COMPRESSING WHAT AN AGENT CAN STEER: CONTROL-AWARE STRUCTURAL ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

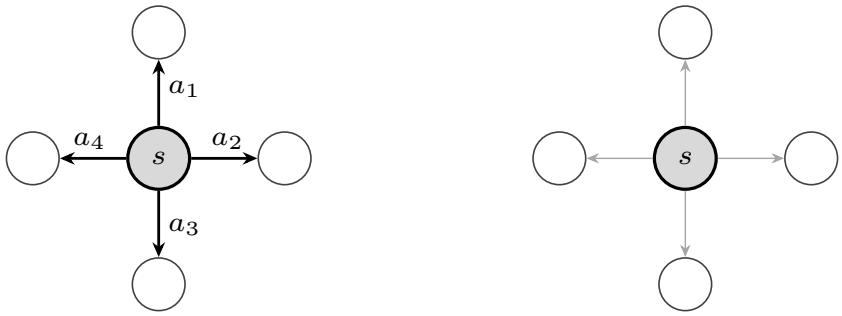
A state abstraction for reinforcement learning should preserve what the agent can control, yet reward-free codelength objectives for compressing transition graphs score how predictable the dynamics are once actions are averaged out, not how far an agent can steer them. We prove that this gap is structural for structural entropy (SE), a random-walk codelength used to build hierarchical state abstractions: the SE codelength of a Markov decision process under *any* encoding tree is a function of the action-marginalised kernel alone, so no SE variant computed from the passive chain can distinguish two MDPs that differ only in their action-conditioned structure. We then define the minimal repair: the *control gain at resolution* T, $\Delta H_{\mathrm{ctrl}}(T) = I(A; m(S') \mid m(S))$, the codelength saved when the block-level code may condition on actions. At the finest partition it equals stationary-averaged uniform-prior empowerment (exactly Klyubin–Salge empowerment in balanced deterministic MDPs), and it is monotone under coarsening. Each estimate is calibrated against a matched null, a within-state action shuffle that preserves the passive chain and all per-state action marginals exactly. On tabular MDP families where controllability is a construction knob *independent* of connectivity, the calibrated estimator recovers the constructed gain to $\sim$0.015 bits (Spearman $\rho = 0.98$, $p \leq 10^{-4}$, in both families), while effective rank is blind where the passive chain is fixed ($\rho = -0.02$); the passive SE baseline breached its preregistered blindness threshold ($\rho = 0.34$, $p = 0.10$), which a post-hoc 100-run extension attributes to sampling noise. In quotient planning on a checkerboard-inversion gridworld (gate **failed**) no reward-free partition objective we tested — passive, action-conditioned, or profile consistency — reliably yields plannable quotients, although plan-preserving partitions exist and the control-aware objective found one (retained 0.95 of optimal return, in 1 of 15 cells). On 15 learned agents both exploratory predictions hold in sign but not significantly. All gates and verdicts are reported verbatim.

1 INTRODUCTION

Compressing an agent's experience into a smaller state space is a central tool in reinforcement learning and world modelling: abstractions shrink planning problems, and learned representations are routinely judged by how compactly they summarise the transition dynamics. The theory of state abstraction is clear about what must survive the compression. Model minimisation and bisimulation (Givan et al., 2003; Ferns et al., 2004), MDP homomorphisms (Ravindran and Barto, 2004) and the abstraction hierarchy of Li et al. (2006) characterise aggregations that preserve values or optimal policies; deep variants learn them from reward and transition losses (Gelada et al., 2019; Zhang et al., 2021). Empowerment (Klyubin et al., 2005; Salge et al., 2014) is the reward-free version: a good state keeps the agent's ability to influence its future.

Reward-free compression objectives measure something else. A family of graph objectives scores a partition of a transition graph by the description length of a random walk on it (Li and Pan, 2016; Rosvall and Bergstrom, 2008), and such objectives have been used to build hierarchical state abstractions for RL (Zeng et al., 2023); the latent transition graph of a learned world model is a natural substrate for them. Once actions are averaged out, however, what remains is a Markov chain, and any score of that chain measures how predictable the dynamics are, not how far an agent

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.



(a) *controlled* state: action $a_i \rightarrow$ neighbour i

(b) *uncontrolled* state: every action $\rightarrow$ uniform

marginal kernel P_β **identical** in (a) and (b), to machine precision
$\Rightarrow$ every passive code length, every effective rank, and every null built from the passive graph is **identical** too (Prop. 1)
control gain ΔH_{ctrl}: **2 bits** in (a), **0 bits** in (b)

Figure 1: A marginal cannot see a conditional. Both states have four actions and the same four successors. In (a) each action selects one successor; in (b) every action gives the uniform distribution over all four. Averaged over any fixed behaviour policy the two are the *same* Markov chain, so every compression-based diagnostic of the passive graph, and every null built from that graph, returns the same number for both — while the ability to act differs by the full two bits.

can steer them: the two states of Figure 1 form the same chain yet differ by two bits of control. A representation can be highly compressible and useless for acting.

This paper asks what it takes to make such an objective control-aware. We take structural entropy (SE; Li and Pan, 2016), the codelength of a stationary random walk under an encoding tree, as the concrete instance, prove that its blindness to control is structural, and repair it minimally by letting the code condition on the executed action. Calibrated against a null that preserves the passive chain exactly, the resulting control gain recovers a controllability knob that no passive measure can see. It does not, by itself, select abstractions that preserve planning: in a preregistered test no reward-free partition objective we tested did, although plan-preserving partitions exist.

Our contributions:

1. **A blindness theorem** (Prop. 1): the SE codelength of an MDP under any fixed encoding tree depends on the behaviour policy only through the action-marginalised kernel. Passive SE cannot see control, for the same reason a marginal cannot see a conditional.
2. **A control-aware codelength with an exact null.** The control gain at resolution T (Def. 1) is the codelength saving when the block-level code may condition on actions. It equals stationary-averaged uniform-prior empowerment at the finest resolution (Prop. 2; Klyubin–Salge empowerment with balanced deterministic fibres, Prop. 3) and is monotone under coarsening (Prop. 4). Its matched null, the within-state action shuffle, preserves the passive chain and per-state action marginals exactly and absorbs the plug-in estimator's finite-sample bias.
3. **Validators by construction**: MDP families where controllability is a knob independent of connectivity, including one whose passive kernel is *identical* at every knob setting, plus a deterministic regression fixture. The calibrated gain tracks the knob in both families, while effective rank is blind where the passive kernel is fixed (Sec. 5.1).
4. **Where it stops**: quotient MDP planning on a gridworld whose action semantics invert on a checkerboard — provably invisible to any passive measure. Its preregistered gate failed: neither the control-aware nor the plain-SE objective reliably yields plannable quotients, and

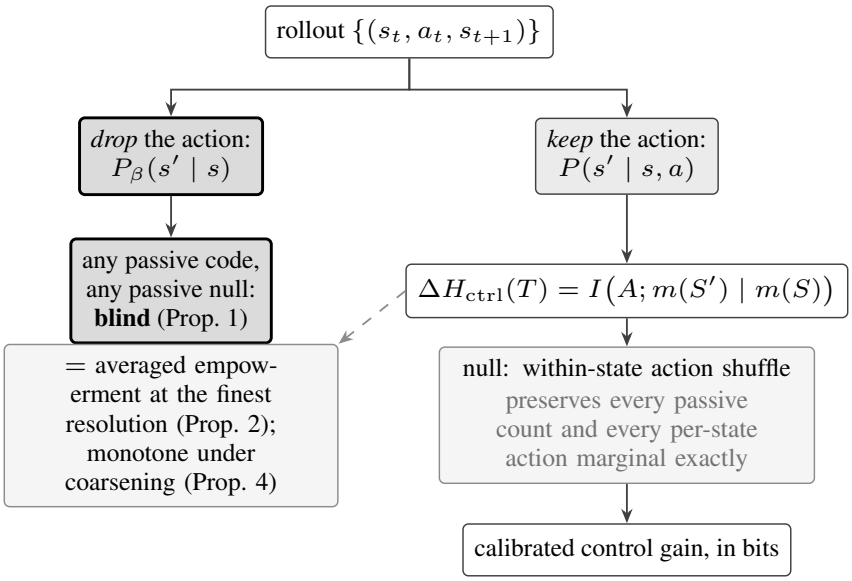


Figure 2: The repair. Keeping the executed action in the conditioning turns the codelength saving into a conditional mutual information, in the same units as the passive quantity. The matched null shuffles action labels within each state and destroys only the action–outcome coupling.

we report why. On 15 TwoRooms agents trained for this study with actions logged, both exploratory predictions hold in sign, neither significantly.

2 Background and setup

Let $M = (\mathcal{S}, \mathcal{A}, P)$ be a finite MDP explored by the uniform-random behaviour policy; $|\mathcal{A}| = A$. The *passive kernel* is $P_u(s' \mid s) = \frac{1}{A}\sum_a P(s' \mid s, a)$, with stationary distribution π. Generalising SE to directed chains by taking volumes from the stationary distribution in place of degrees, the one- and two-level SE of a chain Q under a partition $m : \mathcal{S} \to \{1..K\}$ are defined from π and the block volumes; we write $L(T; Q)$ for the two-level codelength under partition (tree) T. Self-loops shape π but never count as boundary flow.

We observe a rollout $\{(s_t, a_t, s_{t+1})\}_{t<N}$ and use visitation-weighted plug-in entropies (bits). The two passive baselines are *plain SI*, the two-level structural information (SE) of the empirical passive chain, and *effective rank*, the exponentiated entropy of the normalised singular values of the empirical passive kernel (of the learned latents in Sec. 5.3).

3 Theory

Proposition 1 (Passive SE is action-blind)**.** *Fix any encoding tree T and any stationary behaviour policy β. The SE codelength of the walk induced by β under T is a function of the marginal kernel $P_\beta(s' \mid s) = \sum_a \beta(a \mid s)P(s' \mid s, a)$ alone. Consequently, two MDPs with the same P_β receive the same SE (raw or null-calibrated) for every T, regardless of their action-conditioned structure.*

Proof. Every term of the SE codelength (one-level, two-level, or general tree) is an expectation of a per-step cost $c_T(s, s')$ under the stationary edge measure $\pi(s)P_\beta(s' \mid s)$: the within-block code $-\log(\pi_{s'}/V_{m(s')})$ and the boundary terms are functions of (s, s') and T only. The edge measure is linear in P_β, and π itself is a function of P_β. No term depends on $P(\cdot \mid s, a)$ except through the mixture. Any null model that is itself a function of the passive graph (degree-preserving rewiring included) inherits the same invariance. □

Definition 1 (Control gain at resolution T)**.** *For a partition m with blocks B, define the two-level codelength of the* controlled *walk, whose block-level code may condition on the executed action:*

$$L_{\mathrm{ctrl}}(T) = H(m(S') \mid m(S), A) + \sum_B \pi_B \, H\big(\pi_{\cdot|B}\big)\,,$$

$$L_{\mathrm{pass}}(T) = H(m(S') \mid m(S)) + \sum_B \pi_B \, H\big(\pi_{\cdot|B}\big)\,,$$

with entropies under the uniform policy's stationary law; the control gain at resolution T *is*

$$\Delta H_{\mathrm{ctrl}}(T) \;=\; L_{\mathrm{pass}}(T) - L_{\mathrm{ctrl}}(T) \;=\; I(A;\, m(S') \mid m(S)) \;\geq 0.$$

At the finest (singleton) partition we write $\Delta H_{\mathrm{ctrl}} = I(A; S' \mid S) = \sum_s \pi(s)\, I_s(A; S')$.

L_{pass} is the entropy-rate form of the two-level code; the plain-SI baseline of Sec. 5.1 instead uses the cut-based directed SE of Li and Pan (2016). Prop. 1 covers both, and the within-block term cancels in $\Delta H_{\mathrm{ctrl}}(T)$.

Proposition 2 (Identity with uniform-prior empowerment)**.** *$\Delta H_{\mathrm{ctrl}} = \sum_s \pi(s)\, I_s(A; S')$ where I_s is the mutual information of the channel $a \mapsto P(\cdot \mid s, a)$ under the uniform action prior. Hence $\Delta H_{\mathrm{ctrl}} \leq \sum_s \pi(s)\, \mathfrak{E}(s)$, where $\mathfrak{E}(s) = \max_{p(a)} I_s(A; S')$ is the empowerment of Klyubin et al. (2005) (see also Salge et al., 2014), with equality whenever the uniform prior is capacity-achieving (e.g. for symmetric channels).*

Proof. The first equality is the chain rule for conditional mutual information with $S \sim \pi$; the inequality is the definition of channel capacity as a maximum over priors. □

Proposition 3 (Deterministic case)**.** *If $P(\cdot \mid s, a)$ is deterministic with image $\mathcal{R}(s)$ ($|\mathcal{R}(s)| = r_s$ distinct successors) and fibre sizes $k_1, \ldots, k_{r_s}$, then $\mathfrak{E}(s) = \log_2 r_s$ and*

$$\mathfrak{E}(s) - I_s(A; S') \;=\; \mathrm{KL}\left(\left(\tfrac{k_j}{A}\right)_j \,\middle\|\, \mathrm{Unif}(\mathcal{R}(s)) \right) \;\leq\; \log_2 \frac{k_{\max}}{k_{\min}},$$

so ΔH_{ctrl} equals average empowerment exactly whenever fibres are balanced (e.g. distinct deterministic successors: $\Delta H_{\mathrm{ctrl}} = \mathbb{E}_\pi[\log_2 r_s]$).

Proof. For a deterministic channel $H(S' \mid A, s) = 0$, so $I_s = H(S' \mid s) = H((k_j/A)_j)$; capacity is attained by any prior uniform on one representative per fibre, giving $\log_2 r_s$. Subtract. □

Proposition 4 (Monotone under coarsening)**.** *For any partition m, $\Delta H_{\mathrm{ctrl}}(T) \leq \Delta H_{\mathrm{ctrl}}$, and if m refines m' then $\Delta H_{\mathrm{ctrl}}(T') \leq \Delta H_{\mathrm{ctrl}}(T)$.*

Proof sketch. (i) $I(A; m(S') \mid S) \leq I(A; S' \mid S)$ by the data-processing inequality, since $m(S')$ is a function of S'. (ii) Under the uniform policy A is independent of S, so the chain rule gives $I(A; m(S') \mid S) = I(A; m(S') \mid m(S)) + I(A; S \mid m(S'), m(S))$, hence $I(A; m(S') \mid m(S)) \leq I(A; m(S') \mid S)$. Chaining (i) and (ii) through the refinement gives both claims. Equality in (ii) holds iff $I(A; S \mid m(S'), m(S)) = 0$, in particular when all states in a block induce the same block-level channel $a \mapsto p(m(s') \mid s, a)$ — a *control-sufficient* abstraction, the information-theoretic analogue of an MDP homomorphism. □

Remark 1 (Matched null)**.** *The plug-in $\widehat{\Delta H_{\mathrm{ctrl}}}(T)$ is biased upward at finite N. We calibrate against the* within-state action shuffle*: independently for each state s, permute the action labels among the rollout records leaving s. This null preserves every passive count $n(s, s')$ (hence the passive chain, its degrees, its stationary distribution, and every passive SE quantity exactly) and every marginal $n(s, a)$; it destroys precisely the action–outcome coupling that ΔH_{ctrl} claims to measure, and its mean equals the plug-in bias under the null of no coupling. We report $\widehat{\Delta H_{\mathrm{ctrl}}}^{\,\mathrm{cal}} = \widehat{\Delta H_{\mathrm{ctrl}}} - \overline{\widehat{\Delta H_{\mathrm{ctrl}}}_{\mathrm{null}}}$ throughout (10 shuffles).*

Remark 2 (Position)**.** *By Prop. 4, maximising retained control gain at fixed K is a soft, reward-free relaxation of MDP homomorphism (Ravindran and Barto, 2004); unlike bisimulation metrics (Ferns et al., 2004) it needs no reward, and unlike empowerment maximisation it scores a* representation*, not a policy. Successor representations (Dayan, 1993) are passive in the sense of Prop. 1.*

Table 1: All preregistered gates, criteria as registered, observed values and verdicts. $\Delta H^{\mathrm{cal}}_{\mathrm{ctrl}}$: null-calibrated control gain; $\mathrm{SI}^{\mathrm{cal}}$: null-calibrated plain SI; ρ: Spearman against the knob f (G1–G3) or planning success (G6).

Gate	Setting	Criterion (preregistered)	Observed	Verdict
G1	F1	$\rho(\Delta H^{\mathrm{cal}}_{\mathrm{ctrl}}) \geq 0.8$ and per-f means strictly increasing	$\rho = 0.981$; increasing	**PASS**
G2	F1	$\lvert\rho(\mathrm{SI}^{\mathrm{cal}})\rvert \leq 0.3$ and $\lvert\rho(\text{eff. rank})\rvert \leq 0.3$	0.337 and -0.016	**FAIL**
G3	F2	$\rho(\Delta H^{\mathrm{cal}}_{\mathrm{ctrl}}) \geq 0.8$; separation if both baselines $\leq \rho - 0.3$	$\rho = 0.981$; baselines ≤ 0.373	**PASS**
G4	F3	ctrl $>$ plain in $\geq 4/5$ seeds at every K, and median gap ≥ 0.25 at K=12	wins $2/5, 2/5, 1/5$; gap -0.001	**FAIL**
G5	fixture	all checks C1–C6 pass	$6/6$ pass	**PASS**
G6	15 agents	exploratory: (i) $\rho(\Delta H^{\mathrm{cal}}_{\mathrm{ctrl}}) > 0$; (ii) above $\rho(\mathrm{SI}^{\mathrm{cal}})$	0.176 vs. -0.110; $p = 0.54$, 0.70	in sign, n.s.

4 Validators by construction

F1 (mixture-invariant knob). 6×6 torus gridworld, A=4. A fraction f of states is *controlled* (action a moves deterministically to neighbour a); the rest map every action to the uniform distribution over the same four neighbours. The passive kernel is the uniform torus walk for every f — *identical by construction* (verified to machine precision), so by Prop. 1 every passive measure is provably flat in f up to sampling noise, while the true control gain is $2f$ bits (Prop. 3).

F2 (robustness knob). 40-state random MDPs, $P(\cdot \mid s, a) = (1 - f)\, C_s + f\, D_{s,a}$ on a shared 6-state support, with $C_s \sim \mathrm{Dir}(\mathbf{1})$ and $D_{s,a} \sim \mathrm{Dir}(0.3 \cdot \mathbf{1})$; the passive kernel is *not* held fixed, making F2 the honest robustness family for the estimator and its null.

F3 (checkerboard inversion). 6×6 gridworld, slip 0.1, goal with teleport-to-start, three blocked cells per seed. On odd-parity cells the four actions are permuted by the 180° rotation. Since a permutation of actions cannot change the uniform action mixture, the passive kernel is invariant to the inversion (verified exactly): *no passive measure can see the parity axis*. A lifted block policy that merges the two parities is direction-neutral by symmetry.

Fixture. A deterministic regression fixture (supplementary material) packages a minimal validation suite: the identities of Props. 2, 3 and 4 (C1, C2, C4), exact passive invariance of F1/F3 (C3), and thresholds for the calibrated gain (C5) and plain SI (C6) on an f=1 vs f=0 pair (all pass).

5 Experiments

All gates were preregistered before the decisive runs (see supplementary material); 5 seeds per condition; Spearman correlations with 10^4-permutation p-values (floor 10^{-4}); every quantity ships with its matched null. Table 1 lists every gate, its criterion as registered, and its verdict.

5.1 Knob tracking and separation (G1–G3)

Rollouts of N=20,000 steps; knob $f \in \{0, .25, .5, .75, 1\}$; seeds 0–4 (25 runs per family).

F1. The null-calibrated control gain tracks the knob at pooled Spearman $\rho = 0.981$ ($p \leq 10^{-4}$), with per-f means -0.0003, 0.498, 0.988, 1.487, 1.985 bits against true values 0, 0.5, 1, 1.5, 2: the estimator is not merely ordinal, it is *calibrated* to the constructed gain within $\sim$0.015 bits. Effective rank is flat ($\rho = -0.016$, $p = 0.94$). Null-calibrated plain SI came out at $\rho = 0.337$ ($p = 0.103$).

Gate verdicts (preregistered). G1 **PASS** ($\rho = 0.981 \geq 0.8$, means strictly increasing). G2 **FAIL**: the preregistered criterion was $\lvert\rho\rvert \leq 0.3$ for both baselines, and plain SI's 0.337 breaches it. The breach cannot be a real effect: F1 holds the *law* of the passive chain exactly fixed across f (Sec. 4), and SI is a function of the passive rollout alone, so its law does not depend on f and its population correlation with f is zero; the permutation test agrees ($p = 0.103$, not significant). The preregistration error was thresholding a 25-point rank correlation at 0.3 (null s.d. ≈ 0.2) instead of gating

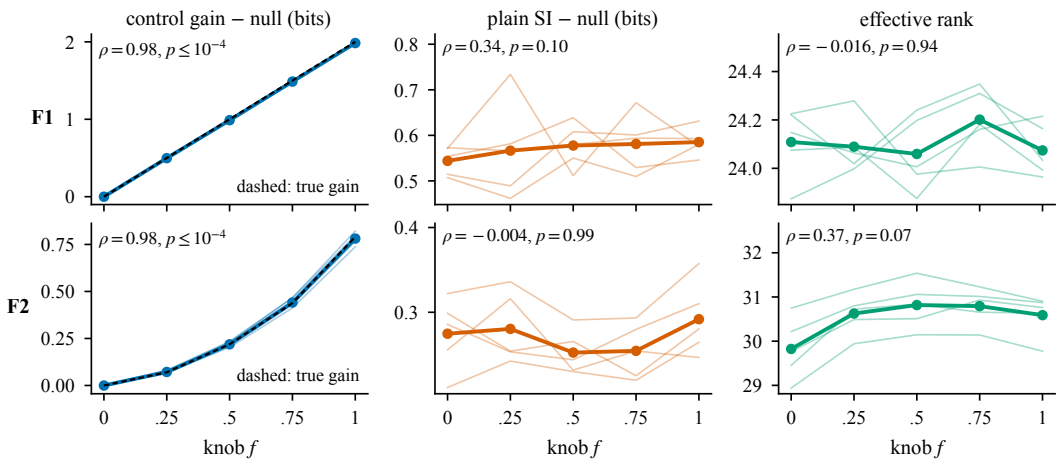


Figure 3: Knob tracking (F1 top, F2 bottom). Thin lines: the 5 seeds; thick: mean. Left: null-calibrated control gain against the true gain (dashed; the two coincide in F1). Middle: null-calibrated plain SI. Right: effective rank. Each panel: pooled Spearman ρ against f over 25 runs (Table 1).

Table 2: Quotient planning on F3: retained control $(J - J_{\text{rand}})/(J^* - J_{\text{rand}})$ per seed, for the passive (P), control-aware (C) and post-hoc profile-consistency (H) objectives; higher is better; 0 is the uniform-random policy, 1 optimal. Bold: the only cell above 0.25. Wins: seeds with C > P on unrounded values, as the gate counts them; seed 0 at K=18 and K=12 is a tie (margin $< 10^{-5}$), so the substantive count is $1/5$ at every K.

	$K = 18$			$K = 12$			$K = 6$		
Seed	P	C	H	P	C	H	P	C	H
0	−0.035	−0.035	−0.023	−0.035	−0.035	−0.035	−0.035	−0.034	−0.035
1	−0.021	−0.037	−0.020	−0.037	−0.001	−0.036	−0.035	−0.037	−0.036
2	0.030	−0.036	−0.036	−0.036	−0.036	−0.019	−0.029	−0.036	−0.036
3	0.228	**0.952**	−0.029	0.216	−0.041	−0.041	−0.038	−0.041	−0.039
4	0.105	−0.048	−0.039	0.007	−0.048	−0.044	−0.037	−0.048	−0.039
Median	0.030	−0.036	−0.029	−0.035	−0.036	−0.036	−0.035	−0.037	−0.036
Wins		2/5			2/5			1/5	

on significance. A *post-hoc, clearly labelled* extension to 20 additional seeds (100 more runs) gives $\rho_{\text{SI}} = -0.043$ ($p = 0.68$) on the fresh sample and $\rho_{\text{SI}} = 0.027$ ($p = 0.76$) pooled over all 125 runs, while $\rho_{\Delta H_{\text{ctrl}}} = 0.980$ throughout — the breach was sampling noise, as invariance requires.

F2 (passive kernel *not* held fixed). Control gain: $\rho = 0.981$ ($p \leq 10^{-4}$), calibrated means again tracking the exact kernel gain ($0.072/0.219/0.441/0.781$ against $0.070/0.219/0.438/0.791$ bits for $f > 0$). Plain SI: $\rho = -0.004$ ($p = 0.99$). Effective rank drifts mildly with the knob ($\rho = 0.373$, $p = 0.066$) — the Dirichlet mixing that raises controllability also sharpens rows of the passive kernel, which effective rank picks up; it remains far below the separation margin. G3 **PASS** ($\rho \geq 0.8$; separation condition met: $0.373 \leq 0.981 - 0.3$).

Per-knob detail and robustness (Table 3, App. B). The null sets the level, not the ranking: the uncalibrated gain ranks the knob identically ($\rho = 0.981$ in both families) but sits 0.012 (F1) and 0.022 bits (F2) above zero at f=0, and the null mean removes that offset at every f. The post-hoc extension to seeds 5–24 also covered F2. There the control gain again gives $\rho = 0.980$ (fresh and pooled) and plain SI stays flat ($\rho = -0.116$, $p = 0.25$ fresh; -0.095, $p = 0.29$ pooled over 125 runs), but effective rank's drift becomes significant ($\rho = 0.250$, $p = 0.013$ fresh; 0.267, $p = 0.004$ pooled): the drift is real but inside the separation margin ($0.267 \leq 0.980 - 0.3$).

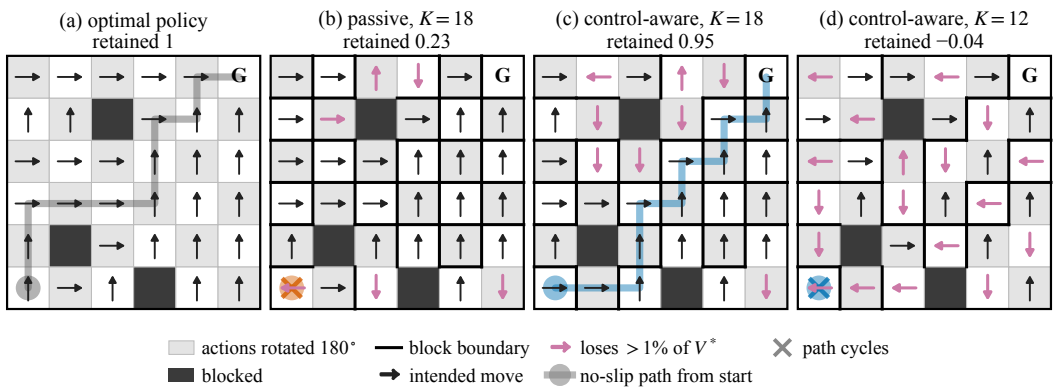


Figure 4: Case study: F3 seed 3, the one instance in which any objective retains more than 0.11 of optimal control (Table 2); chosen because it contains both the success and a failure of the control-aware objective, not because it is typical. Grey cells have the 180° action rotation; thick lines separate blocks of the stored partition. Arrows give the direction each state's lifted action intends to move (optimal policy in (a)); an arrow is marked when that action costs more than 1% of $V^*(\text{start})$ under Q^*. The shaded line follows the lifted policy from the start cell (bottom left) without slip, to the goal G or until it revisits a cell. Policies are recomputed from the stored partitions with the code of the experiment and reproduce the logged returns exactly.

5.2 Quotient planning (G4)

For each F3 instance we draw one uniform rollout (N=40,000), and from the *same counts* extract fixed-K partitions by greedy node-move minimisation of L_{pass} (plain) versus L_{ctrl} (ours), $K \in \{18, 12, 6\}$, identical optimiser and budget. Value iteration on the quotient MDP, lifted block policy, evaluated exactly on the true kernel; we report retained control $(J - J_{\text{rand}})/(J^* - J_{\text{rand}})$.

Gate verdict: G4 FAIL (reported verbatim). The control-aware partition beat the plain partition in only 2/5, 2/5, 1/5 seeds at $K = 18, 12, 6$ (1/5 each without the two seed-0 ties); the median gap at K=12 was -0.001 against the preregistered ≥ 0.25. In absolute terms *both* arms collapse to the random-policy baseline in almost every cell (retained within $[-0.05, 0.23]$), with a single striking exception: the control-aware partition at K=18, seed 3 retained 0.952 of optimal return (Table 2).

Case study (Fig. 4). What decides retained control is the lifted action on the few cells between start and goal, not the partition as a whole. The control-aware partition at K=18 (c) still assigns a costly action to 8 cells, but none lies on the path, which reaches the goal: retained 0.95. The passive partition (b) puts the start cell in one block with its opposite-parity neighbour; the block action, right for the neighbour, drives the start cell into the wall, and only slip moves the agent on (retained 0.23). The same control-aware objective at K=12 (d) merges start with goal, the reward-blind failure described next; 18 cells lose value, the start cell among them (retained -0.04).

Post-hoc analysis (clearly labelled). Four findings explain it:

1. *Reward-blindness merges goal with start.* The goal teleports to the start, so to any reward-free code the two states are dynamically similar; on seed 0 both objectives placed goal and start in one block. Minimally splitting the goal into its own block repaired only the plain arm on that seed ($-0.035 \rightarrow 0.113$) — a partial mechanism.
2. *Even the hand-built "ideal" partition fails.* The region×parity partition at K=18 retains -0.009: with a 180° inversion, a lifted block policy whose action is wrong for a minority of member states does not merely lose value — it reverses drift on the path, and return collapses to the random baseline. Retained return in this family is nearly all-or-nothing in the fidelity of per-state actions.
3. *Profile consistency does not rescue it.* A third arm minimising block-level action-profile inconsistency (an approximate-MDP-homomorphism objective, same optimiser and budget) retained between -0.044 and -0.019 in all 15 cells. No reward-free objective we tested selects plannable quotients here.

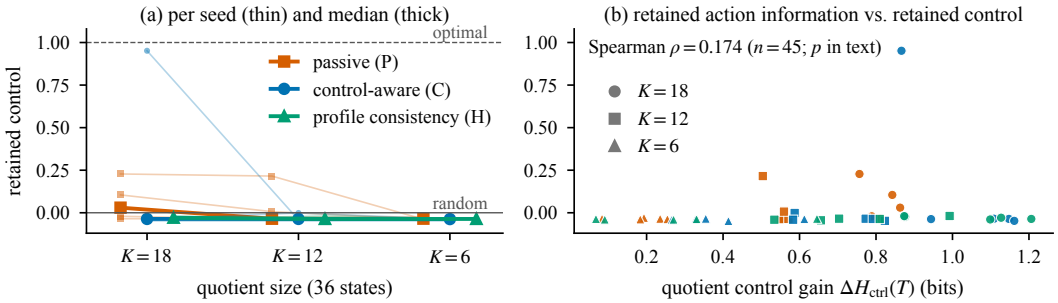


Figure 5: Quotient planning on F3, all 45 partitions. (a) Retained control per seed (thin) and median over 5 seeds (thick) for the three reward-free objectives; 0 is the uniform-random policy, 1 the optimal one. The single high point is C at K=18, seed 3 (Fig. 4c). (b) Quotient control gain against retained control; colour gives the objective as in (a), marker shape the quotient size.

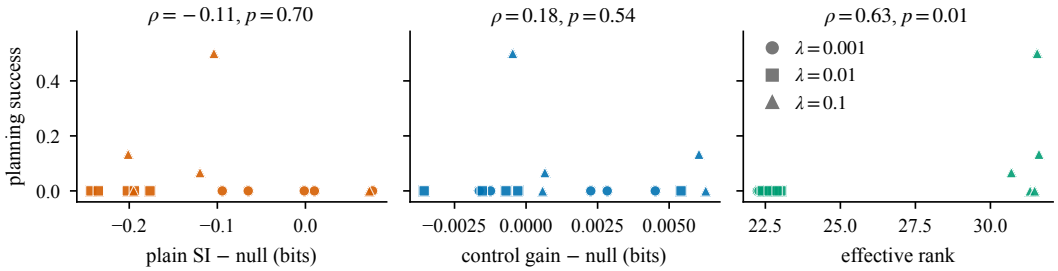


Figure 6: Exploratory check on the 15 TwoRooms agents (G6): planning success against null-calibrated plain SI, null-calibrated control gain and effective rank, with Spearman ρ and permutation p over the 15 agents. Marker shape gives the regulariser weight λ: every successful agent is in the λ=0.1 arm, which also has the highest effective rank.

4. *Plan-preserving partitions exist.* The seed-3 control-aware solution (0.952) is an existence proof; its quotient control gain (0.87 bits) was *lower* than that of the failing control-aware partitions at the same K (0.94–1.16 bits): maximal retained action information is not the same thing as value-coherent structure, and can trade off against it. The control-aware optimiser does reach its own target: its partitions keep more quotient control gain than the plain ones in all 15 cells. But across the 45 partitions of Table 2, quotient control gain and retained control are unrelated (Spearman $\rho = 0.174$, $p = 0.26$; Fig. 5b).

G4 therefore separates two roles: the control gain *measures* the steerable dynamics (Sec. 5.1), but none of the three reward-free partition objectives preserved planning under aggressive abstraction in this adversarially control-sensitive environment. Whether reward-aware abstraction (the ingredient bisimulation metrics add) closes the gap is untested here.

5.3 Exploratory check on 15 learned agents (G6)

G6 trains 15 joint-embedding predictive (JEPA) agents on a continuous TwoRooms task (uniformity regulariser, $\lambda \in \{0.001, 0.01, 0.1\}\times$ seeds 1–5, 16 nuisance dimensions, offline) and scores planning *success*, not planning return. The preregistration calls them re-trained copies of an earlier population; we claim no match to it and analyse them on their own. We log actions during a 20,000-step random rollout, discretise latents to M=40 k-means prototypes, bin the 2-D continuous actions into sign quadrants, and compute both null-calibrated measures on the same prototype process.

Verdict (exploratory; both preregistered predictions hold in sign, neither significantly). Across the 15 agents the null-calibrated control gain correlates with planning success at Spearman $\rho = 0.176$ ($p = 0.54$), and null-calibrated plain SI at $\rho = -0.110$ ($p = 0.70$), so prediction (i), $\rho > 0$, and prediction (ii), control gain above plain SI, both hold in sign (Fig. 6, left and middle). Neither correlation is distinguishable from zero at n=15, and planning success is nonzero for only 3 of the

15 agents, so the ranking rests on few points. Effective rank correlates at $\rho = 0.628$ ($p = 0.011$) on these agents. All three successful agents are in the λ=0.1 arm, which also has the highest effective rank (30.7–31.6, against 22.2–23.0 in the other two arms); with that arm left out no agent succeeds, so every correlation here, effective rank's included, contrasts one arm with the rest rather than tracking a graded trend. The uncalibrated control gain correlates at $\rho = 0.411$ ($p = 0.13$), but the null absorbs almost all of it: raw gains span 0.182–0.197 bits, null means 0.176–0.191, calibrated values -0.004 to 0.006, above twice the null s.d. for 2 of the 15. We read G6 as no evidence either way for the control gain, not as corroboration.

6 Related work

Codelengths of walks. SE (Li and Pan, 2016) and the map equation (Rosvall and Bergstrom, 2008) both score a partition by the description length of a random walk, so both are passive in the sense of Prop. 1. SE has been used to build hierarchical state abstractions for RL from an encoding tree (Zeng et al., 2023); our result bounds what any such tree can see when it is scored by walk codelength alone. SI2E (Zeng et al., 2024) brings actions into SE through a structural mutual information over state–action pairs grouped by policy value, as an exploration bonus; the control gain is reward- and policy-free and scores a partition against an exact null. Degree-preserving rewiring (Maslov and Sneppen, 2002) is the standard null for graph statistics; the within-state action shuffle is its counterpart for the action channel. **State abstraction.** Model minimisation and bisimulation (Givan et al., 2003; Ferns et al., 2004), homomorphisms (Ravindran and Barto, 2004) and the abstraction hierarchy of Li et al. (2006) characterise aggregations that preserve values or optimal policies, and all use reward; deep variants learn such representations with reward and transition losses (Gelada et al., 2019; Zhang et al., 2021). Reward-free multi-step inverse models provably recover the controllable latent state (Lamb et al., 2023); they learn an encoder, whereas the control gain measures, in bits, how much control a given partition keeps. The control gain is reward-free, and G4 shows the cost: no reward-free objective we tested preserved planning on F3. **Empowerment and intrinsic control.** Empowerment (Klyubin et al., 2005; Salge et al., 2014) is the channel capacity from actions to future states; variational estimators make it an intrinsic objective for training policies (Mohamed and Rezende, 2015; Gregor et al., 2016). We use the uniform-prior information as a property of a partition, with an exact null, not as a reward. **Representation diagnostics.** The rank of learned features is used in deep RL to diagnose lost expressivity (Kumar et al., 2021); in F1 effective rank is blind by construction.

7 Limitations

Tabular validators only; the continuous case inherits the discretisation problem wholesale, and our TwoRooms check adopts a k-means prototype discretisation as a convention rather than solving it. The action binning for continuous actions is a convention with the same status. ΔH_{ctrl} is defined under the uniform behaviour policy; policy-dependent variants (e.g. the agent's own policy mix) are enumerated in the appendix but not validated. The greedy fixed-K optimiser carries no optimality guarantee; both arms share it, so comparisons are fair, but absolute codelengths are upper bounds. The G6 agents are scored by planning success, not return; n=15 is exploratory, and their planning success varies in one arm only. Reward-aware arms (one-step bisimulation, Q^*-consistency) were specified in the preregistration after G4 closed but are not reported here, so whether reward-awareness is needed on F3 remains open. One-step gain only; the k-step spectrum is future work.

8 Conclusion

For SE, blindness to control is a theorem, not an accident, and the minimal repair is to let the code condition on actions. The resulting control gain is uniform-prior empowerment written as a codelength saving; it comes with an exact null and, in tabular MDPs, recovers a controllability knob that no passive measure can see. It does not, on its own, select state abstractions that preserve the ability to act: in our checkerboard-inversion test no reward-free partition objective did, although plan-preserving partitions exist. What it supplies is a null-calibrated measurement, in bits, of the action channel a representation keeps.

REPRODUCIBILITY STATEMENT

All formal claims are stated with their assumptions in the main text and proved either there or in the appendix. Every experiment reported here is a closed specification: the constructions, their parameters, the seed set, the null model and its draw count, and the acceptance criteria are given in the experimental sections, and the criteria were fixed before the decisive runs. Code sufficient to regenerate every number and figure, together with the raw logged outputs and the preregistration file recording each criterion and its verdict, is included as anonymous supplementary material.

AI USE STATEMENT

We used generative AI tools on the following tasks from ICLR's required-disclosure list: helping to develop the conceptual framework, helping to formulate mathematical claims, providing ingredients for and assisting in the writing of proofs, proposing and refining hypotheses, giving feedback on experimental design, implementing methods, and helping to interpret results. We did *not* use generative AI tools to generate synthetic data sets reported as measurements, to clean or reformat data sets, for translation, or for qualitative or thematic data analysis. From the recommended-disclosure list we additionally used generative AI tools to draft and edit parts of this paper, to create and edit figures, to identify and summarise related literature, to format references, and to suggest a structure for the paper.

We have reviewed all AI-assisted work. Every theorem statement and proof appearing in this paper was checked line by line by the authors; every numerical value reported was regenerated from the released code and compared against the logged experimental outputs; every preregistered criterion and its verdict was recorded before the corresponding run and is reported verbatim, including the failures. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

A. Li and Y. Pan. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339, 2016.

A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: A universal agent-centric measure of control. *IEEE CEC*, 2005.

C. Salge, C. Glackin, and D. Polani. Empowerment: An introduction. *Guided Self-Organization: Inception*, Springer, 2014.

N. Ferns, P. Panangaden, and D. Precup. Metrics for finite Markov decision processes. *UAI*, 2004.

B. Ravindran and A. G. Barto. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes. *KBCS*, 2004.

P. Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

X. Zeng, H. Peng, A. Li, C. Liu, L. He, and P. S. Yu. Hierarchical state abstraction based on structural information principles. In *IJCAI*, pp. 4549–4557, 2023.

X. Zeng, H. Peng, and A. Li. Effective exploration based on the structural information principles. In *NeurIPS*, 2024. arXiv:2410.06621.

A. Lamb, R. Islam, Y. Efroni, A. Didolkar, D. Misra, D. Foster, O. Ogunmolu, R. Chari, A. Krishnamurthy, and J. Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *TMLR*, 2023. arXiv:2207.08229.

S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In *ISAIM*, 2006.

C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *ICML*, 2019. arXiv:1906.02736.

A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. In *ICLR*, 2021. arXiv:2006.10742.

S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *NeurIPS*, 2015. arXiv:1509.08731.

K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. arXiv:1611.07507, 2016.

A. Kumar, R. Agarwal, D. Ghosh, and S. Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *ICLR*, 2021. arXiv:2010.14498.

A Candidate definitions considered

Three candidate objects suggest themselves for a control-aware SE. (i) *Policy-mix SE*: SE under the agent's policy versus uniform actions — rejected as the headline because it measures a policy, not a representation, and Prop. 1 shows each side separately is still passive. (ii) *Empowerment-weighted SE*: edge weights $\max_a P(s' \mid s, a)$ — rejected because the object is no longer a normalised walk and admits no matched null that separates the action channel from connectivity ($\max$ couples them). (iii) *Control gain*: adopted, generalised from the entropy-rate difference to a per-resolution quantity $\Delta H_{\mathrm{ctrl}}(T)$ so that it composes with SE's encoding-tree machinery and yields a partition objective. The action-shuffle null applies unchanged at every resolution.

B Per-knob values of the knob families

Table 3: Knob families, seeds 0–4. True: exact kernel gain (bits; seed mean on F2). Raw: plug-in $\widehat{\Delta H_{\mathrm{ctrl}}}$. Null: mean of 10 within-state action shuffles. Cal.: raw $-$ null, mean $\pm$ sample s.d. over the 5 seeds. $\mathrm{SI}^{\mathrm{cal}}$ and effective rank: mean $\pm$ s.d. likewise.

	f	True	Raw	Null	Cal.	$\mathrm{SI}^{\mathrm{cal}}$	Eff. rank
F1	0	0	0.012	0.012	-0.000 ± 0.001	0.544 ± 0.031	24.11 ± 0.14
	0.25	0.5	0.509	0.012	0.498 ± 0.011	0.566 ± 0.106	24.09 ± 0.11
	0.5	1	1.000	0.012	0.988 ± 0.009	0.578 ± 0.050	24.06 ± 0.15
	0.75	1.5	1.498	0.012	1.487 ± 0.002	0.581 ± 0.064	24.20 ± 0.14
	1	2	1.996	0.012	1.985 ± 0.001	0.585 ± 0.031	24.07 ± 0.11
F2	0	0	0.022	0.022	-0.000 ± 0.002	0.275 ± 0.043	29.82 ± 0.69
	0.25	0.070	0.094	0.022	0.072 ± 0.006	0.280 ± 0.043	30.63 ± 0.45
	0.5	0.219	0.241	0.022	0.219 ± 0.013	0.252 ± 0.026	30.82 ± 0.53
	0.75	0.438	0.464	0.022	0.441 ± 0.024	0.254 ± 0.032	30.79 ± 0.42
	1	0.791	0.803	0.022	0.781 ± 0.032	0.292 ± 0.043	30.59 ± 0.47