.3623
39	0.7436/0.4375	0.6667/0.4375	1.0000/0.3623
52	0.7500/0.4375	0.6731/0.4375	1.0000/0.3623
74	0.7432/0.4375	0.6757/0.4375	1.0000/0.3623
104	0.7500/0.4375	0.6731/0.4375	1.0000/0.3623
148	0.7500/0.4375	0.6689/0.4375	1.0000/0.3623

Table 11: Cross-world transfer of the fixed $N = 32$, $H = 8$ fit. Each cell is exact true-world regret; the five worlds are disclosed legal mechanisms, not a sampled distribution.

World fixture	full13	no_endpoint	no_products
fixture	0.4375	0.4375	0.3623
emission offset	0.4898	0.4898	0.4146
weak emission	0.5283	0.5283	0.4241
fully observed	0	0	0
biased transition	1.0566	1.0566	0.8105

establish IID performance, confidence intervals, neural representation quality, population generalization, or minimax performance.

F.5 Systematic parameter-grid transfer

The ladder varies one declared resource axis at a time. We add a separate mechanism stress test to distinguish that ladder from transport across the world family. The grid has 45 labelled cells: four five-point one-dimensional axes (emission offset, emission strength, transition persistence, and transition bias) and a 5×5 Cartesian slice over emission strength and persistence. Axis endpoints overlap, so these cells represent 33 unique legal parameter tuples, not 45 unique worlds. The fixed supplied-fit branch uses the alternating public records 0/5 at $N = 32$; the matched branch fits each cell from its exact thirteen population moments. Together they produce $45 \times 3 \times 2 = 270$ exact deployment rows, in addition to eight supplied-prefix identifiability rows. No record is sampled, shuffled, or generated in either branch.

Table 12 reports the fixed-fit transfer. Entries are finite-grid mean/max exact regrets, and the final column counts cells with a unique best arm. The full and endpoint-product arms often have the same controller on this supplied fit, whereas the product-free arm exposes a different transfer boundary.

Table 12: Fixed supplied-fit transfer on 45 labelled cells (33 unique mechanisms). Each arm entry is a family-row mean/max exact regret, with equal weight 1/Cells within that row (5 or 25); strict winners count unique minimum-regret cells. No aggregate over all 45 cells is implied. Deterministic supplied-input diagnostic.

Family	Cells	full13 mean/max	no_endpoint mean/max	no_products mean/max	Strict winner
Emission offset	5	0.4573/0.5556	0.4573/0.5556	0.3821/0.4804	product-free 5/5
Emission strength	5	0.6563/1.7500	0.6563/1.7500	0.5435/1.4492	product-free 4/5; 1 tie
Transition persistence	5	0.2625/0.4375	0.2625/0.4375	0.2494/0.4619	product-free 2/5; 3 ties
Transition bias	5	0.7394/2.1133	0.7394/2.1133	0.6642/1.6963	product-free 3/5; 2 ties
$\rho\times$ persistence	25	0.3938/1.7500	0.3938/1.7500	0.3741/1.8477	product-free 8/25; 17 ties

Table 13: Exact-moment matched transfer on the same labelled grid. Each arm entry is a family-row mean/max exact regret, with equal weight 1/Cells within that row (5 or 25). The strict-winner column gives the cells in which the full interface is the unique minimum-regret arm. No aggregate over all 45 cells is implied. Deterministic supplied-input diagnostic.

Family	Cells	full13 mean/max	no_endpoint mean/max	no_products mean/max	Full13 strict winner
Emission offset	5	0/0	0.0860/0.1114	0.4573/0.5556	5/5
Emission strength	5	0/0	0.1757/0.5453	0.6563/1.7500	4/5; 1 tie
Transition persistence	5	0/0	0.0704/0.1248	0.2625/0.4375	4/5; 1 tie
Transition bias	5	0/0	0.0163/0.0816	0.1054/0.4375	1/5; 4 ties
$\rho\times$ persistence	25	0/0	0.1228/0.5453	0.3938/1.7500	16/25; 9 ties

The matched branch gives a clean identifiability check: the complete moment interface reaches zero exact regret on every disclosed cell, while the two ablations retain transfer loss. The fixed branch shows the complementary failure mode: transporting one fitted controller across legal mechanisms can produce regret as high as 2.1133 on the transition-bias slice. These are finite disclosed-mechanism diagnostics, not estimates of a task population or evidence of universal dominance. Their purpose is to separate fit recovery from model-to-world transfer before interpreting any world-model claim.

F.6 Dense mechanism boundary audit

The coarse 45-cell grid can miss narrow failure bands. We therefore repeat the same fixed supplied fit on an 11-by-11 emission-strength/persistence slice and an 11-point transition-bias axis. The two axes share one zero-bias endpoint, giving 132 labelled cells and 131 unique legal mechanisms, with 792 exact deployment rows across the three masks and an exact-moment matched branch.

Table 14: Dense mechanism grid. Entries are family-row mean/max exact regret, with equal weight 1/Cells within each row (121 or 11); the matched branch is included to separate fitting from transport. The 132 labelled cells are a disclosed finite grid, not a sampled task population.

Family	Cells	full13 mean/max	no_endpoint mean/max	no_products mean/max
Strength × persistence (fixed fit)	121	0.3341/1.7500	0.3341/1.7500	0.3191/1.8477
Transition bias (fixed fit)	11	0.6854/2.1133	0.6854/2.1133	0.6102/1.6963
Strength × persistence (matched moments)	121	0.000006/0.0003	0.0987/0.5453	0.3341/1.7500

The product-free mask is the strict winner on 50 of the 121 fixed-fit strength/persistence cells and 7 of the 11 transition-bias cells, whereas the matched branch makes the complete interface nearly exact. The dense screen therefore sharpens the transport boundary without changing the scientific interpretation: these are disclosed labelled mechanisms, not draws from a world distribution.

F.7 Horizon-by-world transport grid

The earlier ladder varied horizon on one fixture, while the coarse world grid fixed $H = 8$. The horizon-transport packet crosses these axes: the supplied $N = 32$ fit is held fixed, and each of 45 labelled mechanisms is evaluated at $H = 1, \ldots, 8$ for all three masks, producing 1,080 exact rows. The $H = 1$ row is an explicit initial-reward baseline with no action layer.

Table 15: Horizon transport on the 45-cell world grid. Entries are mean/max exact regret for the complete mask, with equal weight over the 45 cells at each horizon; the final column counts strict product-free winners among the three masks. Deterministic supplied-input diagnostic.

H	Mean regret	Max regret	Product-free strict winners
1	0.0000	0.0000	0/45
2	0.0625	0.2500	22/45
3	0.1257	0.5000	2/45
4	0.1909	0.8125	22/45
5	0.2562	1.1250	2/45
6	0.3220	1.4531	22/45
7	0.3879	1.7813	2/45
8	0.4538	2.1133	22/45

The monotone increase in the complete-mask transport summary is consistent with horizon amplification on this fixed supplied fit, while the alternating strict-winner pattern is a finite-grid control result rather than a horizon rate. The receipt records controller resources and exact evaluator identity; no episode is sampled and no population or minimax claim is promoted.

F.8 Policy-class decomposition and planner controls

The world-grid transfer rows still leave one confound: a poor fitted controller can lose value because it misfits the mechanism, because its finite belief mesh is insufficient, or because a history-dependent policy is needed. We therefore freeze the same 45 labelled cells, the alternating supplied record-0/record-5 fit at $N = 32$, and $H = 8$, and evaluate four independent controls with the exact evaluator. fixed0 and fixed1 repeat one action; open_loop is the best of all 2^7 observation-independent action sequences; true_model_grid plans from the exact mechanism on the full13 pointer mesh; and the public-history oracle is the evaluator reference.

The fixed-action-1 control is exactly tied with fixed-action-0 on every reported family because the disclosed grid is action-symmetric. The true-model planner is exact on all 45 cells at the same finite pointer budget, so the planner discretization is not the dominant error in this packet. In contrast, the supplied-fit full13 controller reaches mean/max regret

Table 16: Policy-class decomposition on the same labelled world grid. Each entry is a family-row mean/max exact regret (equal weight 1/Cells; the slash means mean/max). Four controls therefore give $4 \times 45 = 180$ exact evaluations, all passing the evaluator identity check. This is a deterministic supplied-input diagnostic.

Grid family	fixed action 0	fixed action 1	best open-loop	true-model grid planner
Emission offset (5)	0.4573/0.5556	0.4573/0.5556	0.4573/0.5556	0/0
Emission strength (5)	0.6563/1.7500	0.6563/1.7500	0.6563/1.7500	0/0
Transition persistence (5)	0.2625/0.4375	0.2625/0.4375	0.2625/0.4375	0/0
Transition bias (5)	0.7394/2.1133	0.7394/2.1133	0.1054/0.4375	0/0
$\rho\times$ persistence (25)	0.3938/1.7500	0.3938/1.7500	0.3938/1.7500	0/0

.7394/2.1133 on the transition-bias slice, whereas the best open-loop policy has .1054/.4375. This control therefore attributes the large transfer gap to model-to-world transport rather than to an intrinsic requirement for history conditioning. The result is a deterministic decomposition, not a neural baseline, compute-matched benchmark, IID estimate, or population ranking; its receipt is released in 02_experiments/20260924_policy_class_ablation_v1/.

F.9 Negative results and falsification checks

The implementation campaign was designed to expose cases in which a stronger moment interface does not improve the finite controller. Table 17 collects the retained negative outcomes and the claim each one rules out. This is the appropriate interpretation of the experiments: they test whether the constructive interface is sufficient for the stated mechanism, not whether it wins a benchmark suite.

Table 17: Negative results retained as falsification checks. Every row uses a fixed supplied set or a disclosed exact mechanism. The 459 tied-regret and 36 hash-difference counts are separate, non-partitioning diagnostics.

Check	Observed outcome	Claim ruled out
15-world broad replay	full13 wins 2, loses 6, ties 7	More moments do not guarantee a lower regret on every finite cell.
32 one-record cells	Endpoint and product-free masks coincide on all 32 controllers	A one-triple screen cannot establish endpoint-product necessity.
528 supplied pairs	All three regrets tie on 459 pairs; hashes differ on 36	Parameter identifiability and control relevance are not equivalent.
Mesh ladder	ρ moves by up to 0.0192, while $H = 8$ regret is unchanged	A fitted-parameter change is not itself a deployment improvement.
Horizon ladder	Product-free regret is non-monotone across $H = 1, \ldots, 8$	A single finite mechanism does not prove a horizon scaling law.
Five-world transfer	Biased-transition row favors product-free (0.8105 vs. 1.0566)	A fit that works on one fixture is not universally safe across legal mechanisms.

These failures are not repaired by selecting the most favorable row: each is kept in the public receipt and is referenced by the process claim-to-evidence map. In particular, the complete moment interface is a constructive upper bound in the theorem, not an empirical dominance statement.

F.10 Reproducibility scope

The evidence has three independent integrity layers. First, the exact evaluator checks the public-history oracle and verifies controller identity on every reported row. Second, normal and optimized executions produce byte-identical JSON receipts; the mechanism, pair-screen, ladder, and world-grid test suites report 10, 3, 7, and 4 passing tests, respectively. Third, every table row is generated from a disclosed record list, exact moment vector, or explicit legal world and is included in the public process bundle. The evaluator is bounded at $H \leq 8$ and 50,000 stored controller nodes; rows beyond that limit are marked not evaluated rather than silently truncated.

This scope supports reproduction of the deterministic diagnostics and the theorem's stated finite mechanisms. It does not support confidence intervals, population averages, neural representation comparisons, or claims about general-purpose world-model performance.

G Monte Carlo calibration of the reset-triple learner

The deterministic screens in the main paper establish algebraic and exact pointer invariants, but they intentionally do not estimate behavior on independent training streams. We therefore ran a separate local Monte Carlo calibration with the released float64 fitter and the exact public-history evaluator. The runner hardcodes four fixed legal worlds, $N \in \{1, 4, 16, 64, 256\}$, 32 replicate streams per cell, and $H = 8$; the public protocol is a retrospective execution record rather than a preregistration. The arms share each replicate's training triples, so the arm comparisons are paired. The evaluator's public-history oracle is computed once per world; all fitted controllers are scored with an independent exact Fraction occupancy pass.

Table 18: Per-world Monte Carlo calibration. Entries are mean exact public-history regret ± sample SD over 32 paired streams. The worlds are fixed mechanisms, not an IID population.

World	N	full13	no_endpoint	no_products
interior	1	0.424±0.077	0.437±0.111	0.437±0.111
	4	0.354±0.226	0.353±0.197	0.415±0.210
	16	0.358±0.195	0.361±0.193	0.380±0.239
	64	0.345±0.166	0.336±0.185	0.435±0.230
	256	0.265±0.216	0.268±0.194	0.371±0.227
weak emission	1	0.058±0.082	0.055±0.079	0.055±0.079
	4	0.066±0.085	0.081±0.087	0.060±0.083
	16	0.067±0.082	0.052±0.076	0.071±0.079
	64	0.093±0.086	0.082±0.084	0.092±0.075
	256	0.067±0.079	0.058±0.073	0.073±0.075
weak transition	1	0.197±0.480	0.161±0.401	0.161±0.401
	4	0.226±0.421	0.273±0.433	0.248±0.408
	16	0.229±0.386	0.264±0.411	0.295±0.393
	64	0.024±0.118	0.024±0.118	0.156±0.299
	256	0.036±0.153	0.039±0.153	0.165±0.299
near boundary	1	2.365±0.747	2.316±0.708	2.316±0.708
	4	1.401±1.107	1.480±1.164	2.070±1.167
	16	0.449±0.728	0.784±1.029	1.733±1.365
	64	0.146±0.101	0.152±0.098	1.530±1.442
	256	0.084±0.098	0.086±0.100	1.290±1.160

The near-boundary and weak-transition mechanisms show the clearest reduction with N, and removing both product families is particularly harmful near the boundary. In the weak-emission mechanism every arm has a small value gap, which is expected because the observation channel carries little actionable signal. No arm wins on every fixed mechanism or budget. The released paired_differences.csv gives replicate-level deltas with descriptive 95% intervals; it is the appropriate file for checking the paired comparison rather than treating the 1,920 rows as independent observations.

The receipt contains 1,920 arm-level rows, 640 paired training streams, source hashes, model/controller hashes, exact regrets, and an independent occupancy check. Its status is EXPLORATORY_LOCAL_COMPLETED_NO_ADMISSION: the stdlib random source has no external signer or native D10 admission. This calibration therefore supplies corresponding implementation evidence for the learner/planner, but it does not validate the minimax theorem, population generalization, latent superiority, or a neural world-model benchmark.

H Exploratory visual world-model audit

The theorem studies a binary controlled latent process, so it is not by itself an empirical neural world-model benchmark. We nevertheless ran a separate pixel-level audit to test

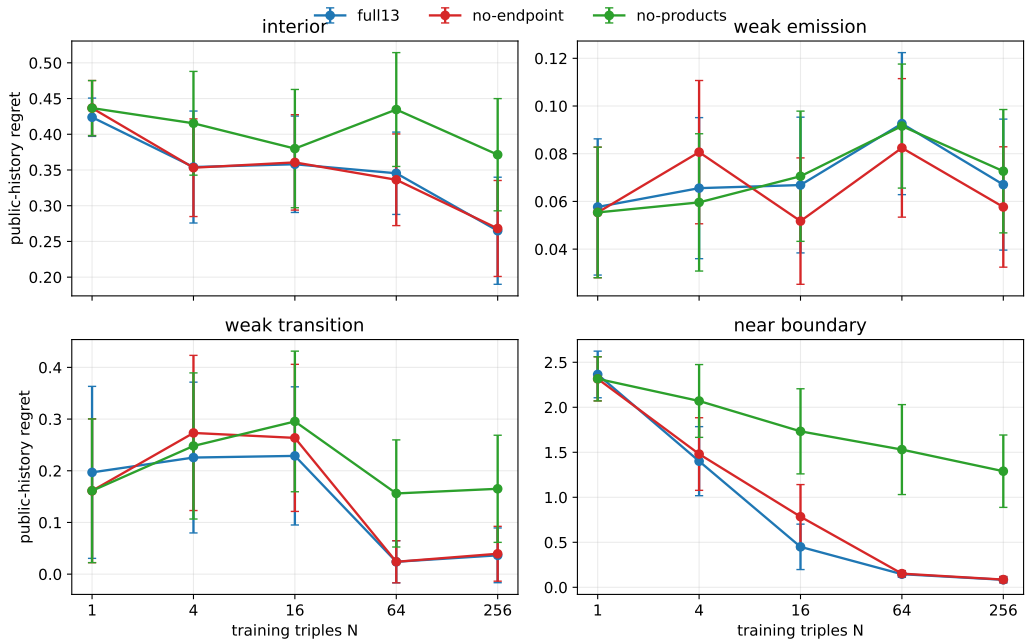

Figure 1: Monte Carlo calibration curves for the exact reset-triple learner. Error bars are descriptive normal-approximation intervals over 32 replicate streams. The fixed $H = 8$, $N \leq 256$ grid is a finite calibration and does not fit or validate an asymptotic rate.

whether the surrounding research workflow can train an action-conditioned model, expose its prediction failure, and evaluate the learned model in a fresh closed loop. The receipt is explicitly EXPLORATORY_NOT_ADMITTED: no authenticated D10 executor or external signer was available, and none of these rows is used to support the theorem.

The environment renders a 16×16 grayscale frame containing a soft agent blob and a target marker. The hidden state is position and velocity; action $a \in \{-1, 0, +1\}$ changes the inertial transition, and reward is normalized target tracking. Each model receives two consecutive frames, encodes them with a convolutional network, predicts an action-conditioned latent next state, decodes the next frame, and predicts immediate reward. The compact arm has an 8-dimensional latent; the wider local comparator is explicitly named wide16 under the declared CPU budget. Planning enumerates action sequences at depths 1, 3, and 5. A model-free arm is included only as a diagnostic because its labels come from a one-step expert, while random and matched-start greedy reference controls provide negative and reference controls.

The expanded screen contains three fixed world parameter settings, training sizes $N \in \{32, 128, 512\}$, three seeds, twenty training epochs, and sixteen held-out episodes per cell. Episodes are independently sampled within those fixed settings; they are not draws from a world population. It emits 243 closed-loop rows, 216 open-loop rollout rows, and 54 unique training histories, with two representative aliases retained for traceability. The evaluator compares every arm to a matched-start one-step greedy reference with a velocity penalty. Its normalized shortfall is exactly zero by construction for the reference arm; it is not regret to a proven optimal controller. At fixed planning depth 3, the pooled shortfalls for $(N = 32, 128, 512)$ are $(0.0914, 0.3805, 0.2429)$ for latent8, $(0.2471, 0.2553, 0.2731)$ for wide16, $(0.0111, 0.0067, 0.0068)$ for privileged imitation, and $(0.4030, 0.3800, 0.3843)$ for random. The latent curve is therefore non-monotone in this budget. The world-separated latent8 depth-1/3/5 means are $(0.1583, 0.1599, 0.1599)$ for near_target, $(0.2941, 0.3094, 0.2594)$ for far_target, and $(0.1938, 0.2455, 0.2455)$ for noisy_far_target. All intervals in the released summaries are descriptive t intervals over fixed cells, not population confidence intervals. These results expose the current failure boundary: the expert-label arm is not compute-

matched, the reference is heuristic, and neither more data nor a wider latent is consistently better.

Table 19: Expanded visual screen: clipped relative shortfall at planning depth 3, pooled over the three fixed worlds and three seeds. Parentheses show sample standard deviation over the nine fixed world/seed cells.

N	latent8	wide16	privileged imitation	random
32	0.0914 (0.1514)	0.2471 (0.3368)	0.0111 (0.0118)	0.4030 (0.0826)
128	0.3805 (0.2576)	0.2553 (0.2356)	0.0067 (0.0054)	0.3800 (0.0691)
512	0.2429 (0.2437)	0.2731 (0.2506)	0.0068 (0.0060)	0.3843 (0.0731)

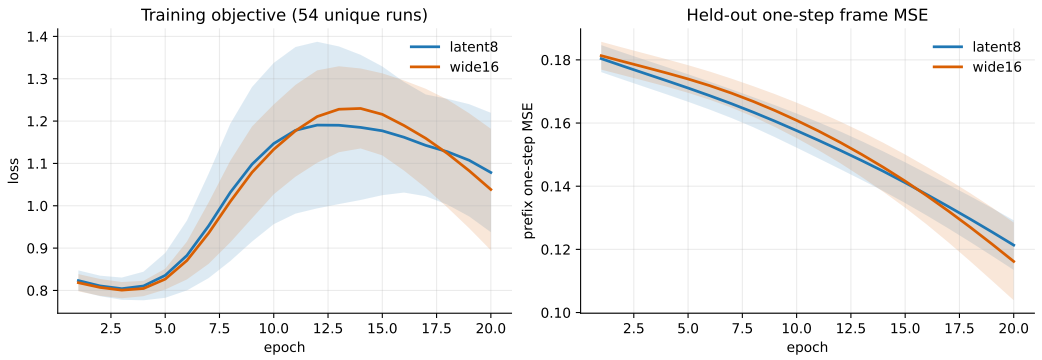

Figure 2: Actual visual-model training and held-out one-step frame prediction curves from all 54 unique receipt histories (mean with sample standard-deviation band). The curves use the twenty-epoch CPU budget and are retained as a diagnostic rather than selectively summarized; the wider arm is not assumed to be better merely because it has a larger latent.

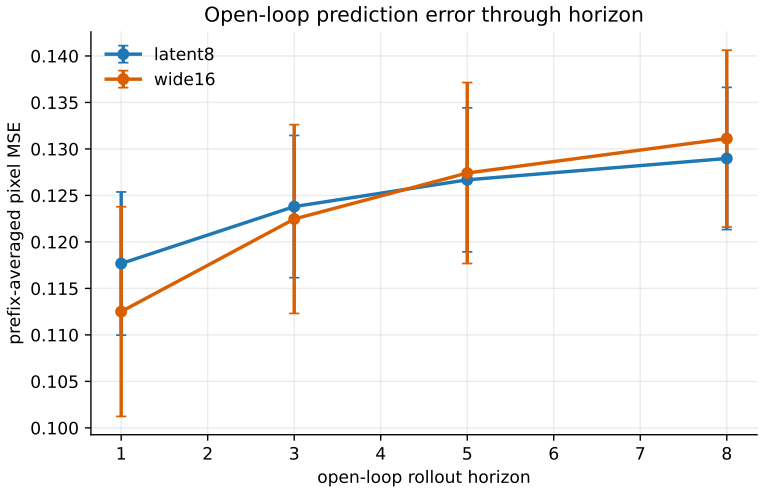

Figure 3: Open-loop prefix-averaged pixel prediction error through the indicated horizon, averaged over the disclosed worlds, data sizes, and seeds. This is the mean of one-step errors accumulated along each prefix, not an endpoint MSE or a claim of compounding error; bars are sample standard deviations across receipt rows.

All raw values, CSV exports, model-card metadata, environment fingerprint, tests, and plotting source are in 02_experiments/20260925_visual_neural_world_model_v2/. The verification receipt records 243 control rows, 216 rollout rows, 54 unique histories, exact-zero reference-arm shortfall, and increasing prefix-mean rollout error at horizons $(1, 3, 5, 8)$. A

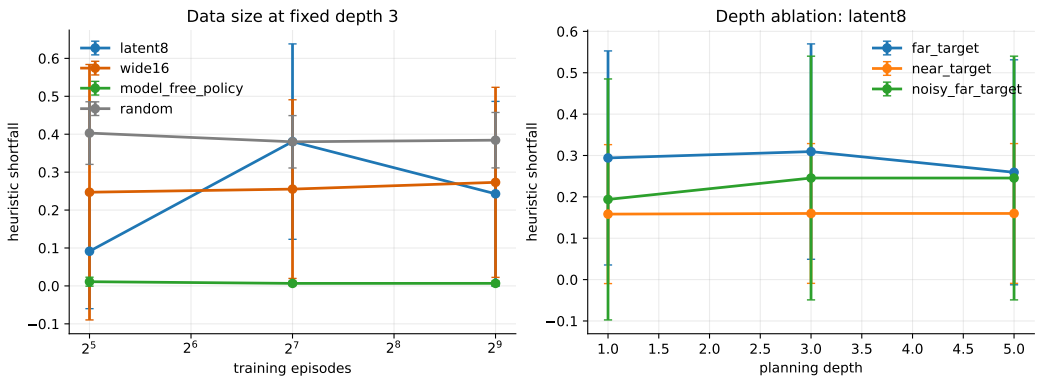

Figure 4: Fresh closed-loop clipped relative shortfall. The left panel fixes planning depth 3 and varies training episodes; the right panel varies planning depth for each fixed world setting. Error bars are sample standard deviations, not confidence intervals. The near-zero imitation diagnostic uses privileged one-step expert labels and is not evidence for a fair neural comparison.

future admitted run must replace the CPU-budgeted comparator with a declared capacity, remove expert labels from the primary baseline, add more seeds and training points, and attach a trusted executor receipt before any main-paper claim is upgraded.

I Paired held-out-world visual campaign

The v3 follow-up was frozen after the earlier exploratory screen and before looking at its control curves. Its training data are nested random-action prefixes with $N \in \{64, 256, 1024\}$, three seeds, and 19 transitions per episode. A model is trained on exactly one of the three training worlds and is then evaluated on that same world and on three fixed held-out parameter worlds: mid_target, high_inertia, and noisy_mid_target. Every arm within a cell reuses the same 24 initial states and observation-noise tape. The run uses 200 fixed Adam updates, minibatches of 128 transitions, a 12-step control horizon, and planning depths 1, 3, and 5. Checkpoints are saved for every fit and reloaded in a separate replay check.

The primary comparator is a reward-only convolutional predictor, reward_greedy; it sees pixels, actions, and observed rewards from the random replay buffer, but no hidden state or expert action labels. The metric reported below is the signed return gap to the matched one-step reference. It is deliberately not clipped: a negative value would mean that an arm exceeded the heuristic, while zero is the reference by construction. The reference is a reproducible heuristic, not a dynamic-programming optimum.

Table 20: Per-held-out-world control gaps at $N = 256$. Values are means over the three training worlds and three seeds for each held-out evaluation world; latent arms use planning depth 3 and reward_greedy uses depth 0.

Held-out world	latent8	wide16	reward_greedy	random
mid_target	1.024	0.881	0.758	3.728
high_inertia	0.868	0.721	0.652	4.050
noisy_mid_target	1.154	1.041	0.824	4.282

At $N = 256$, matched-reference action agreement on the same three rows is $(0.459, 0.484, 0.440)$ for latent8, $(0.477, 0.500, 0.451)$ for wide16, and $(0.566, 0.626, 0.524)$ for reward_greedy; the reward-only arm is therefore closer to the heuristic even though it has no multi-step latent planner. Endpoint frame MSE is also not a monotone proxy for control: on the held-out aggregate at $N = 256$, latent8 at horizons $(1, 3, 5, 8)$ gives

$(0.087, 0.085, 0.086, 0.087)$ and wide16 gives $(0.077, 0.077, 0.081, 0.082)$. These are endpoint errors on an independent random-action trajectory, not closed-loop returns.

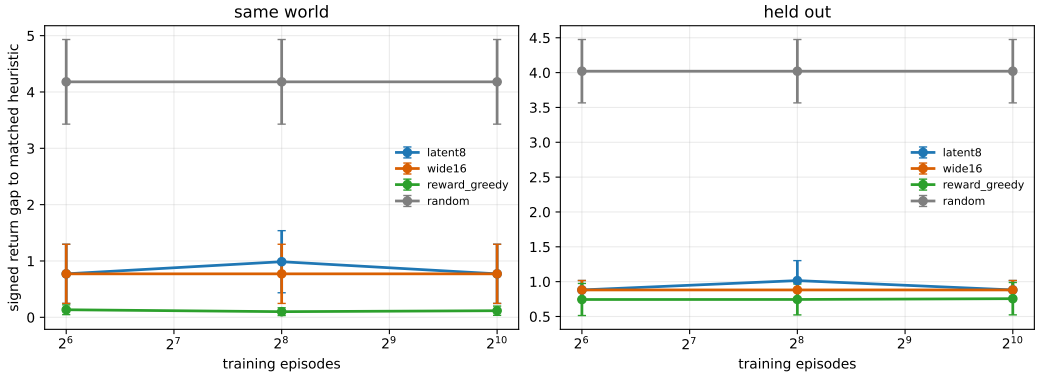

Figure 5: Paired control transfer curves. Error bars are sample SD over the fixed train-world/evaluation-world/seed cells. The held-out panel is a transfer stress test, not a population generalization curve.

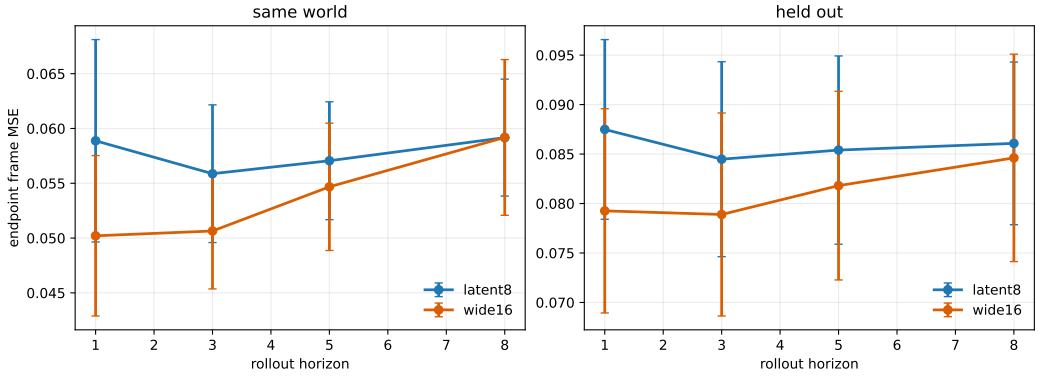

Figure 6: Endpoint frame prediction error at horizons 1, 3, 5, and 8. The figures separate same-world and held-out evaluation and retain the non-monotone finite-budget behavior.

The public v3 receipt contains 972 cell-level control rows, 864 endpoint-rollout rows, 81 checkpoints, 729 logged update rows, and full per-episode action/reward traces. Its verification checks arm and split sets, exact zero gap for the reference, checkpoint reload parity, and the expected row counts. No D10 executor or external signer was available; the status remains exploratory. The appropriate conclusion is that the current latent planners do not beat a simple reward-only pixel baseline under this budget, and that a wider latent does not repair the held-out transfer gap.

I.1 Checkpoint action-collapse audit

The repeated control values warranted a direct behavioral audit of the saved models. Without retraining or selecting a checkpoint, we reloaded all 81 v3 fits, replayed the same matched-start tapes, and compared every learned action and reward trace with the stored receipt. This gives 324 learned-policy cells and 7,776 episode traces with exact parity. We also evaluated fixed actions -1, 0, and $+1$ from the same starts. The table reports the fixed-world cell mean after averaging within each seed; "constant" means that all 24 episodes and 12 actions in that cell used one action. The $+1$ column is the learned return minus the always-$+1$ return, so zero indicates behavioral equivalence on the fixture.

The latent and wide models therefore often reproduce a fixed rightward policy because all initial positions lie to the left of the target. This explains why some con-

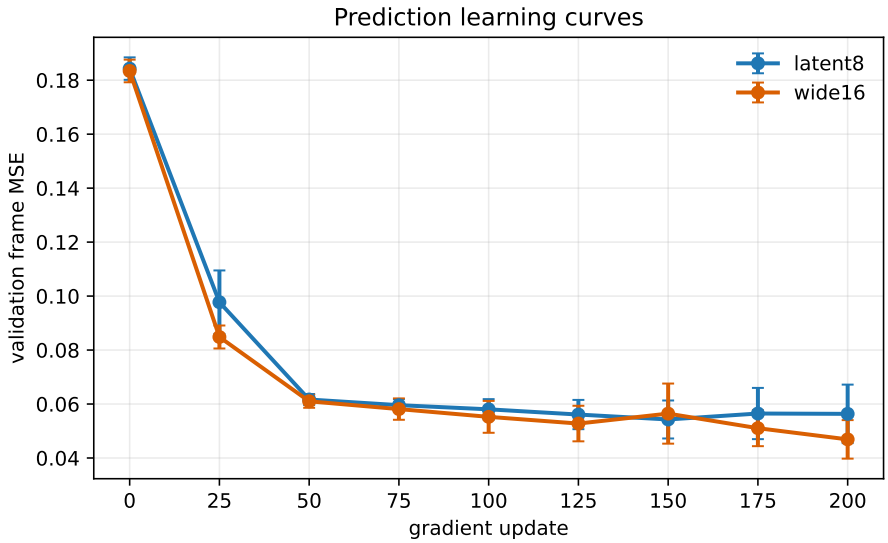

Figure 7: Validation frame-MSE trajectories for all 54 latent model fits. The curves document optimization progress; they do not justify selecting a checkpoint by test return.

Table 21: Saved-checkpoint action audit. The audit is post-hoc and reuses the original v3 tapes; it is not an independent test set.

Split	N	Arm	Constant +1 cells	Learned − fixed +1
Held out	64	latent8	27/27	0.000 ± 0.000
Held out	64	wide16	27/27	0.000 ± 0.000
Held out	256	latent8	21/27	-0.134 ± 0.119
Held out	256	wide16	27/27	0.000 ± 0.000
Held out	1024	latent8	27/27	0.000 ± 0.000
Held out	1024	wide16	27/27	0.000 ± 0.000
Same world	256	latent8	7/9	-0.215 ± 0.203
Same world	256	wide16	9/9	0.000 ± 0.000
Same world	256	reward_greedy	0/9	0.671 ± 0.054

trol curves are exactly tied and why planning depth does not change the reported action. It is evidence of policy-level action collapse on this small fixture, not a general statement about the encoder or latent representations. The complete 324-cell CSV, fixed-control traces, checkpoint hashes, and replay receipt are released under 02_experiments/20260925_visual_policy_collapse_audit_v1/; the audit remains local, post-hoc, and not admitted.

J Complete proof packets

The following self-contained packets expand the upper transfer, all adaptive lower mechanisms, and the positive three-state boundary. They are included here so that the supplement does not depend on a repository-local Appendix L, Appendix T, or an unpublished proof file. The wrapper keeps the upper and lower packets on the same audited constant convention.

K A complete two-state learning upper bound

This appendix gives the upper-bound construction used in Theorem 1. It is an information-theoretic finite-state argument, not a neural training result. The proof has four interfaces: thirteen bounded reset-triple moments, a legal one-dimensional contrast mesh, reward-specific transfer on signed prefixes, and a finite stored-pointer belief controller. All policies below are total public-history policies, including histories assigned zero probability by a fitted candidate.

K.1 Model, policy space, and target

Write $S, X \in \{-1, +1\}$ and $A \in \{0, 1\}$. A legal model is described by $(\mu, \rho, b_0, \beta_0, b_1, \beta_1)$ with

$$\Pr(X = x \mid S = s) = \frac{1 + x(\mu + \rho s)}{2}, \qquad \mathbb{E}[S' \mid S = s, A = a] = b_a + \beta_a s, \tag{1}$$

where $\rho \geq 0$, $|\mu| + \rho \leq 1$, and $|b_a| + |\beta_a| \leq 1$. Each episode starts from an independent uniform hidden state. The learner receives N reset episodes, each with three observations and two effective actions, and then outputs a policy for a fresh H-observation deployment. Rewards are $(1 + X_t)/2$ and are counted before the action at that time.

Let $V_M(\pi)$ denote expected deployment return and let V_M^* be the maximum over total observable-history policies. The upper construction below obeys, for $H \geq 8$ and $N \geq 1$,

$$\mathbb{E}[V_M^* - V_M(\widehat{\pi})] \leq \min\left\{H, \sqrt{22}\, H^{3/2}\sqrt{\mathbb{E}w} + \varepsilon_{\text{plan}}, K_H \mathbb{E}w + \varepsilon_{\text{plan}}\right\}, \tag{2}$$

where w is the fitted raw-moment discrepancy, $K_H = H(2H-1)(3H-1)/2$, and $\varepsilon_{\text{plan}}$ is the finite-grid planning error. We next define one learner for which $\mathbb{E}w \leq C_{\text{fast}} N^{-1/2}$.

K.2 Thirteen moments and their algebra

Under an intervention word (a, b) define

$$\mu = \mathbb{E}[X_0], \qquad P_a = \mathbb{E}[X_1 \mid do(a)], \qquad Q_a = \mathbb{E}[X_0 X_1 \mid do(a)],$$
$$R_{ab} = \mathbb{E}[X_2 \mid do(a, b)], \qquad S_{ab} = \mathbb{E}[X_0 X_2 \mid do(a, b)].$$

The two actions in each position give thirteen raw quantities. Define

$$C_a = P_a - \mu, \quad M_a = Q_a - \mu P_a, \quad D_{ab} = R_{ab} - \mu, \quad N_{ab} = S_{ab} - \mu R_{ab}. \tag{3}$$

Uniform initialization and conditional emission independence give the exact identities

$$C_a = \rho b_a, \qquad M_a = \rho^2 \beta_a, \qquad D_{ab} = C_b + \beta_b C_a, \qquad N_{ab} = \beta_b M_a. \tag{4}$$

These identities are observable moment identities; no hidden-coordinate identification or division by ρ is assumed.

For two legal models, let w be the maximum absolute discrepancy between the thirteen raw moments, let $e = |\mu - \mu'|$, and let v be the maximum discrepancy among C, M, D, N. Since raw moments lie in $[-1, 1]$,

$$e \leq w, \qquad |\Delta C|, |\Delta D| \leq 2w, \qquad |\Delta M|, |\Delta N| \leq 3w, \qquad v \leq 3w. \tag{5}$$

K.3 Concentration under the reset-triple contract

The logger used by the upper bound samples both actions independently and uniformly. The per-episode features are

$$X_0, \quad 2\mathbf{1}\{A_0 = a\}X_1, \quad 2\mathbf{1}\{A_0 = a\}X_0 X_1,$$

$$4\mathbf{1}\{A_0 = a, A_1 = b\}X_2, \quad 4\mathbf{1}\{A_0 = a, A_1 = b\}X_0 X_2,$$

for $a, b \in \{0, 1\}$. They are unbiased for the thirteen quantities and are bounded in $[-4, 4]$. Let E_N be their maximum empirical-coordinate error. A log-sum-exp bound over the two-sided tails gives, for every $\lambda > 0$,

$$\mathbb{E}E_N \leq \frac{\log(26)}{\lambda} + \frac{8\lambda}{N}, \qquad \mathbb{E}E_N \leq \sqrt{\frac{32\log(26)}{N}}.$$

No independence among the thirteen features within an episode is required.

K.4 Fast legal-model fit

Set $m = \lceil 13\sqrt{N} \rceil$ and enumerate only the contrast values $r_j = j/m$, $j = 0, \dots, m$. At a mesh point $r > 0$, clip the empirical offset to $\mu_r \in [-1 + r, 1 - r]$ and center all empirical moments using that same offset. Put

$$K = \max\{r^2, |\widehat{C}_0|, |\widehat{C}_1|\}.$$

If $K \le r^2 + 10^{-15}$, initialize $\widehat{\beta}_b$ from $\widehat{M}_b/r^2$; otherwise use the maximizing index c (with $c = 0$ on a tie) and initialize from $(\widehat{D}_{cb} - \widehat{C}_b)/\widehat{C}_c$. Clip the initializer to $[-1, 1]$. For each action, project $(\widehat{C}_b, \widehat{\beta}_b)$ onto

$$\frac{|C|}{r} + |\beta| \le 1 \tag{6}$$

in the weighted sup norm $\max\{|\Delta C|, K|\Delta\beta|\}$, then set $b = C/r$. At $r = 0$ set $b = \beta = 0$ and perform no division.

For completeness, the projection is explicit. With $A = |\widehat{C}_b|$ and $B = K|\widehat{\beta}_b|$, take

$$t = \max\left\{0, A - r, B - K, \frac{KA + rB - rK}{K + r}\right\},$$

and set

$$C = \operatorname{sgn}(\widehat{C}_b)(A - t)_+, \qquad \beta = \operatorname{sgn}(\widehat{\beta}_b)\frac{(B - t)_+}{K}.$$

The candidate is legal by construction. Reconstruct all thirteen raw moments and select the candidate with smallest maximum residual; ties retain the first candidate in the mesh enumeration order.

Here is the residual calculation used by the fast fit. Fix a mesh value r and a legal comparison model with the same contrast and maximum raw discrepancy δ from the empirical vector. Centering gives errors at most δ for C, D and at most 2δ for M, N. Either initializer satisfies

$$K|\widehat{\beta}_b - \beta_b^*| \le 3\delta.$$

The weighted projection therefore gives

$$|\widehat{C}_b - C_b^*| \le 4\delta, \qquad K|\widehat{\beta}_b - \beta_b^*| \le 6\delta, \qquad |\widehat{M}_b - M_b^*| \le 6\delta.$$

Substitution into the four identities in (4) bounds the reconstructed raw residuals by $4, 10, 16, 28$ times δ for the corresponding moment families.

Here is the constant bookkeeping behind this last display. For the fixed mesh point, the fitted and comparison models use the same offset, so the projection bounds imply the following raw-coordinate residuals:

raw family	P	Q	R	S
maximum error	4δ	10δ	16δ	28δ

Thus a conservative uniform residual bound for the selected candidate is 29δ; this envelope includes the shared centering and offset coordinates in the selection residual. At the true contrast, choose $r^* = \lfloor m\rho \rfloor/m$ and retain the true transition coefficients. Because $|\mu_{r^*}| + r^* \le 1$, this comparison point is legal; the clipping error in its offset is at most E_N. The raw polynomial perturbation and offset clipping give an empirical comparison residual

$$\delta_* \le 13h + 14E_N, \qquad h = 1/m.$$

Candidate selection and one final triangle inequality therefore give

$$w \le 29\delta_* + E_N \le 29(13h + 14E_N) + E_N = 377h + 407E_N.$$

For reference, the raw moment polynomials in the legal coordinates are

$$\begin{aligned} P_a &= \mu + \rho b_a, \\ Q_a &= \mu^2 + \mu\rho b_a + \rho^2\beta_a, \\ R_{ab} &= \mu + \rho b_b + \beta_b\rho b_a, \\ S_{ab} &= \mu^2 + \mu\rho b_b + \mu\beta_b\rho b_a + \beta_b\rho^2\beta_a. \end{aligned}$$

Each monomial is unit bounded on the legal domain. Telescoping its factors along the mesh perturbation gives coordinate Lipschitz constants $(1, 3, 8, 6, 13)$ for (μ, P, Q, R, S), respectively. This is the bound used above for δ_*. Consequently

$$\mathbb{E}w \le C_{\text{fast}} N^{-1/2}, \qquad C_{\text{fast}} = 29 + 407\sqrt{32\log(26)}. \tag{7}$$

The fit is a legal-model approximation, not a parameter-recovery theorem.

K.5 Signed-prefix transfer

For a public history, let p be its probability mass and let r be ρ times its hidden signed mass. A physical prefix satisfies $p \ge 0$ and $|r| \le \rho p$, and initially $(p, r) = (1, 0)$. Emission x followed by action a is represented by

$$B_{a,x} = \frac{1}{2}\begin{pmatrix} 1+\mu x & x \\ C_a(1+\mu x) + xM_a & \beta_a(1+\mu x) + xC_a \end{pmatrix}, \qquad g_\mu(p, r) = \frac{(1+\mu)p + r}{2}. \tag{8}$$

The same public policy probability multiplies each branch in the two models. Writing $R_j = \sum_h |r_h|$ before observation j and $q = \max_a(|C_a|, |M_a|)$, physicality and (8) imply

$$R_j \le \min\{\rho, 2jq\}. \tag{9}$$

If $q > 0$, choose a C_c or M_c attaining q. The moment identities imply

$$(\beta_b - \beta'_b)C_c = \Delta D_{cb} - \Delta C_b + \beta'_b \Delta C_c, \qquad (\beta_b - \beta'_b)M_c = \Delta N_{cb} + \beta'_b \Delta M_c,$$

and hence $\max_b |\beta_b - \beta'_b| \le 3v/q$.

Insert $B - B'$ once along physical prefixes from M and retain fitted suffixes from M'. Summing over observations and the common policy kernels gives

$$\sum_h |\Delta p_h| \le e, \qquad \sum_h |\Delta r_h| \le (3 + 6j)v + 2e, \tag{10}$$

and, using $R_j \le \rho$ instead,

$$\sum_h |\Delta r_h| \le 3v(1 + 1/s) + 2e, \qquad s = q/\rho, \tag{11}$$

when $\rho, q > 0$. The inserted vector is an algebraic signed vector; it is not assumed to be a posterior under M'.

For a fixed continuation policy with k future rewards, let J_+ and J_- be its values from the two hidden signs under M'. Since each reward lies in an interval of width ρ', $|J_+ - J_-| \le k\rho'$ and $0 \le J_\pm \le k$. Thus for arbitrary signed input

$$J(p, r) = \frac{p(J_+ + J_-)}{2} + \frac{r(J_+ - J_-)}{2\rho'}, \qquad |J(p, r)| \le k|p| + \frac{k}{2}|r|. \tag{12}$$

At $\rho' = 0$, (12) follows by approaching along the legal path $\rho'_\eta = \eta$, $\mu'_\eta = (1 - \eta)\mu'$ and taking the finite-policy limit. This also covers deterministic emission corners.

K.6 Reward transfer and the signal split

Telescoping the finite history tree with true prefixes and fitted suffixes, the immediate reward row costs $e/2$ per time. An insertion before the jth future transition has $k = H - j - 1$ remaining rewards, so

$$d_H(M, M') := \sup_\pi |V_M(\pi) - V_{M'}(\pi)| \le A_H e + L_H v, \tag{13}$$

$$A_H = \frac{H(2H-1)}{2}, \qquad L_H = \frac{H(H-1)(2H-1)}{4}. \tag{14}$$

For $\rho, q > 0$, using (11) gives the sharper signal-sensitive form with coefficient $(3/4)H(H-1)$ on $v(1 + 1/s)$.

Within a single true model, after an observation let m be the hidden posterior mean. The next observation mean differs from μ by at most

$$\rho|b_a + \beta_a m| \le |C_a| + |M_a|/\rho \le 2s. \tag{15}$$

The initial reward is policy-independent, so every policy has regret at most $2(H-1)s$. If $\widehat{\pi}$ is an exact optimizer for M', inserting the true and fitted optima gives $\mathrm{Regret}_M(\widehat{\pi}) \le 2d_H$. Combining (5) and (14) yields

$$\text{Regret} \le K_H w, \qquad K_H = \frac{H(2H-1)(3H-1)}{2}, \tag{16}$$

$$\text{Regret} \le 11H^2 w/s, \tag{17}$$

$$\text{Regret} \le \sqrt{22}\, H^{3/2}\sqrt{w}. \tag{18}$$

The final line balances the first policy-span bound with the second transfer bound. If $\rho = 0$ or $q = 0$, observations are policy-independent and every policy has the same value, so these cases require no division.

K.7 Finite stored-pointer belief controller

For a fixed fitted model, let $p = \Pr(S = +1 \mid h)$ be the post-observation belief, write $p_+ = p$ and $p_- = 1 - p$, and let k count future rewards. Define

$$O_x(s') = \frac{1 + x(\mu + \rho s')}{2}, \qquad T_a(s' \mid s) = \frac{1 + s'(b_a + \beta_a s)}{2},$$

$$z_{a,x,s'}(p) = O_x(s') \sum_s T_a(s' \mid s) p_s, \qquad q_{a,x}(p) = \sum_{s'} z_{a,x,s'}(p),$$

$$\tau_{a,x}(p) = z_{a,x,+}(p)/q_{a,x}(p), \qquad F_k(p) = \max_a \sum_x q_{a,x}(p) \left[\frac{1+x}{2} + F_{k-1}(\tau_{a,x}(p))\right],$$

with $\tau = 0$ when $q = 0$ and $F_0 = 0$. Let $G = \{0, 1/M, \dots, 1\}$ and let Π be nearest-grid projection with a fixed tie rule. At each grid point store both observation successors $\mathrm{next}(g, a, x) = \Pi(\tau_{a,x}(g))$ and the maximizing action under the recursion

$$u_k(g) = \max_a \sum_x q_{a,x}(g) \left[\frac{1+x}{2} + u_{k-1}(\mathrm{next}(g, a, x))\right].$$

The controller executes these stored pointers; it does not re-filter with the unknown true model. Impossible fitted branches still receive a stored successor and a total action definition.

Every fixed policy subtree with k future rewards is $k\rho$-Lipschitz in the initial belief. If $\delta = 1/(2M)$, induction over the stored pointers gives both the planning-table error and the executed-policy error at most $\rho\delta k(k-1)/2$. Accounting for the initial projection gives

$$0 \le F_k(p) - J_k^{\Pi(p)}(p) \le \rho\delta k(k+1). \tag{19}$$

Take $M = \max\{1, H(H-1)m/2\}$ and initialize from the first observation. For $H = 1$ this is one inert interval and there is no action layer. Since $\rho \le 1$, the fitted-model execution loss is at most

$$\varepsilon_{\text{plan}} \le 1/m. \tag{20}$$

K.8 Conclusion and complexity

The same executed learner satisfies all three bounds in (18). Jensen and (7) give

$$\mathbb{E}\,\text{Regret} \le \min\left\{H, \sqrt{22C_{\text{fast}}}H^{3/2}N^{-1/4} + \frac{1}{m}, C_{\text{fast}}K_H N^{-1/2} + \frac{1}{m}\right\}. \tag{21}$$

Because $1/m \le N^{-1/2} \le F(N, H)$ for $H \ge 8$, this implies the main-paper constant $C = 1 + \max\{\sqrt{22C_{\text{fast}}}, 3C_{\text{fast}}\}$.

Moment accumulation costs $O(N)$ and evaluation of the $m+1$ contrast candidates costs $O(\sqrt{N})$ with constant fitting workspace. The stored belief table costs $O(H^3\sqrt{N})$ arithmetic operations and entries. Thus the total arithmetic complexity of this fast fit and planner is

$$O\left(N + \sqrt{N} + H^3\sqrt{N}\right),$$

For genuine empirical inputs, exact fitted parameters and grid values have polynomial bit height; this is a bit-complexity statement, not a practical wall-clock guarantee.

Scope. The proof uses known uniform resets, binary alphabets, two hidden states, homogeneous kernels, one fresh deployment episode, and the observed reward $(1+X)/2$. It does not establish parameter recovery, unknown-reset control, arbitrary latent rank, fixed-floor matching lower bounds, neural performance, or representation superiority.

L Adaptive lower bound for reset-triple control

This section proves the lower side of the reset-triple theorem. The proof uses complete adaptive likelihoods, so the training logger may adapt within and across episodes. The deployment policy may use its complete observable history and may infer a hidden sign during deployment.

L.1 Experiment and policy quantifiers

Hidden states and observations are $S, X \in \{-1, +1\}$ and actions are $A \in \{0, 1\}$. A legal homogeneous model has conditionally independent emissions and transitions

$$\Pr(X = x \mid S = s) = \frac{1 + x(\mu + \rho s)}{2}, \qquad \rho \geq 0, \quad |\mu| + \rho \leq 1, \tag{22}$$

$$\Pr(S' = s' \mid S = s, A = a) = \frac{1 + s'(b_a + \beta_a s)}{2}, \qquad |b_a| + |\beta_a| \leq 1. \tag{23}$$

Every training episode starts from an independent uniform hidden reset and has record $(X_0, A_0, X_1, A_1, X_2)$. There are N episodes. The learner may select actions using earlier records, the current observable prefix, and private randomness, but it observes neither hidden states nor a model oracle. Deployment starts from one further independent uniform reset, has H observations and $H-1$ effective actions, and has reward $Y_j = (1+X_j)/2$ before action A_j and its transition. There is no reset inside deployment.

Policies are total on every observable history, including histories having zero probability under a particular model. Let

$$V_M(\pi) = \mathbb{E}_M^{\pi}\left[\sum_{j=0}^{H-1} Y_j\right], \qquad R_{\mathcal{C}}(N, H) = \inf_{\mathcal{A}} \sup_{M \in \mathcal{C}} \mathbb{E}\left[V_M^* - V_M(\pi_{\mathcal{A}})\right], \tag{24}$$

where V_M^* is the optimum over total observable-history policies. The class $\mathcal{C}_{\mathrm{sym}}$ fixes only $\mu = 0$; nonzero action drifts remain legal. The full class $\mathcal{C}_{\mathrm{gen}}$ additionally allows the unknown offset. Define

$$F(N, H) = \min\{H, H^{3/2}N^{-1/4}, H^3N^{-1/2}\}. \tag{25}$$

Lower-bound statement. For all integers $H \geq 8$ and $N \geq 1$,

$$R_{\mathcal{C}_{\mathrm{sym}}}(N, H) \geq \frac{F(N, H)}{176160768}. \tag{26}$$

By class inclusion the same lower applies to $\mathcal{C}_{\mathrm{gen}}$. The hard models below may depend on N, H while remaining in the same fixed class.

L.2 Adaptive record overlap

Let $\sigma \in \{0,1\}$ label a pair of models. For an intervention word (a,b) let $p_\sigma^{ab}(x)$ be the complete observation-triple mass for $x \in \{-1,+1\}^3$. Conditional on a fixed previous training record, the logged one-episode law has the factorization

$$P_\sigma(x,a,b) = w(x,a,b)p_\sigma^{ab}(x), \qquad w(x,a,b) = \pi_0(a \mid x_0)\pi_1(b \mid x_0,a,x_1), \tag{27}$$

with

$$\sum_{a,b} w(x,a,b) = 1, \qquad \sum_{x,a,b} w(x,a,b) = 8. \tag{28}$$

The action kernels are common at a common observable history. Thus (27) is a causal likelihood factorization; it does not condition on an adaptive action sequence and reinterpret it as an intervention.

Suppose every triple atom obeys $p_\sigma^{ab}(x) \geq 1/64$ and every pairwise atom difference obeys $|p_0^{ab}(x) - p_1^{ab}(x)| \leq B$. Omitting zero-weight atoms, $\mathrm{KL}(P_0\|P_1) \leq \chi^2(P_0,P_1)$ gives

$$\mathrm{KL}(P_0\|P_1) \leq \sum_{x,a,b} w(x,a,b)\frac{(p_0^{ab}(x) - p_1^{ab}(x))^2}{p_1^{ab}(x)} \tag{29}$$

$$\leq 512B^2. \tag{30}$$

Place all exogenous algorithmic randomness, including randomness used during deployment, in a seed sampled before interaction. It is independent of the model and contributes zero KL. Conditional on that seed and any common past record, the fresh reset makes the same one-episode calculation applicable. The KL chain rule and Pinsker's inequality therefore imply, for the full training-record laws λ_0, λ_1,

$$\mathrm{KL}(\lambda_0\|\lambda_1) \leq 512NB^2, \qquad \tau_T := \mathrm{TV}(\lambda_0,\lambda_1) \leq 16B\sqrt{N}. \tag{31}$$

No unconditional iid assertion about adaptive logs is used.

We also use a common-part overlap inequality. For nonnegative measurable losses $g_0(T), g_1(T)$ satisfying $g_0(T)+g_1(T) \geq 2G$ pointwise, let ν have density $\min(d\lambda_0/d\eta, d\lambda_1/d\eta)$ with respect to $\eta = \lambda_0 + \lambda_1$. Since ν has mass $1-\tau_T$ and is dominated by both record laws,

$$\frac{\mathbb{E}_{\lambda_0} g_0 + \mathbb{E}_{\lambda_1} g_1}{2} \geq \frac{1}{2}\int (g_0 + g_1)\, d\nu \geq (1-\tau_T)G. \tag{32}$$

True conditional regrets are nonnegative. Whenever a proxy loss is used below, its nonnegativity is established before invoking (32).

L.3 Absorbing action-label pair

Set $\rho = 1/2$, $\delta = 1/(64\sqrt{N})$, and $\mu = 0$. In model σ, action σ is safe with $(b,\beta) = (0,1)$; the other action is unsafe with $(b,\beta) = (-\delta, 1-\delta)$. State -1 is absorbing under both actions, while state $+1$ survives with probability one under the safe action and with probability $1-\delta$ under the unsafe action. These transitions are legal and emissions have probabilities $1/4$ and $3/4$.

Let $D_j = \{S_0 = +1, S_j = -1\}$. The known-model optimum is $V_\sigma^* = H/2$, attained by always using the safe action. For every policy, the exact identity

$$r_\sigma(\pi) := V_\sigma^* - V_\sigma(\pi) = \rho \sum_{j=1}^{H-1} \Pr_\sigma(D_j) \tag{33}$$

holds, since $S_j = S_0 - 2\mathbf{1}\{D_j\}$ and $\mathbb{E}[X_j] = \rho\mathbb{E}[S_j]$.

For a fixed training intervention word, couple the two initial signs and emissions until hidden paths separate. Conditional on initial $+1$ and no separation, exactly one model has hazard δ at each step; separation over two steps is at most $\frac{1}{2}[1-(1-\delta)^2] \leq \delta$. Consequently every

triple atom differs by at most δ. The emission floor gives $p_\sigma^{ab}(x) \geq 1/64$, so (31) yields $\tau_T \leq 1/4$ for fully adaptive training.

Fix a common training record and any full-history deployment policy. Share the initial sign, policy randomness, and emissions while both copies remain $+1$. Before the first death their histories and actions coincide, and exactly one selected action is unsafe. The conditional probability of first death at each such step is δ. After death, allow arbitrary divergence with the correct marginal laws; the dead copy cannot recover. Therefore

$$\Pr_0(D_j) + \Pr_1(D_j) \geq \frac{1}{2}[1 - (1-\delta)^j], \tag{34}$$

and

$$r_0(\pi_T) + r_1(\pi_T) \geq \frac{\rho}{2} \sum_{j=1}^{H-1} [1 - (1-\delta)^j]. \tag{35}$$

This is an all-feedback-policy statement. Since $1 - (1-\delta)^j \geq j\delta/(1+j\delta) \geq \frac{1}{2}\min\{j\delta, 1\}$, the indices $j = \lceil H/2 \rceil, \ldots, H-1$ give

$$\sum_{j=1}^{H-1} [1 - (1-\delta)^j] \geq \frac{H}{12} \min\{H\delta, 1\}. \tag{36}$$

Combining (32)–(36), taking the worst model and then the minimax infimum, gives

$$R_{\mathcal{C}_{\mathrm{sym}}}(N, H) \geq \frac{H}{128} \min\left\{ \frac{H}{64\sqrt{N}}, 1 \right\} \geq \frac{1}{8192} \min\left\{ H, \frac{H^2}{\sqrt{N}} \right\}. \tag{37}$$

In particular this is $H/8192$ throughout $N \leq H^2$.

L.4 Symmetric flips and deployment-time sign inference

Assume $H \geq 64$ and $H^2 \leq N \leq H^4/4096$. Set

$$\delta = \frac{1}{64H}, \qquad \rho = \frac{1}{2}\sqrt{H}\, N^{-1/4}, \qquad \mu = 0. \tag{38}$$

Action σ is the identity and action $1-\sigma$ flips the hidden sign with probability δ. Thus $b_a = 0$, $\beta_{\mathrm{safe}} = 1$ and $\beta_{\mathrm{flip}} = 1 - 2\delta$. The pair is legal, $\rho \leq 1/2$, $\delta H = 1/64$, and $H\rho^2 \geq 16$.

For a fixed action word (a, b), uniform reset and zero drift imply that all odd moments vanish and

$$8p_\sigma^{ab}(x) = 1 + \rho^2 \beta_{\sigma,a} x_0 x_1 + \rho^2 \beta_{\sigma,b} x_1 x_2 + \rho^2 \beta_{\sigma,a} \beta_{\sigma,b} x_0 x_2. \tag{39}$$

Each beta changes by 2δ and each product by at most 4δ. Thus atom differences are at most $\rho^2\delta$, all atoms are at least $1/64$, and (31) gives

$$\mathrm{KL}(\lambda_0 \| \lambda_1) \leq 512 N \rho^4 \delta^2 = \frac{1}{128}, \qquad \tau_T \leq \frac{1}{16}. \tag{40}$$

For model σ, let $F_k(p)$ be the optimal value before the next observation with k rewards remaining and belief $p = \Pr(S = +1)$; let $G_k(p)$ be the value after a current observation with k future rewards. With $F_0 = G_0 = 0$,

$$T_{\mathrm{safe}}(p) = p, \qquad T_{\mathrm{flip}}(p) = \delta + (1-2\delta)p, \tag{41}$$

$$q_x(p) = \frac{1 + \rho x(2p-1)}{2}, \qquad p_x = \frac{p(1+\rho x)}{2q_x(p)}, \tag{42}$$

$$G_k(p) = \max_a F_k(T_a(p)), \tag{43}$$

$$F_k(p) = \sum_{x \in \{-1,+1\}} q_x(p) \left[\frac{1+x}{2} + G_{k-1}(p_x) \right]. \tag{44}$$

For any fixed total feedback policy π over k observations, $J_{\pi,k}(p)$ is affine in p. A union bound on conditional flip hazards gives $\mathbb{E}[S_t \mid S_0 = +1] \geq 1 - 2t\delta$ and $\mathbb{E}[S_t \mid S_0 = -1] \leq -1 + 2t\delta$. Hence

$$J_{\pi,k}(1) - J_{\pi,k}(0) \geq L_k := k\rho[1 - \delta(k-1)] \geq k\rho/2, \tag{45}$$

and, by evaluating a maximizing policy at the smaller belief, $F_k(p_2) - F_k(p_1) \geq L_k(p_2 - p_1)$ for $p_2 \geq p_1$. Since $T_{\text{safe}}(p) - T_{\text{flip}}(p) = \delta(2p-1)$, a wrong action loses at least $\delta L_k|2p-1|$ in the Bellman comparison (with equality allowed at the zero-sign tie).

Fix a common training record and the learner policy. Let $P_{\sigma,j}$ be the joint law of (S_j, h_j), where $h_j = (X_0, A_0, \ldots, A_{j-1}, X_j)$, and put $d_P(h) = P(+1, h) - P(-1, h)$. Define

$$A_{\sigma,j}(P) = \sum_h \left[\pi_j(1 - \sigma \mid h)(d_P(h))_+ + \pi_j(\sigma \mid h)(-d_P(h))_+\right]. \tag{46}$$

Its summands are 1-Lipschitz in d_P, so $|A_{\sigma,j}(P) - A_{\sigma,j}(Q)| \leq 2\,\mathrm{TV}(P, Q)$. Zero-mass histories need no division.

Let R_j use the same initial sign, emissions and policy but no transitions. Coupling until the first actual flip gives $\mathrm{TV}(P_{\sigma,j}, R_j) \leq j\delta$. At this common reference, opposite action labels imply

$$A_{0,j}(P_{0,j}) + A_{1,j}(P_{1,j}) \geq A_{0,j}(R_j) + A_{1,j}(R_j) - 4j\delta = \mathbb{E}_R|\mathbb{E}_R[S_0 \mid h_j]| - 4j\delta. \tag{47}$$

Under R_j, the observations are conditionally iid with mean ρS_0. Majority-sign estimation and Chebyshev imply

$$\mathbb{E}_R|\mathbb{E}_R[S_0 \mid h_j]| \geq 1 - \frac{2}{(j+1)\rho^2}. \tag{48}$$

For $J_H = \{\lceil H/4\rceil, \ldots, \lfloor H/2\rfloor\}$, $|J_H| \geq H/8$, $j + 1 \geq H/4$, and $H - j - 1 \geq H/4$. Since $H\rho^2 \geq 16$ and $4H\delta = 1/16$, the right-hand side of (47) is at least $1/4$ for every $j \in J_H$.

The actual Bellman losses telescope along the learner trajectory:

$$r_\sigma(T) = \sum_{j=0}^{H-2} \mathbb{E}_{P_\sigma}\left[G_{H-j-1}(p_{\sigma,j}) - F_{H-j-1}(T_{\sigma,A_j}(p_{\sigma,j}))\right]. \tag{49}$$

Each summand is nonnegative. Keeping only $j \in J_H$ and using (45)–(47) gives, pointwise in T,

$$r_0(T) + r_1(T) \geq \frac{\rho H}{16384}. \tag{50}$$

Applying (32) and taking the minimax infimum proves

$$R_{\mathcal{C}_{\text{sym}}}(N, H) \geq \frac{H^{3/2}N^{-1/4}}{131072}, \qquad H \geq 64, \quad H^2 \leq N \leq H^4/4096. \tag{51}$$

The deployment policy is fully history-dependent; useful inference of the hidden sign is explicitly allowed.

L.5 Clipped weak signal: the large-budget regimes

Assume $H \geq 8$ and $N \geq H^4$. Put $\alpha = 1/8192$ and

$$\rho = \alpha \min\{\sqrt{H}N^{-1/4}, H^2N^{-1/2}\}, \qquad \mu = 0, \tag{52}$$

$$b_0 = b_1 = 1/H, \qquad \beta_0 = 1 - 2/H, \quad \beta_1 = 1 - 3/H. \tag{53}$$

Model σ assigns β_0 to action σ and β_1 to the other action. If $\beta = 1 - c/H$, $c \in \{2, 3\}$, the two off-diagonal transition probabilities are $(c-1)/(2H)$ and $(c+1)/(2H)$, so the model is legal. The smallest transition probability is of order $1/H$; no fixed transition floor is used. Also $\rho \leq \alpha/\sqrt{H} < 1/64$.

For a fixed intervention word (a, b), write $C = \rho/H$, $M_a = \rho^2\beta_a$, $D_{ab} = C(1 + \beta_b)$, and $K_{ab} = \rho^2\beta_a\beta_b$. The complete triple law is the Walsh expansion

$$\begin{aligned} 8p^{ab}_\sigma(x) = 1 &+ Cx_1 + D_{ab}x_2 + M_a x_0 x_1 + K_{ab}x_0x_2 \\ &+ (C^2 + M_b)x_1x_2 + CM_a x_0x_1x_2. \end{aligned} \tag{54}$$

Indeed, $\mathbb{E}S_1 = 1/H$, $\mathbb{E}S_2 = (1+\beta_b)/H$, $\mathbb{E}S_0S_1 = \beta_a$, $\mathbb{E}S_0S_2 = \beta_a\beta_b$, $\mathbb{E}S_1S_2 = 1/H^2 + \beta_b$, and $\mathbb{E}S_0S_1S_2 = \beta_a/H$. These moments and the independent emission channel give all eight atoms.

Across the two labels, β changes by $1/H$ and a product changes by at most $2/H$. Since $|C| = \rho/H$, every atom difference is at most

$$B = \frac{\rho/H^2 + 5\rho^2/H}{8}, \qquad p_\sigma^{ab}(x) \geq \frac{1}{64}. \tag{55}$$

Using (31) and the two branches of (53),

$$\tau_T \leq 2\sqrt{N}\left(\frac{\rho}{H^2} + \frac{5\rho^2}{H}\right) \leq 2(\alpha + 5\alpha^2) < \frac{1}{4}. \tag{56}$$

It remains to compare the labels under deployment. Given a common training record T, let P_σ be the true deployment path law. Let Q_σ retain the same reset, transitions, and policy but drive the policy with independent fair observations. The two Q_σ hidden laws may differ; only the observation/action-word law is common and independent of transition randomness. In the P_σ/Q_σ likelihood ratio, only emissions change. Since

$$k(\rho) = \frac{1+\rho}{2}\log(1+\rho) + \frac{1-\rho}{2}\log(1-\rho) \leq \rho^2, \tag{57}$$

the chain rule and Pinsker imply

$$\mathrm{TV}(P_\sigma, Q_\sigma) \leq \rho\sqrt{H/2} \leq \frac{\alpha}{\sqrt{2}} =: \tau_D. \tag{58}$$

The true value and the physical-reward proxy

$$\overline{V}_\sigma(\pi) = \frac{H}{2} + \frac{\rho}{2}\mathbb{E}_{Q_\sigma}^{\pi}\left[\sum_{t=0}^{H-1} S_t\right] \tag{59}$$

therefore satisfy $|V_\sigma(\pi) - \overline{V}_\sigma(\pi)| \leq H\rho\tau_D$. The test function has range length $H\rho$, so this is one whole-path TV comparison, not a sum of H reward-span bounds.

Let the comparator always choose the β_0 action. Its hidden mean is $m_t^* = (1-\beta_0^t)/2$, and its true and proxy values coincide because its actions ignore observations. For an arbitrary common action word under Q_σ, write $m_{\sigma,0} = 0$ and $m_{\sigma,t+1} = 1/H + \beta_{\sigma,a_t} m_{\sigma,t}$. If $d_{\sigma,t} = m_t^* - m_{\sigma,t}$, then

$$d_{\sigma,t+1} = \beta_{\sigma,a_t} d_{\sigma,t} + \mathbf{1}\{a_t \neq \sigma\}\frac{m_t^*}{H} \geq 0. \tag{60}$$

Summing the two labels and then the rewards gives

$$\frac{\rho}{4}\sum_{t=0}^{H-1}(d_{0,t} + d_{1,t}) \geq \frac{\rho}{24}\sum_{j=0}^{H-2}(1-\beta_0^j)(1-\beta_1^{H-1-j}). \tag{61}$$

For $j \in \{\lceil H/4\rceil, \ldots, \lfloor H/2\rfloor\}$, there are at least $H/8$ indices, and $1-\beta_0^j \geq 1-e^{-1/2}$, $1-\beta_1^{H-1-j} \geq 1-e^{-3/4}$. Hence the right hand side of (61) is at least $C_{\mathrm{gap}} H\rho$, where

$$C_{\mathrm{gap}} = \frac{(1-e^{-1/2})(1-e^{-3/4})}{192} \geq \frac{1}{1344}. \tag{62}$$

The inequality follows from $1-e^{-x} \geq x/(1+x)$ at $x = 1/2, 3/4$.

For every common record and policy, define the nonnegative proxy loss $g_\sigma(T) = V_\sigma^{\mathrm{lo}} - \overline{V}_\sigma(\pi_T)$. The preceding bound gives $g_0(T) + g_1(T) \geq 2C_{\mathrm{gap}} H\rho$. Applying (32), then subtracting the deployment-transfer error, shows that every learner has average true regret at least

$$H\rho\,[C_{\mathrm{gap}}(1-\tau_T) - \tau_D] \geq \frac{C_{\mathrm{gap}}}{2} H\rho. \tag{63}$$

Since $C_{\mathrm{gap}}/2 \geq 1/2688$ and $H\rho = \alpha\min\{H^{3/2}N^{-1/4}, H^3N^{-1/2}\}$,

$$R_{\mathcal{C}_{\mathrm{sym}}}(N,H) \geq \frac{\min\{H^{3/2}N^{-1/4}, H^3N^{-1/2}\}}{22020096}, \qquad N \geq H^4. \tag{64}$$

L.6 Budget bridges and endpoint completion

At fixed H, $R_{\mathcal{C}}(N,H)$ is nonincreasing in N, since a learner may discard extra training episodes. At $N = H^4$, (64) gives $\sqrt{H}/22020096$, and therefore the same lower persists for every $N \le H^4$. If $H^4/4096 \le N \le H^4$, then $H^{3/2}N^{-1/4} \le 8\sqrt{H}$; the clipped lower bridges the middle range for $8 \le H < 64$. For $H \ge 64$, (51) covers the remaining interval $H^2 \le N \le H^4/4096$. Combining this with (37) for $N \le H^2$ and the two branches of (64) for $N \ge H^4$ proves (26). The rate function has joins at $N = H^2$ and $N = H^6$; H^4 is only a proof construction boundary.

M Positive three-state reset-triple obstruction

We next give a separate rank-three boundary example. It is an explicit $H = 4$ obstruction, not a rank-three rate or an arbitrary-horizon theorem.

M.1 Models and clean triple laws

The hidden state set is $\{0,1,2\}$, actions are $\{0,1\}$, and observations are $\{-1,+1\}$. Each episode starts from $p = (1,1,1)/3$. Let the clean observation labels be $(-1,-1,+1)$ and define the clean transition maps

	a=0	a=1
M	$(0,1,0)$	$(0,2,0)$
M'	$(1,0,0)$	$(0,2,0)$

where an entry is the successor of current states $0,1,2$ respectively. For the four action words, the clean triples generated from initial states $0,1,2$ are:

action word	M			M'		
	$s_0=0$	$s_0=1$	$s_0=2$	$s_0=0$	$s_0=1$	$s_0=2$
00	$\langle -,-,-\rangle$	$\langle -,-,-\rangle$	$\langle +,-,-\rangle$	$\langle -,-,-\rangle$	$\langle -,-,-\rangle$	$\langle +,-,-\rangle$
01	$\langle -,-,-\rangle$	$\langle -,-,+\rangle$	$\langle +,-,-\rangle$	$\langle -,-,+\rangle$	$\langle -,-,-\rangle$	$\langle +,-,-\rangle$
10	$\langle -,-,-\rangle$	$\langle -,+,-\rangle$	$\langle +,-,-\rangle$	$\langle -,-,-\rangle$	$\langle -,+,-\rangle$	$\langle +,-,-\rangle$
11	$\langle -,-,-\rangle$	$\langle -,+,-\rangle$	$\langle +,-,-\rangle$	$\langle -,-,-\rangle$	$\langle -,+,-\rangle$	$\langle +,-,-\rangle$

Thus the two members have the same multiset of triples for every fixed intervention word; their complete clean triple laws coincide.

Replace each clean transition and emission by

$$T_{m,a}(s,j) = \frac{3}{4}\mathbf{1}\{j = f_{m,a}(s)\} + \frac{1}{12}, \tag{65}$$

$$O(+1 \mid s) = \frac{(1,1,7)_s}{8}, \qquad O(-1 \mid s) = 1 - O(+1 \mid s). \tag{66}$$

Transition entries are $5/6$ or $1/12$, and emission entries are $1/8$ or $7/8$. Equivalently, each transition retains the clean map with probability $3/4$ and resets uniformly with probability $1/4$, while each clean label is flipped independently with probability $1/8$.

Condition on all internal reset locations. The hidden path splits into independent clean segments beginning from the uniform distribution, each with at most three labels. The clean table and its marginals make the segment laws equal for M and M'. Mixing over the identical reset flags and then applying the common label-flip channel preserves equality. Hence all four complete noisy interventional triple laws agree exactly.

M.2 Adaptive training indistinguishability

For a prior training record z and private seed U, the likelihood of an episode path is

$$p_m(x_0,x_1,x_2 \mid do(a_0,a_1))\pi_0(a_0 \mid x_0,z,U)\pi_1(a_1 \mid x_0,a_0,x_1,z,U). \tag{67}$$

The first factor is equal for the two models and the policy factors are common. Fresh uniform resets give, by induction over episodes, identical complete adaptive training laws for every finite N, including algorithms with cross-episode adaptation. Integrating over U does not change equality.

M.3 Full-history Bellman recursion

Use unnormalized joint member-state masses $v_{m,s}$ before an observation. Let D_x be the block-diagonal emission matrix and T_a the block-diagonal member-preserving transition matrix. With reward $r(x) = (1 + x)/2$, define

$$F_0(v) = 0, \quad F_h(v) = \sum_{x\in\{-1,+1\}} \left[r(x) \sum_{m,s} (vD_x)_{m,s} + \max_a F_{h-1}(vD_xT_a) \right]. \tag{68}$$

At the start $v = (\lambda p, (1-\lambda)p)$. Induction on h shows that (68) is optimal over all observable-history policy trees, including policies that update the member posterior during deployment. Randomization cannot improve the pointwise maximum.

M.4 Exact $H = 4$ minimax calculation

At the last decision, action 1 improves the final reward by $9b_1/16$ in either member, so $A_2 = 1$ is optimal at every history. Set

$$r = (1, 1, 7)/8, \qquad T_1r = (3, 12, 3)/16, \qquad w = r + T_1r = (5, 14, 17)/16. \tag{69}$$

At the preceding decision, the action advantages (action 1 minus action 0) are

$$\Delta_M(b) = \frac{9b_1}{64}, \qquad \Delta_{M'}(b) = \frac{9(4b_1 - 3b_0)}{64}. \tag{70}$$

The initial observation probabilities are 3/8 and 5/8 and the hidden state posteriors are

$$b(+1) = (1, 1, 7)/9, \qquad b(-1) = (7, 7, 1)/15. \tag{71}$$

After either first action, the predictive laws before the second observation are:

X_0, A_0	$\Pr(S_1 = 0, 1, 2 \mid X_0, A_0)$
$+, 0$	$(3/4, 1/6, 1/12)$
$+, 1$	$(3/4, 1/12, 1/6)$
$-, 0$	$(29/60, 13/30, 1/12)$
$-, 1$	$(29/60, 1/12, 13/30)$

States 0 and 1 have the same emission law, so observing X_1 leaves the member posterior unchanged in each row. At prior λ the mixed advantage is

$$\Delta_\lambda(b) = \frac{9[(4 - 3\lambda)b_1 - 3(1 - \lambda)b_0]}{64}. \tag{72}$$

At $\lambda = 67/72$, the resulting optimal middle actions are:

X_0, A_0	known M	known M'	$\lambda = 67/72$
$+, 0$	1	0	1
$+, 1$	1	0	0
$-, 0$	1	1	1
$-, 1$	1	0	either

For the three rewards at times $1, 2, 3$, define

$$u_{M,0} = \frac{(35, 62, 83)}{64}, \quad u_{M',0} = \frac{(62, 35, 83)}{64}, \quad u_1 = \frac{(35, 71, 83)}{64}. \tag{73}$$

Direct dot products show that the advantage of choosing $A_0 = 1$ over $A_0 = 0$ is positive for both members and the mixture, namely

X_0	known M	known M'	$\lambda = 67/72$
$+$	$1/64$	$1/16$	$3/128$
$-$	$21/320$	$57/256$	$21/320$

Thus the full-history optimum is attained by $A_0 = 1, A_2 = 1$, with known M using $A_1 = 1$ and known M' using $A_1 = 0$. The resulting values are

$$V_M^* = \frac{39}{32}, \qquad V_{M'}^* = \frac{363}{256}, \qquad V_{67/72} = \frac{1251}{1024}. \tag{74}$$

Consider the two Bayes-optimal policies

$$\begin{aligned} Q &: A_0 = A_2 = 1, \quad A_1 = 0, \\ P &: A_0 = A_2 = 1, \quad A_1 = 1 \text{ iff } X_0 = -1. \end{aligned}$$

Their return and regret table is

policy	V_M	$V_{M'}$	$V_M^* - V_M$	$V_{M'}^* - V_{M'}$
Q	$309/256$	$363/256$	$3/256$	0
P	$2487/2048$	$2703/2048$	$9/2048$	$201/2048$

The prior-67/72 Bayes value is 1251/1024, while

$$\frac{67}{72} V_M^* + \frac{5}{72} V_{M'}^* - V_{67/72} = \frac{67}{6144}. \tag{75}$$

Equation (75) lower-bounds the maximum regret of every deployment policy, and equality is attained by mixing P with probability $1/9$ and Q with probability $8/9$ once per deployment episode. Consequently, for every finite N,

$$R_{\{M,M'\}}(N, 4) = \frac{67}{6144}. \tag{76}$$

M.5 Controlled Hankel rank

Let $D_+ = \mathrm{diag}(r)$ and $D_- = I - D_+$. Before an observation, take compatible prefixes $\varnothing, (+,1), (-,1)$ and suffixes $\varnothing, +, (+,1,+)$. Their reachable rows are

$$R = \mathrm{rows}\left[\frac{(1,1,1)}{3}, \frac{(9,1,2)}{32}, \frac{(29,5,26)}{96}\right], \qquad \det R = -\frac{1}{64}, \tag{77}$$

and the suffix columns are $C = [1, r, D_+ T_1 r]$ with $\det C = -27/512$. The compatible word-probability block is

$$K = RC = \begin{pmatrix} 1 & 3/8 & 3/32 \\ 3/8 & 3/32 & 81/4096 \\ 5/8 & 9/32 & 231/4096 \end{pmatrix}, \qquad \det K = \frac{27}{32768}. \tag{78}$$

Every compatible controlled word factors through the three hidden coordinates, so rank is at most three; (78) gives rank at least three. Both noisy processes therefore have controlled Hankel rank exactly three. This rank statement does not assert a one-step revealing condition, a full-column-rank emission matrix, or invertible transitions.

Scope. The obstruction is at $H = 4$. It gives neither an arbitrary-horizon three-state rate nor a worst-case minimal identifying window. Equality of short laws alone would not imply a control gap; the Bellman calculation above supplies that missing ingredient.