WHETHER VS. WHEN: SEPARATING SHORTCUT FORMATION FROM EVIDENCE-TIMING SENSITIVITY IN VALUE-BASED ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep reinforcement learning agents can overweight early experience, a phenomenon studied as primacy bias, but whether a shortcut forms and how strongly its formation depends on the timing of disambiguating evidence are distinct properties that this framing can conflate. We separate them in a controlled visual task where a salient, spurious feature and a subtler, reward-relevant feature agree except at planted evidence cells. Holding the content and quantity of evidence fixed, we relocate the same 40 evidence-cell visits from the first to the last 400 steps of a 4000-step value-based online Double DQN training window. This temporal shift changes shortcut formation from 0 of 30 to 30 of 30 paired seeds, with a mean paired margin difference of -6.17 (95% CI $[-6.41, -5.93]$). The split survives interventions targeting action coverage, exploration, evidence dose, integration time, replay-sampling exposure, and optimizer-update count, although inward placement and closer matching of integration time attenuate it. The effect also persists across a smaller vision transformer, a structurally distinct corridor environment, and fully organic, action-driven visitation (0/30 vs. 30/30), though realized exposure there is endogenous and unequal rather than matched. In contrast, matched fixed-dataset supervised and bootstrapped-TD learners show substantially smaller timing effects, while a matched on-policy PPO agent shows no timing-dependent split despite forming the same shortcut. Removing experience replay also closes the split, but batch size and gradient variance change simultaneously, so this does not identify replay as the unique cause. A rewind intervention further points to the role of accumulated network state in how evidence is incorporated. Together, these results show that shortcut formation is not unique to online reinforcement learning, but its dependence on evidence timing is substantially amplified in the value-based online-RL regime studied here, while leaving the precise mechanism unresolved.

1 INTRODUCTION

A reinforcement learning agent's behavior can come to depend on an easy, spurious feature rather than a harder, task-relevant one: a shortcut. This raises two properties that are easy to conflate but are not the same. *Formation* is whether such a shortcut becomes the agent's behavioral rule at all. *Timing sensitivity* is how strongly that formation depends on *when*, during training, disambiguating evidence about the task-relevant feature arrives, holding the content and quantity of that evidence fixed. A system in which a shortcut forms only when disambiguating evidence arrives late is timing-sensitive in a way that a system in which the same shortcut forms regardless of evidence order is not, even though both systems form the shortcut. This paper keeps these two properties lexically and empirically distinct throughout.

The anchor result is stark. Holding a fixed budget of 40 evidence-cell visits and a fixed 4000-step training window, relocating those visits from positions $[0, 400)$ to $[3600, 4000)$ changes an online Double DQN agent's shortcut formation from 0 of 30 seeds to 30 of 30. This is not a rediscovery of primacy bias (Nikishin et al., 2022): content, count, and total training length are held fixed, and relocating the evidence produces a categorical reversal, though, as we detail next, it necessarily changes several downstream aspects of training besides evidence position itself.

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 construction this sensitive to timing invites an objection: the extreme-window anchor bundles evidence position with several other differences between EARLY and LATE: annealed ϵ-greedy exposes the two orderings to different action-conditioned data, LATE's evidence closes exactly when the formation window ends, and the exploration schedule differs across the orderings' entire trajectories, not only at the evidence steps. We test each directly: forcing identical action coverage, closely approximating integration time, and flattening the schedule. The split survives the first; the latter two each attenuate the raw contrast substantially ($\approx$36% and $\approx$24%) but leave a large residual, so none individually explains the effect.

Prior work on primacy bias does not hold a fixed evidence budget constant while varying only timing, nor compare the resulting sensitivity against a matched non-RL regime. We build a task where a grid cell carries a salient feature A and a subtler feature B that agree except at a small, planted set of cells, so a later shift makes reliance on A harmful, and present the identical evidence sequence to an online RL agent, a fixed-dataset classifier, and a fixed-dataset bootstrapped-TD scaffold. The pattern is a dissociation: online RL shows the categorical split above, while the fixed-dataset controls form the shortcut in a majority of seeds under both orderings with far less ordering effect, confirmed by a direct paired interaction test (by formation-rate gap, max 6.7 points vs. 50 + at every value-based RL point measured). The dissociation reproduces under a small vision transformer ($\approx$15% of the parameter count) and a structurally distinct one-dimensional environment (categorical for TinyViT, large but not fully categorical for Corridor).

Together, these results show that shortcut formation is not unique to online RL, but the size of its dependence on evidence timing is dramatically larger under the value-based online-RL protocol we test than in the matched fixed-dataset settings, and is absent under a tested on-policy alternative. Our contributions:

1. **Conceptual separation.** We distinguish shortcut formation from its sensitivity to evidence timing and introduce an evidence-ordering protocol that reports formation and timing sensitivity separately.
2. **Cross-paradigm dissociation.** Across matched evidence sequences, the value-based online-RL protocol shows a dramatically stronger timing dependence, by formation-rate gap, than fixed-dataset supervised and isolated bootstrapped-TD learning, confirmed by paired interaction tests and a targeted sweep stress-testing the supervised operating point; this is a protocol-level comparison, not an attribution to a single ingredient.
3. **Robustness and scope.** The dissociation reproduces across a different architecture, a structurally distinct environment, and fully organic action-driven visitation, and is unchanged under a $2\times$ dose reduction, but attenuates when the evidence windows move inward; replay exposure and raw update count are each ruled out as sufficient alone, and removing replay entirely (confounded with batch size) eliminates it. A matched on-policy PPO agent does not reproduce the dependence, delimiting the phenomenon to the value-based online-RL regime tested rather than to online interaction generally.
4. **Diagnosing bundled confounds.** The extreme-window anchor bundles evidence position with action-coverage asymmetry, integration time, and exploration schedule; individually intervening on each attenuates the gap substantially, but none closes the split, leaving a large residual most consistent with path dependence rather than any single tested ingredient.

2 RELATED WORK

Primacy bias, plasticity, and online-RL nonstationarity. Nikishin et al. (2022) show deep RL agents disproportionately fit early interactions in a way that damages later learning, mitigated by periodically resetting later layers; follow-up work links this to update-to-data ratios (D'Oro et al., 2023; Fedus et al., 2020) and to value-function divergence rather than early-data overfitting (Hussing et al., 2024). Achille et al. (2019) show early-training input deficits cause a lasting deficit; Dohare et al. (2024) document longer-horizon plasticity loss under continual training. None holds a fixed evidence quantity constant while varying only timing, or compares the sensitivity to a matched non-RL regime.

Shortcut learning and data ordering. Geirhos et al. (2020) document that supervised SGD readily exploits salient, easy features over harder, more predictive ones; Pezeshki et al. (2021) and Puli et al.

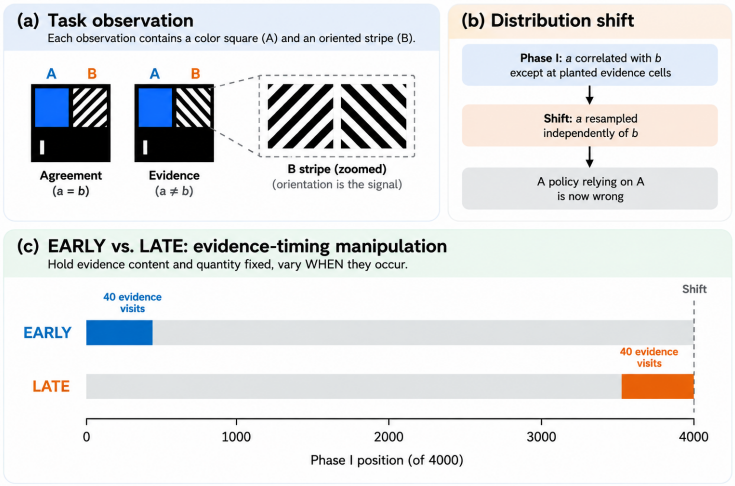

Figure 1: Task and evidence-timing manipulation. (a) Feature A (salient color) and B (subtler diagonal stripe) at fixed regions, with a magnified inset of the stripe-orientation difference. (b) A later shift resamples A independently of B. (c) The same 40 evidence-cell visits occupy the first or last 400 of 4000 positions, holding evidence content and quantity fixed while varying WHEN; relocating evidence also changes downstream training, not only ordering (Section 1).

(2023) trace this to gradient dynamics under cross-entropy alone, and curriculum learning (Bengio et al., 2009) shows presentation order on a fixed set changes what a supervised learner converges to. This establishes that shortcut *formation* is not RL-specific, consistent with our fixed-dataset controls; we add that formation's *timing sensitivity* differs sharply between paradigms by formation-rate gap, surviving a targeted sweep stress-testing the supervised operating point.

Replay, update ratios, and positioning. Higher update-to-data ratios can destabilize training (van Hasselt et al., 2018); offline RL studies a fixed, pre-collected dataset with no further environment interaction (Levine et al., 2020). Our TD control isolates bootstrapping alone on such a dataset; a null result speaks to this construction's sufficiency, not bootstrapping's role in general. Prior work has not, to our knowledge, compared a fixed, temporally-relocated evidence budget across online RL, matched fixed-dataset learning, and a structurally distinct environment, nor tested whether the sensitivity survives intervention on factors that co-vary with evidence position (action coverage, integration time, exploration schedule, replay itself). Our contribution is this conceptual separation and controlled protocol, not a new primacy-bias, shortcut-learning, or architecture result.

3 Controlled Framework

3.1 Task and evidence-timing intervention

The environment is a 6×6 grid; each cell renders two binary features (Figure 1a): A, a large high-contrast color block, and B, a diagonal-stripe texture whose orientation alone determines reward. During an initial Phase I of $L{=}4000$ positions, A equals B everywhere except a fixed, planted set of 9 cells; at the end of Phase I, a shift resamples A independently of B over the whole grid and holds it fixed thereafter, so a policy relying on A becomes wrong. An example is *evidence* if its cell has $A{\neq}B$, and *agreement* otherwise.

The evidence-timing manipulation holds the evidence budget fixed at 40 evidence-cell visits and L fixed, varying only *where* those visits occur: Early places them in $[0, 400)$, Late in $[3600, 4000)$, with the remaining 3960 positions drawn from the agreement pool. These 40 visits revisit a fixed, per-seed-randomized pool of 9 unique cells, identical between orderings for a given seed, so the conditions differ in *when* a cell is visited, not *which* cells exist. Every paired seed is verified to share an identical content hash, a different order hash, exactly 40 visits, and identical label counts (Appendix J). The anchor uses a corrected construction with dedicated evidence/agreement RNG

streams (exact content-hash equality, all 30 seeds); a shared-RNG issue in the original version is detailed in Appendix G.

3.2 BEHAVIORAL RELIANCE AND THE FORMATION CRITERION

At each checkpoint (every 200 steps), four synthetic probes spanning $(A, B) \in \{0, 1\}^2$ give the harvest margin (the difference between the two harvest-action Q-values or logits); $\mathrm{Rel}(A)$ and $\mathrm{Rel}(B)$ are the finite-difference effect on this margin of flipping each feature with the other fixed. Initialization probes show near-zero, effectively symmetric margins ($\mathrm{Rel}(A) \approx 0.002$, $\mathrm{Rel}(B) \approx -0.003$, three orders of magnitude below trained margins), so the formation threshold is not explained by an intrinsic probe offset (Appendix N). The continuous quantity $\mathrm{Rel}(A) - \mathrm{Rel}(B)$ is our primary behavioral measurement, defined identically across every paradigm and architecture and carrying the statistical weight of our conclusions. **Formation** is a compact binary summary, mean($\mathrm{Rel}(A)$, last 5 pre-shift checkpoints) $>$ mean($\mathrm{Rel}(B), \ldots$), measured strictly before the shift so it reflects whether the shortcut formed, not whether it can later be corrected. For the LATE anchor condition, two of these five checkpoints (steps 3600, 3800) fall inside LATE's own evidence window ($[3600, 4000)$); we retain this definition uniformly rather than special-casing LATE, but this overlap means LATE's arm is not measured in a period cleanly insulated from ongoing evidence exposure, which is one reason the integration-time analysis (Section 4.3) is necessary rather than assumed. As a checkpoint-sensitivity check, the categorical split is insensitive to this choice of window: using the single final pre-shift checkpoint alone gives 0/30 vs. 28/30, and the final two checkpoints give 0/30 vs. 30/30 (Appendix S).

3.3 LEARNING PARADIGMS

The identical evidence sequence is presented to the three core paradigms used for the cross-paradigm comparison, holding architecture fixed and using the same Adam optimizer with paradigm-appropriate learning rates (Appendix L); PPO is added separately as a boundary condition (Appendix O). Update count is not held fixed and is itself manipulated in Section 4.2. This is a protocol-level comparison rather than a one-component-at-a-time ablation: replacing online RL changes data generation, action-reward coupling, replay, bootstrapping, and update dynamics all at once, so the test is whether timing sensitivity survives removing the online loop, not whether it can be pinned on one ingredient.

The **online RL** agent (CNN, 154,214 parameters, Double DQN, target network, Huber loss) has live epsilon-greedy action selection and a uniform-sampling replay buffer; the evidence-timing controller determines only which cell the agent occupies, never which action it takes or what reward it receives. An earlier exact-trajectory yoking check offered weak, indirect support for this scripted-visitation construction (Appendix A); Section 4.3 instead reports a direct, confirmatory test with fully organic, action-driven movement. The **fixed-dataset supervised** control removes the RL loop entirely, training $K{=}4$ deterministic passes by cross-entropy with no environment, reward, or replay. The **fixed-dataset TD** control asks whether bootstrapping alone recovers the online timing dependence, adding a bootstrapped Double DQN target on a fixed, precomputed, two-sided self-loop transition set ($s'{=}s$, never terminal), with no replay or live data generation; this isolates a single, self-referential form of bootstrapping, not multi-step credit assignment or bootstrapping combined with replay. Full architectures, hyperparameters, and the invalidated one-sided TD-v1 pilot are in Appendices H–L.

To test whether the dissociation depends on grid topology, image observations, or visual-frequency salience, we built **Corridor Foraging**: a one-dimensional, 12-station corridor with 6-dimensional vector observations and an MLP agent, where A is a linearly-decodable one-hot pair and B is the XOR of two bits redrawn on every observation. Evidence budget, architecture, and paradigms are scaled analogously (Appendix K).

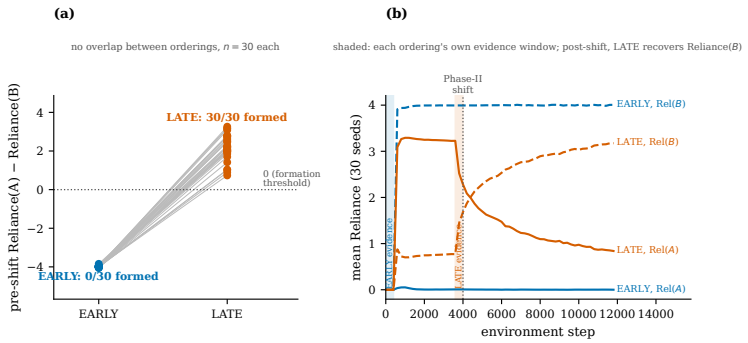

Figure 2: Corrected online-RL anchor (CNN, n=30). (a) Paired per-seed pre-shift margin, no overlap, binary formation counts annotated directly. (b) Mean Reliance trajectories, evidence windows shaded, Phase-II shift marked; LATE shows increased Reliance(B) only after the shift.

4 RESULTS

4.1 EVIDENCE TIMING DECISIVELY ALTERS ONLINE-RL FORMATION

With total Phase-I duration and evidence count fixed, the position of the 40 evidence-cell visits within Phase I produces a categorical difference in formation. Figure 2a shows the pre-shift $(\mathrm{Rel}(A) - \mathrm{Rel}(B))$ margin per seed: every EARLY seed sits at a large negative value and every LATE seed at a positive one, with no overlap, so the binary formation criterion's placement at zero is not a fragile choice near either distribution's edge (exact per-arm intervals in Appendix M). This gives Wilcoxon $p = 1.86 \times 10^{-9}$ (the minimum attainable exact value at n=30 with unanimous sign agreement, recurring throughout not by coincidence), Cohen's $d_z = -9.18$, mean paired difference -6.17, 95% CI $[-6.41, -5.93]$. Across 30 paired seeds, EARLY yields 0/30 formed shortcuts and LATE yields 30/30 (McNemar $p = 1.86 \times 10^{-9}$; Figure 2a). Figure 2b shows LATE seeds do recover B-reliance after the shift: the timing effect governs whether the shortcut *forms* during the evidence window, not whether it can later be corrected. Formation is behaviorally consequential: LATE takes an immediate post-shift reward hit relative to EARLY (21.0 vs. 32.0, $p = 1.7 \times 10^{-6}$) and remains below EARLY even after 8000 post-shift steps (30.7 vs. 32.0, $p = 1.3 \times 10^{-4}$), though 28/30 LATE seeds recover B-reliance within that window. Pooled across all 60 anchor seeds, the continuous pre-shift margin correlates strongly with this cost (Spearman $\rho = -0.81$, $n = 60$, $p = 3.9 \times 10^{-15}$), a predominantly between-condition association rather than an independently powered within-condition test.

A shared-RNG construction issue in the original run caused a small EARLY/LATE content-multiset mismatch; re-running with dedicated evidence/agreement RNG streams restored exact content-hash equality for all 30 seeds and reproduced the result essentially unchanged ($d_z = -9.18$ vs. the original run's -9.77), robust to the correction (Appendix G).

This raw contrast bundles evidence position with action-conditioned data, integration time, and exploration schedule (Section 1); Figure 3 previews four of the controls tested in Section 4.3.

4.2 FORMATION PERSISTS, BUT LARGE TIMING SENSITIVITY IS NOT REPRODUCED IN FIXED-DATASET CONTROLS

Transplanting the identical evidence sequence onto a fixed-dataset classifier, with no environment, replay, or reward, still forms the shortcut readily: 56.7% (17/30) EARLY vs. 60.0% (18/30) LATE (McNemar p=1.0; continuous margin d_z=0.078, 95% CI $[-0.092, 0.144]$, excluding an RL-sized effect but not an exactly zero one). A single pass (K=1) saturates both orderings at ceiling, not a masked larger effect. A targeted 3×3 learning-rate×pass-count sweep stress-testing this CNN grid-task supervised operating point (9 cells, 540 runs, not exhaustive; full record in Appendix E) finds no cell with a value-based-RL-sized effect: the largest formation-rate gap is 6.7 points (K=2, lr= 3×10^{-4}), far below every value-based RL gap (50+); at a different cell, the largest $|d_z|$ is 1.75, narrowing under RL's most-attenuated forms (2.13–2.80; Table 7). This 6.7-point figure is

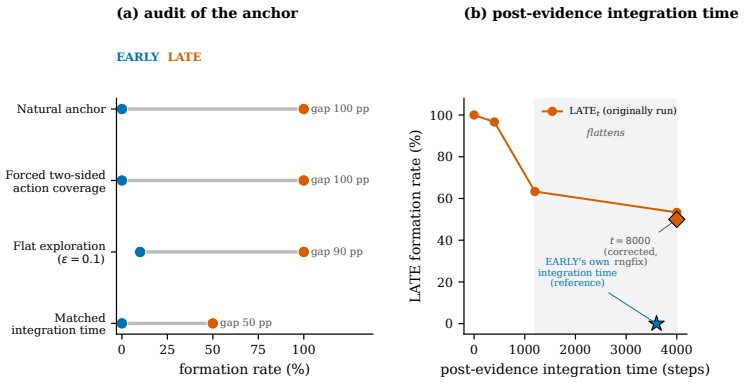

Figure 3: Auditing the anchor (CNN, n=30 per arm). (a) EARLY/LATE formation rate, natural anchor and three interventional controls, not a fully-crossed decomposition; attenuations are not additive. (b) LATE formation rate across the full integration-time lead-time family, dropping sharply to 1200 steps and flattening thereafter; EARLY's own integration time/formation is a reference point, not a fifth condition; the RNG-corrected t=8000 point is distinguished from the originally-run family.

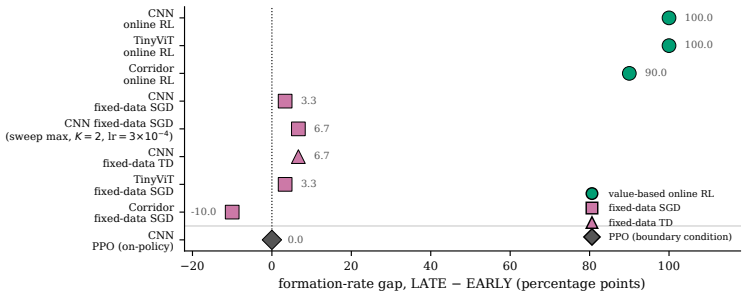

Figure 4: Cross-paradigm and cross-environment dissociation, formation-rate gap, n=30 per condition, marker shape encoding paradigm: every value-based online-RL point is at 90+ points; every fixed-dataset control is within $\pm$10, including the CNN sweep's own largest-gap cell (6.7, distinct from Corridor's larger, oppositely-signed -10); PPO (boundary condition, not another value-based replication) sits at 0. The corresponding direct paired interaction test on the continuous margin is supporting evidence, reported numerically above and in Appendix A, Table 3, rather than as a separate panel.

CNN-grid-task-specific: Corridor's supervised gap is larger and oppositely signed (10 points; Section 4.3). An isolated self-loop Double-DQN-bootstrap control likewise fails (80.0% vs. 86.7%, gap 6.7 points, $d_z = -0.27$) and does not generalize to bootstrapping with replay, live data, or multi-step credit assignment.

The seed-paired bootstrap on this gap-of-gaps excludes zero by a wide margin against both supervised ($[87, 107]$ points) and TD ($[83, 100]$ points); a direct paired interaction test on the continuous margin agrees (all four architecture/environment pairings $p = 1.86 \times 10^{-9}$; full means and CIs in Appendix A, Table 3), though its raw scale differs across paradigms, so we rely on the gap bootstrap as primary.

B remains highly decodable from the supervised control's features even where behavior relies on A, so shortcut reliance need not reflect loss of the unused feature from the representation (Appendix N).

Replay exposure and update count are insufficient explanations. EARLY's evidence is resampled $\approx$40$\times$ more often than LATE's before the shift; boosting LATE's exposure to EARLY's level did not rescue a single seed (Holm-p=1.0), while throttling EARLY's down to LATE's level raised formation from 0/30 to 10/30 (Holm-p=0.0059), a real but partial contribution. Matching raw update count (4804) across conditions reached via different elapsed experience still left formation sharply

Table 1: Interventional controls on the anchor (CNN online RL, n=30 paired seeds each row). Gap is LATE minus EARLY formation rate, percentage points; d_z is the paired standardized effect size on the continuous margin. Full CIs and p-values: Appendix A, Table 3.

Condition	EARLY	LATE	Gap	d_z
Natural anchor	0/30	30/30	100.0	-9.18
Forced two-sided action coverage	0/30	30/30	100.0	-6.48
20-event dose (half budget)	0/30	30/30	100.0	-8.59
Flat exploration (ϵ=0.1)	3/30	30/30	90.0	-2.13
Interior window	0/30	22/30	73.3	-5.13
Matched integration time (t=8000)	0/30	15/30	50.0	-2.80
No-replay (batch size 1)	29/30	29/30	0.0	-0.22

different (6.7% vs. 90.0%, Holm-p=1.25×10^{-6}, d_z= $-$ 1.39), and a shift-timing probe rules out LATE evidence being simply too close to the shift to matter (Cochran's Q=34.38, p=1.6×10^{-7}; Appendix B). Removing replay entirely, in a batch-size-1 online update that otherwise preserves the Double-DQN target, optimizer, and evidence construction, does eliminate the split: EARLY and LATE each form in 29/30 seeds (McNemar p=1.00; margins 2.694 vs. 2.490, mean paired diff. -0.204, 95% CI $[-0.542, 0.135]$, d_z= $-$ 0.22; Appendix Q) – but batch size and gradient variance change simultaneously, so this is a replay/batch-size ablation, not evidence that replay alone is causal.

4.3 GENERALITY AND ROBUSTNESS

Organic, action-driven visitation (Appendix P). Every result above uses a scripted controller that places the agent regardless of its chosen action. We re-ran the anchor with the controller removed entirely (`env.yoked_controller=None`): movement is fully action-driven, and only the temporal availability of the same 9 planted cells is gated to EARLY's or LATE's window. The categorical split reproduces (0/30 vs. 30/30, n=30, McNemar p=1.86×10^{-9}; margins -3.986 vs. 2.378, mean paired diff. 6.364, 95% CI $[6.095, 6.633]$, d_z=8.47). Visitation is now endogenous, so realized exposure is unequal, not matched (LATE's median 26 visits vs. EARLY's 99.5; 8/30 LATE seeds see zero); this is not an artifact driving the result, since the 22 nonzero-exposure LATE seeds also all form (22/22), indistinguishable in margin from the 8 zero-exposure seeds (Mann-Whitney p=0.50).

Table 3 reports the full comparison (Appendix A). Under TinyViT (≈15% of the CNN's parameters), online RL again splits categorically (0/30 vs. 30/30) while its supervised control shows a small, non-significant difference (53.3% vs. 56.7%); by gap, TinyViT's RL split (100 points) dwarfs its supervised control's (3.3), interaction test confirming (p=1.86×10^{-9}). Corridor Foraging (different topology, modality, and salience) shows the same qualitative pattern at smaller magnitude: RL splits 3/30 vs. 30/30, and its supervised control is a small effect *opposite* in sign (66.7% vs. 56.7%, p=0.0019); by gap, Corridor's RL split (90) again dwarfs its supervised control's (10, oppositely signed), interaction test confirming (p=1.86×10^{-9}). Reliance scales differ by construction, so we do not compare raw d_z across architectures or environments; the dissociation reproduces under both, within the settings tested.

Boundary of online-RL generality. Having shown the dissociation holds across architectures and environments, we next ask whether it depends on the RL algorithm itself. A matched on-policy PPO actor-critic (Schulman et al. (2017); Appendix O) answers this as a boundary condition, not another replication: PPO formed the shortcut under both EARLY and LATE (30/30 vs. 30/30; 0-point gap, p=1.00; Table 3), with the continuous margin likewise near zero (mean diff. -0.145, 95% CI $[-0.442, 0.152]$, d_z= $-$ 0.18), so this is not a binary-ceiling artifact. This is an absence of the early-evidence benefit, not robustness against formation. PPO and DQN differ in several coupled respects (replay, on- vs. off-policy updating, value- vs. actor-critic learning), so we do not identify which accounts for this boundary.

Interventional controls. Table 1 collects the anchor alongside every individually tested intervention on dose, placement, action coverage, integration time, exploration, and replay; full statistics, CIs, and p-values are in Appendix A, Table 3.

Halving the evidence dose and forcing identical two-sided action coverage at the 40 evidence steps each leave the split essentially unchanged, though paired reanalysis of the latter against the natural anchor detects a small, real, LATE-localized reduction ($\approx$3.5% of the anchor's magnitude, p=0.001), not a sufficient explanation on its own. Moving the same windows inward toward the Phase-I midpoint attenuates the split, an asymmetric placement effect reported without inferring mechanism. Integration time and a flat exploration schedule each attenuate the raw contrast substantially ($\approx$36% and $\approx$24%, respectively) but leave a large residual; the full lead-time family (Figure 3b) drops sharply to 1200 steps and flattens thereafter, and a duration-only control leaves formation at ceiling, confirming this is about integration time, not total duration. Because the flat-exploration change alters the whole trajectory rather than isolating evidence-step actions alone, it shows EARLY's protection is schedule-dependent, not an independent fourth mechanism. Only removing replay entirely, itself confounded with batch size (above), closes the split. These attenuations are not additive and must not be summed (Section 5).

5 DISCUSSION

The phenomenon. Formation and its timing sensitivity are separable properties of a learning system: every paradigm forms the shortcut under at least one ordering, so formation itself is not RL-specific; it is the *size of the gap* between orderings that differs by paradigm. The anchor's split is not a clean, isolated estimate of "temporal placement, all else held fixed", since EARLY and LATE also differ in evidence-step action-conditioned data, post-evidence training time, and exploration regime – exactly what Table 1's interventions test one at a time.

What survives, and why it matters. Each control fails to abolish the effect alone, though that does not mean every confound is jointly eliminated, and these effects overlap and are not additive (Section 4.3). One bundled factor stands out: network maturity at evidence time is the one ingredient no control here removes independently, since relocating the window always changes it too. We call the surviving residual *path dependence*: what the agent has already done, not any single ingredient, shapes what a fixed later evidence budget can still change – a description of the data, not an identified mechanism. Candidate mechanisms, none established here, include reduced plasticity after early convergence (Dohare et al., 2024), dormant representations, and ordinary training-history non-stationarity. The stakes are practical: shortcut-mitigation conclusions validated only in fixed-dataset or offline settings are tested here in a regime where this dependence is much weaker, and 2 of 30 LATE anchor seeds never recover B-reliance within 8000 post-shift steps.

Two mechanistic probes, both bounded, plus a boundary condition. Removing replay entirely eliminates the split under a batch-size-1 online update, but batch size and gradient variance change simultaneously, so this is a replay/batch-size ablation, not evidence that replay alone is causal (Appendix Q). A learning-state rewind at step 3600 (online/target networks reset to their initial state, replay/RNG/environment state held fixed) leaves a large interaction on the continuous margin that grows specifically during LATE's evidence window (-3.267, 95% CI $[-3.668, -2.866]$ at step 3800; Appendix R), consistent with accumulated learner state affecting how late evidence is processed; replay is preserved throughout, so this does not isolate replay from network state or establish consolidation/plasticity. Separately, a matched on-policy PPO agent reproduces none of this sensitivity while still forming the shortcut under both orderings, so whatever produces the DQN effect is not present merely by virtue of a live data stream; PPO and DQN differ in several coupled respects (replay, on/off-policy updating, value- vs. actor-critic learning), so this boundary does not identify which is responsible.

Formation occurs broadly across paradigms, but its timing-dependence is dramatically larger under the value-based online-RL regime studied here (Section 4.2), shrinking under several explanations but disappearing under none, and is absent under a tested on-policy alternative.

6 LIMITATIONS

Both environments are purpose-built to make this comparison identifiable, bounding how far the results generalize. The mechanism behind the residual timing sensitivity is not identified: action-coverage asymmetry, integration time, exploration schedule, replay exposure, and update count each leave a large residual (Table 1), without determining which factor is responsible; this residual is

evidence for path dependence, not an identified mechanism, and the attenuation percentages are descriptive and must not be summed. LATE's arm is not measured in a period cleanly separated from its own evidence exposure, since two of its five checkpoints fall inside its own evidence window (Section 3.2). Network maturity and post-evidence integration time are not separated: moving the window later always changes both together, and no experiment fully decomposes position, integration time, and duration (the PRIOR×POST probe, Appendix B, bears on this but is confounded).

Three follow-up interventions each bound, rather than close, the mechanistic question. Movement during Phase I is scripted in the anchor construction; an organic, action-driven replication (Section 4.3) confirms the split survives, but realized exposure is then endogenous and unequal (LATE median 26 visits vs. EARLY's 99.5, 8/30 LATE seeds seeing zero; Appendix P), not matched – the 22 nonzero-exposure LATE seeds also all forming does not make it matched. Removing replay in a batch-size-1 online update eliminates the split entirely (Appendix Q), but is confounded with the accompanying batch-size/gradient-variance change, so it does not establish replay as uniquely causal. A learning-state rewind at step 3600, preserving replay/RNG/environment state (Appendix R), leaves a large, timing-selective interaction consistent with accumulated learner state affecting late-evidence processing, but since replay is preserved throughout, it does not isolate replay from network state or establish consolidation/plasticity.

Fixed-dataset timing effects are small but not exactly zero (Appendix E); the TD control is one narrow, self-loop scaffold, not a generalization to bootstrapping with replay or live data. Timing sensitivity is not placement-invariant: only the extreme-window construction and its exact replications produce a categorical split, and the interior-window breakdown (n=8 unformed against n=22 formed; Appendix N) is too small to characterize further. This program was exploratory before confirmatory (Appendix M). We make no claim of universal RL specificity, only that the measured dependence is dramatically larger, by formation-rate gap, in the value-based online-RL regime tested than in the matched fixed-dataset settings, surviving each individually tested intervention without their joint contribution being established; nor do we establish which algorithmic difference separates this regime from PPO (Section 4.3) – replay, update structure, and on- vs. off-policy learning remain coupled, and other on-policy algorithms or replay-augmented PPO were not tested.

ETHICS STATEMENT

This work uses only synthetic gridworld and vector-observation simulations, with no human subjects, personal data, or foreseeable dual-use concern.

REPRODUCIBILITY STATEMENT

All code, raw per-seed artifacts, and figure/table-generation scripts are in the supplementary material. Section 3 specifies the environments, controller, and paradigms; the appendix details the integrity checks (Appendix J), RNG correction (Appendix G), the PPO boundary condition (Appendix O; code in `agent/ppo_model.py`, `experiments/train_ppo.py`; artifacts in `results_ppo_evidence_temporal/`), organic visitation (Appendix P; `results_organic_evidence/`), the replay-necessity ablation (Appendix Q; `results_noreplay/`), the learning-state rewind (Appendix R; `results_plasticity_reset/`), and every architecture, hyperparameter, and seed (Appendices K–L). All reported numbers are computed directly from raw artifacts.

AI USE STATEMENT

We used a generative AI system to help us ideate and refine both our own ideas and its own suggestions on experimental design, implementation, analysis, and manuscript/figure drafting. It implemented the environments and agents that the human authors ultimately settled on, computed and documented all reported statistics, and helped diagnose problems with our proposed ideas several times before reaching the final design. All implementations were reviewed and approved by the human authors; several were revised or rejected after scrutiny. We did not use generative AI to fabricate results or citations: every numerical claim was cross-checked against raw artifacts, every citation against the cited paper's own text, and all AI-assisted code reviewed. The authors take responsibility for the final content of this work.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *International Conference on Learning Representations*, 2023.

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3061–3071. PMLR, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Marcel Hussing, Claas Voelcker, Igor Gilitschenski, Amir-massoud Farahmand, and Eric Eaton. Dissecting deep rl with high update ratios: Combatting value divergence. In *Reinforcement Learning Conference*, 2024.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16828–16847. PMLR, 2022.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021.

Aahlad Puli, Lily H Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. *arXiv preprint arXiv:2308.12553*, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

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

Hado van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

A COMPLETE EXPERIMENTAL TIMELINE AND FULL RESULTS TABLES

The research program proceeded in the following order, each stage motivated by a specific gap left open by the previous one. Experiment numbers below match the project's internal numbering, referenced elsewhere in this appendix. Table 2 summarizes the causal arc; Table 3 gives the full numerical record of every load-bearing formation/timing-sensitivity contrast reported in the main text.

1. **Organic-vs-rebalanced pilot** ($n = 10$ paired seeds): an agent with live control over its own visitation shows far more persistent post-shift reliance on feature A than an agent whose visitation is exogenously rebalanced (Cohen's d up to -11.6 on visitation entropy, $d = 1.06$ on residual post-shift Reliance(A)).
2. **Exact-trajectory yoking** ($n = 20$, historical, superseded): replaying an organic run's own Phase-I trajectory off-policy and matching Phase-II movement to that seed's own empirical statistics shows no detectable difference from the organic outcome at $n = 20$ (all $p > 0.10$, a weak equivalence-style negative result at this sample size, not proof of equivalence). This manipulates post-shift visitation breadth, not the temporal-placement construction itself, so it was at the time suggestive but not direct support for treating a fixed, externally-defined visitation script as compatible with online RL. It has since been superseded by a direct confirmatory test with fully organic, action-driven Phase-I movement and the actual evidence-timing construction (Experiment 12 below, Section 4.3, Appendix P); we retain this earlier check only as historical context for the controller design in Section 3.3, not as evidence alongside the confirmatory result.
3. **Evidence dose-response** ($n = 20$ per condition): directly and causally manipulating Phase-I disagreement exposure (independent of any agent policy) shows that even modest evidence exposure is close to sufficient to prevent later shortcut reliance, in a threshold-like rather than smoothly graded pattern.
4. **Evidence threshold sweep** ($7 \times 20 = 140$ runs): refines the threshold to a noisy transition zone spanning roughly 0.25–1% exposure.
5. **Count-vs-rate** ($n = 20$ per duration): an initial design varying Phase-I duration at fixed evidence count finds an apparent rate effect (Cohen's $d_z = -0.60$ to -0.96), later superseded once evidence position was manipulated directly (Experiment 6) without needing a rate-based account.
6. **Experiment 6, temporal placement, the anchor result** ($n = 30$): manipulating only the position of a fixed 40-visit evidence budget within a fixed 4000-step window produces the 0/30 versus 30/30 formation split reported in Section 4.1. Originally run with a shared-RNG construction later found to produce a small evidence/agreement content mismatch; corrected and re-run (Appendix G).
7. **Experiment 7, shift-timing probe** ($n = 30$ per condition): the same LATE evidence becomes progressively more effective given more lead time before the shift (formation drops from 100% to 53.3% as lead time grows from 0 to 4000 extra steps, Holm-$p = 0.0029$ at $t = 5200$ and 0.00049 at $t = 8000$; trend across lead times, Cochran's $Q = 34.38$, $p = 1.6 \times 10^{-7}$), ruling out a permanently-inert account of late evidence (Section 4.2).
8. **Experiment 8, replay-sampling manipulation** ($n = 30$ per condition): direct, bidirectional manipulation of replay-sampling weight (Section 4.2).
9. **Experiment 9, consolidation timing** ($n = 30$ per condition): a PRIOR $\times$ POST factorial isolating elapsed pre-evidence training time from post-evidence lead time; PRIOR necessarily bundles lower epsilon and more pre-evidence updates, so this is reported as a bounded, confounded probe, not an independent identification of elapsed experience as an ingredient.
10. **Experiment 10, environment experience vs. update count** ($n = 30$ per condition): the update-count-matched contrast reported in Section 4.2.
11. **Experiment 11, agreement-cell diversity** (target $n = 30$, 179/180 runs completed): inconclusive at the resulting $n = 29$, no contrast surviving correction; reported as inconclusive, not as a null result, and not load-bearing for any claim in the main text.
12. **Fixed-dataset supervised control (H5)** ($n = 30$): Section 4.2.
13. **Fixed-dataset TD, invalidated pilot**: a one-sided construction diagnosed as a design failure before any confirmatory run (Appendix H).
14. **Fixed-dataset TD, repaired** ($n = 30$): Section 4.2.
15. **TinyViT architecture generality** ($n = 30$ per condition, 4 conditions; online-RL arm later corrected and independently rerun, Appendix G): Section 4.3.

16. **Corridor Foraging, second environment** ($n = 30$ per condition, 4 conditions: RL and supervised, EARLY and LATE): Section 4.3.
17. **Evidence-budget robustness** ($n = 30$): halving the anchor's evidence budget to 20 visits in the same extreme windows, same corrected dedicated-RNG construction, reproduces the categorical split (0/30 vs. 30/30) with an essentially unchanged effect size; Section 4.3.
18. **Window-placement robustness** ($n = 30$): moving the same 40-visit, 400-step-wide windows inward, symmetric about the Phase-I midpoint, attenuates but does not eliminate the split (0/30 vs. 22/30); Section 4.3.
19. **Forced action-coverage control** ($n = 30$): a post-hoc audit found that annealing ϵ gives EARLY and LATE evidence measurably different action-conditioned data (48.4% vs. 96.7% one-sided harvests matching the shortcut cue); forcing an identical, exact two-sided action sequence at the 40 evidence steps for both orderings leaves the categorical split intact (0/30 vs. 30/30), with a small ($\approx$3.5% of the total effect), LATE-localized reduction on paired reanalysis; Section 4.3.
20. **Organic, action-driven visitation** ($n = 30$): the scripted controller removed entirely, movement fully action-driven, only evidence-cell content temporally gated; reproduces the categorical split (0/30 vs. 30/30) despite endogenous, unequal realized exposure; Section 4.3, Appendix P.
21. **Replay-necessity ablation** ($n = 30$): experience replay removed entirely, batch size 1, single fresh transition per update, identical Double-DQN target and optimizer otherwise; eliminates the split (29/30 vs. 29/30), confounded with the accompanying batch-size/gradient-variance change; Section 4.2, Appendix Q.
22. **Learning-state rewind** ($n = 30$ per arm, 2×2 factorial plus a zero-evidence control): online and target networks reset to their seed-specific initial state at step 3600, replay/RNG/environment/controller state held fixed; a large, timing-selective interaction on the continuous margin, not a replay-isolation or plasticity result; Section 5, Appendix R.

B FULL RECORD OF THE CAUSAL-ELIMINATION EXPERIMENTS

Table 4 gives the complete numerical record behind the replay-exposure, update-count, consolidation-timing, and shift-timing-probe results summarized narratively in Section 4.2; each row is a single pre-specified comparison, not a selection among several attempted analyses.

C POST-EVIDENCE INTEGRATION TIME: FULL LEAD-TIME CURVE AND DURATION-ONLY CONTROL

Defining integration_time = phase1_steps − evidence_window_end, Table 5 gives the full monotonic relationship between post-evidence integration time and formation, using the existing shift-timing family (Appendix A, Experiment 7) together with the corrected anchor and the new confirmatory `late_t8000_rngfix` condition. $\text{LATE}_{t=8000}$ is the closest available match to EARLY's own 3600-step integration time (within 400 steps) among the four LATE_t lead-time points; this comparison point was fixed by this pre-stated closest-match criterion, not selected after inspecting which of the four candidate points gives the most favorable comparison. The curve drops sharply from 0 to 1200 integration steps (100%→96.7%→63.3%) and then flattens substantially from 1200 to 4000 steps (63.3%→53.3%/50.0%), so most of the attenuation happens early in the lead-time family and little further reduction occurs beyond 1200 steps. $\text{LATE}_{t=8000}$ is also the point of lowest formation, and hence smallest gap, among the four tested lead times: the pre-stated closest-integration-time criterion and the most-attenuated point in this family happen to coincide, which we disclose for transparency; this does not change that a large residual (a 50-point gap) remains even at this most-attenuated point.

The construction-control check comparing the corrected $\text{LATE}_{t=8000}$(rngfix) confirmatory run against the originally run $\text{LATE}_{t=8000}$(orig.) construction, paired by seed, finds no significant difference (mean diff. $+0.05$, 95% CI $[-0.43, 0.54]$, Wilcoxon $p = 0.82$, $d_z = 0.04$): the correction does

Table 2: Experimental roadmap. Each row answers one question in the causal arc; n is the number of paired seeds.

Question	Design	n	Outcome
Does evidence timing control RL formation?	EARLY vs. LATE, fixed count/window, CNN-DQN, corrected construction	30	0/30 vs. 30/30
Is replay exposure sufficient?	boost LATE / throttle EARLY exposure	30	insufficient alone
Is raw update count sufficient?	match updates across differing experience	30	insufficient alone
Does fixed-dataset SGD reproduce formation and timing dependence?	identical content/order, fixed dataset, cross-entropy	30	forms; much weaker timing dependence
Does self-loop bootstrapping restore timing dependence?	fixed two-sided transitions, Double DQN target	30	forms; much weaker timing dependence
Does the dissociation survive a different architecture?	CNN → TinyViT, same task/schedule	30	reproduces
Does the dissociation survive a different environment?	grid → Corridor Foraging, vector obs., MLP	30	reproduces (smaller RL magnitude; supervised control shows a small, oppositely-signed effect)
Is the extreme-window split robust to a 2× dose reduction?	20 rather than 40 visits, same extreme windows, CNN-DQN	30	0/30 vs. 30/30, unchanged
Does the split survive moving the windows inward?	EARLY-MID/LATE-MID at $[800, 1200)/[2800, 3200)$, CNN-DQN	30	0/30 vs. 22/30, attenuated
Is the annealing-ϵ action-coverage asymmetry sufficient to explain the split?	force an identical, exact two-sided action sequence at the 40 evidence steps for both orderings	30	0/30 vs. 30/30, intact; small (≈3.5%) real reduction on paired reanalysis
Does the split survive closely approximating post-evidence integration time?	extend LATE's Phase I to 8000 steps (same $[3600, 4000)$ window), giving LATE 4000 vs. EARLY's 3600 post-evidence integration steps (closest available match, within 400 steps), corrected dedicated-RNG construction	30	0% vs. 50.0%, raw contrast reduced ≈36%, large residual remains
Does the split survive a flat, non-annealed exploration schedule?	$\epsilon = 0.1$ fixed for the entire run (both orderings), instead of annealing 1.0→0.02	30	10.0% vs. 100%, raw contrast reduced ≈24%; only manipulation to break EARLY's 0/30 floor
Does the split survive fully organic, action-driven visitation?	scripted controller removed entirely, movement action-driven, only evidence-cell content temporally gated	30	0/30 vs. 30/30, despite endogenous, unequal realized exposure
Does removing experience replay close the split?	batch size 1, single fresh transition per update, replay never sampled, Double-DQN target/optimizer otherwise unchanged	30	29/30 vs. 29/30, closes the split; confounded with the accompanying batch-size/gradient-variance change
Does a learning-state rewind reveal a timing-selective effect?	online/target networks reset to their initial state at step 3600, replay/RNG/environment state preserved	30	large, timing-selective interaction on the continuous margin; does not isolate replay from network state

not change this result, resolving the construction caveat that previously qualified the integration-time reanalysis (`confirmatory_late_t8000_rngfix_report.md`).

The two ZERO$_t$ conditions run the same duration-manipulation without any evidence at all: formation stays at ceiling (30/30) at both durations, ruling out "training longer, on its own, changes formation" as an alternative account of why LATE$_{t=8000}$'s formation rate is lower than LATE$_{t=4000}$'s. This supports attributing the reduction to post-evidence integration time rather than total Phase-I duration, without an independent, fully-crossed isolation of the two (see the following paragraph).

We do not orthogonally decompose evidence position, integration time, and total Phase-I duration in a single fully-crossed design; LATE$_{t=8000}$ changes total duration and relative position within Phase I simultaneously with integration time, and this is the closest matched comparison this project's controller family supports without a substantially larger new experiment (Section 6).

D EXPLORATION SCHEDULE: FLAT-EPSILON CONTROL, FULL RECORD

Replacing the annealed schedule (ϵ: 1.0 → 0.02 over the first 3000 of 4000 Phase-I steps) with a flat $\epsilon = 0.1$ for the entire run, in both orderings, on the corrected dedicated-RNG construction, gives formation of 3/30 (10.0%) for EARLY and 30/30 (100%) for LATE (McNemar $p = 1.49 \times 10^{-8}$, Wilcoxon $p = 2.61 \times 10^{-8}$, $d_z = -2.127$, mean paired difference -4.666, 95% CI $[-5.451, -3.881]$). This is the only manipulation tested anywhere in this project, across

the natural anchor, the 20-visit dose-robustness condition, the interior-window condition, and the forced-action-coverage control, under which EARLY is not categorically 0/30.

Comparing this flat-epsilon run to the natural annealed-epsilon anchor directly, paired by seed (per-seed margin correlation $r = 0.386$, weaker than the forced-action-coverage control's $r = 0.97$ because flat epsilon changes far more of the training trajectory than a targeted 40-step action override): paired $t = -4.06$, $p = 0.00034$, Wilcoxon $p = 9.98 \times 10^{-7}$, mean(natural−flat) $= -1.505$, an attenuation of the natural anchor's -6.172 magnitude by ≈24.4%, larger than the forced-action-coverage control's ≈3.5%. Because flat epsilon changes the exploration policy for the entire Phase-I and Phase-II trajectory rather than only the 40 evidence-step actions, this larger attenuation is not an independent confirmation of a distinct fourth mechanism beyond action coverage: a flat schedule is a superset intervention that includes, but is not limited to, equalizing evidence-step action coverage, so its effect should not be added to the ≈3.5% action-coverage figure.

E SUPERVISED LEARNING-RATE/PASS-COUNT GRID: FULL RECORD

To test whether the reported K=4, lr=10^{-4} CNN grid-task supervised operating point understates supervised timing sensitivity, we ran the full grid $K \in \{1, 2, 4\} \times \text{lr} \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, 30 seeds per cell per ordering (540 runs total, 0 failures). Two cells reproduce already-reported numbers as an internal consistency check: K=4, lr= 10^{-4} gives 56.7%/60.0% (the main text's reported supervised control, exactly), and K=1, lr= 10^{-4} gives 100%/100% (the main text's K=1 saturation check, exactly).

The largest formation-rate gap over the grid is 6.7 points (K=2, lr= 3×10^{-4}); the largest $|d_z|$ is 1.75 (K=2, lr= 10^{-3}, where both orderings sit at an identical near-floor rate (6.7%) yet the continuous margin differs significantly, illustrating why the continuous margin rather than the binary rate alone is the primary readout). Table 7 compares this grid's best case against the CNN grid-task value-based RL operating points tested in this section's controls (natural anchor, forced action-coverage, flat exploration, matched integration time); TinyViT and Corridor RL operating points, and the separate PPO boundary condition (Appendix O), are reported in Table 3. Across every value-based RL operating point measured in this paper, by formation-rate gap this CNN grid-task supervised control's largest measured timing sensitivity remains far below RL's effect; by standardized effect size, the supervised best case ($|d_z| = 1.75$) remains smaller than every value-based RL operating point, though the gap narrows considerably under RL's most-attenuated forms ($|d_z| = 2.13$–2.80).

This is a targeted 3×3 sweep of the CNN grid-task operating point, not an exhaustive search; the pattern (gap capped well below 10 points within this sweep, $|d_z| < 2$ at every cell) is clear enough that further cells are unlikely to change the qualitative conclusion, though we do not claim exhaustiveness or that 6.7 points is the largest gap among every supervised control in this paper (Corridor's is larger, at 10 points, oppositely signed).

Corridor's supervised control (Appendix K) uses lr= 10^{-3}, the same rate at which this grid's own three lr= 10^{-3} cells ($K \in \{1, 2, 4\}$) all show a 0.0-point gap. That the same nominal learning rate yields a 0-point gap on the grid task but a non-zero, oppositely-signed 10-point gap on Corridor indicates this small supervised effect is task- or environment-specific rather than an artifact of a particular learning-rate choice; we did not re-run the grid task's supervised control at Corridor's full K-range or vice versa, so this comparison is suggestive, not a matched cross-environment sweep.

F MOVE-OVERRIDE FORENSIC AUDIT: SYMMETRIC, NOT LOAD-BEARING

The scripted evidence-timing controller (`TemporalEvidenceController`) is active for all 4000 Phase-I steps, not only the 40 evidence steps, and unconditionally overrides the agent's position regardless of the action passed to the environment's `step()`, never reading the action actually taken for a `MOVE`. Consequently, for a `MOVE` action, the stored transition's next state is generally unrelated to the natural, un-scripted successor of the current state. We quantify this directly at full scale (all 30 seeds, both orderings, 0 failures; Table 8).

This effect is real, large (≈97%), and, critically, essentially identical between EARLY and LATE at every breakdown (differences of 0.0-0.7 percentage points, well within the ≈0.4-3% seed-to-seed standard deviation): it is a genuine, previously undocumented design property of the entire scripted-

Phase-I controller family, worth disclosing, but symmetric between orderings and therefore not a source of the EARLY/LATE asymmetry under audit. Its most plausible downstream consequence is diffuse: gradient signal from inconsistent movement transitions flows through the network's shared convolutional backbone (movement and harvest Q-values share all layers up to the final linear head), potentially adding noise to learned visual features generally, but this affects both orderings equally by construction and so cannot explain the categorical split.

G SHARED-RNG CONSTRUCTION ISSUE AND CORRECTED ANCHOR REPLICATION

The original run of Experiment 6 (the online-RL anchor, Section 4.1) constructed the EARLY/LATE evidence-timing controller with a single shared `numpy RandomState` stream for both the evidence-cell draw and the agreement-cell draw. Because EARLY and LATE interleave evidence and agreement draws in different orders (evidence draws cluster at the start of the sequence under EARLY and at the end under LATE), and because `randint` consumes a state- and range-dependent number of underlying draws under rejection sampling, this shared stream does not guarantee that the two orderings' resulting (cell, label) sequences form an identical multiset, even though the same number of evidence and agreement examples is drawn from the same pools in both conditions. We found this by direct reconstruction: recomputing the exact per-step sequence for all 30 seeds under the original construction shows content-hash equality fails in 30 of 30 seeds, with label counts differing by up to 25 examples out of 4000 across seeds (seed 0's late-window label sum differs from its early-window label sum by 4). This defect is present only in the online-RL arm's original construction; the fixed-dataset supervised and TD arms (Section 3.3) already use a dedicated agreement-draw RNG stream (seed offset `+6`, distinct from the evidence-draw stream's `+5`) and satisfy exact content-hash equality as originally run, verified directly.

We corrected the online-RL construction by giving the agreement-cell draw its own dedicated stream, identical in spirit to the fixed-dataset arms' existing construction, and re-ran all 30 paired seeds with every other hyperparameter, the environment, the network, the optimizer, the replay buffer, the target-sync schedule, and the total step budget held exactly fixed. This corrected construction achieves exact content-hash equality, order-hash inequality, and identical label counts for all 30 of 30 seeds (verified by direct reconstruction, not only by inspection of the code path). Before this confirmatory run, we verified by code inspection that the correction is an opt-in, default-off change: every other condition and experiment in this paper leaves the corresponding flag unset and is provably unaffected, and two fresh runs of the unmodified code path produced byte-identical logs, confirming no unrelated nondeterminism was introduced.

The corrected replication reproduces the original result closely: formation is 0/30 versus 30/30 in both the original and corrected runs; the corrected run's effect size is $d_z = -9.18$ (mean paired difference -6.17, 95% CI $[-6.41, -5.93]$) against the original run's $d_z = -9.77$ (mean paired difference -6.21, 95% CI $[-6.44, -5.99]$). The result is robust to the corrected construction and does not materially change under it; this does not mean the defect had no effect at all, only that the categorical result and its effect size reproduce within a small margin once corrected. We report the corrected run as the primary anchor result throughout the main text and retain the original run's numbers here for transparency.

TinyViT integrity audit and corrected replication. The TinyViT RL condition (Section 4.3) originally used the identical EARLY/LATE evidence-timing construction as the original, uncorrected CNN run: the environment and controller are seeded and constructed identically regardless of network architecture, and `-network` selects only the Q-network class, touching no RNG stream used by the controller. We verified this by reconstructing the exact per-step sequence for all 30 TinyViT seeds (cross-checked against the real `disagree_cells` recorded in TinyViT run metadata, which matched the reconstruction exactly): content-hash equality failed in 30 of 30 seeds, with a maximum label-count discrepancy of 25 out of 4000 across seeds, the same magnitude as the original CNN defect.

We then independently reran TinyViT online RL, both orderings, all 30 seeds, using the same dedicated evidence/agreement RNG streams already validated for the CNN anchor (Appendix G), with every other hyperparameter, the architecture, the optimizer, and the total step budget held exactly fixed. A preflight check before training, and a post-run check against the resulting real artifacts,

both confirm content-hash equality, label-count equality, evidence-cell-set equality, and order-hash inequality at 30 of 30 seeds. All 60 runs (30 seeds $\times$ 2 orderings) completed with no failures, retries, or substituted seeds.

The corrected replication reproduces the original categorical result (0/30 versus 30/30) with an essentially unchanged paired mean difference and overlapping confidence intervals: corrected mean paired difference -5.25, 95% CI $[-5.52, -4.99]$, versus the original run's -5.44, 95% CI $[-5.61, -5.26]$ (Cohen's $d_z = -7.12$ corrected versus -11.05 original; both extreme by conventional standards). The architecture-level dissociation is therefore robust to the RNG correction. We report the corrected run as the primary TinyViT result throughout the main text and retain the original run's numbers here for transparency. Corridor Foraging's controller (both arms) was built with a dedicated agreement-draw stream from the start and independently verified at 0 of 30 content-hash failures (Appendix K).

H TD-V1: INVALIDATED CONSTRUCTION

An initial fixed-dataset TD construction assigned, at every scripted cell, a single deterministic transition using the ground-truth-correct harvest action: $(s, \text{harvest}(\text{true } B), +1, s, \text{done} = \text{False})$, with no corresponding transition for the incorrect action. At a single sanity seed, probe accuracy remained near chance (0.50) for nearly the entire training run, Reliance(A) exceeded Reliance(B) by 3 to 5 times throughout, and B-decodability plateaued at 0.75, far below the approximately 0.99 seen in every other reported condition in this paper. The diagnosis: a Huber-loss TD update only pulls the *taken* action's value toward its target, so a one-sided dataset never contrasts the two harvest values the way cross-entropy's softmax does over both classes; the untrained harvest logit never receives gradient pressure, and the network defaults to the more salient feature. This is a structural design failure, not a scientific null result, and no number from this construction appears anywhere in this paper's results. It motivated the two-sided repair described in Section 3.3, which explicitly re-establishes the two-sided contrast that the invalidated design lacked.

I TINYVIT ARCHITECTURE DETAILS

TinyViT tokenizes a $32 \times 32 \times 3$ observation into 16 non-overlapping 8×8 patches. Each patch is flattened to a 192-dimensional vector and linearly projected to an embedding dimension of 32. A fixed (non-learned) sinusoidal positional encoding is added to each patch embedding. Two standard transformer encoder layers follow (2 attention heads each, feedforward dimension 64, ReLU activation, no dropout), followed by mean pooling over the 16 token embeddings and a single linear layer to the six-dimensional action or logit output. Total parameter count is 23,462, compared to 154,214 for the CNN (`QNet`: two convolutional layers, two fully connected layers). We verified directly from the environment's rendering geometry that feature A occupies patches $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ and feature B occupies patches $(0,2)$, $(0,3)$, $(1,2)$, $(1,3)$ in the 4×4 patch grid, a clean, non-overlapping 2×2 block for each feature, so no single patch mixes pixels from both features.

J INTEGRITY METHODOLOGY

For every paired run in every experiment reported in the main text, we verify, directly from the constructed dataset or logged run metadata rather than from the controller's specification: (1) a content hash over the sorted multiset of (cell, label) pairs is identical between EARLY and LATE; (2) an order hash over the sequence as constructed is different between EARLY and LATE; (3) the evidence count equals exactly 40 (grid task) or 24 (Corridor Foraging) in both orderings; (4) label counts are identical between orderings; (5) for the RL arms, the disagreement-cell layout, optimizer hyperparameters, and total environment/update budget are identical between orderings for a given seed; (6) for the supervised and TD arms, the total number of optimizer updates is identical between orderings (252 for both H5 and TD, by construction); (7) for the TD arm additionally, every underlying example produces exactly two transitions with an exact $+1/-1$ reward pairing, every transition has $s' = s$, and no transition is terminal. The corrected grid-task RL anchor, the corrected TinyViT RL replication, and Corridor Foraging (both arms) pass all checks at 0 of 30 failures, verified by direct reconstruction rather than code inspection alone. The originally run (uncorrected) constructions for

both the grid-task RL anchor and the TinyViT RL condition failed the content-hash check in 30 of 30 seeds, with a maximum label-count discrepancy of 25 of 4000; both were subsequently corrected and independently rerun with dedicated evidence/agreement RNG streams, and both corrected reruns pass all checks (Appendix G). The 20-visit dose-robustness and interior-window-placement conditions (Section 4.3) use the same corrected, dedicated-RNG construction from the start and are independently verified to pass all checks at 0 of 30 failures. All other load-bearing experiments reported in Table 3 pass every check at 30/30. The one exception in this paper's full record is the agreement-cell diversity experiment (Experiment 11, not load-bearing), which reached 179/180 planned runs after a single seed was lost to a repeated, externally caused low-memory termination with no in-guest cause found across two forensic passes (no kernel out-of-memory event, no control-group memory limit or pressure, no resource-limit restriction, and no memory-management daemon active), and was frozen at that sample rather than further retried.

Scope of these checks: visitation-script integrity, not action-conditioned transition integrity. Checks (1)-(7) above verify that the *scripted* construction, meaning which cell the controller places the agent at and the ground-truth label there, is correct and matched between orderings. They do not verify what the agent's own ϵ-greedy policy does once placed there: the RL arms' controller overrides only position, never the chosen action, so an evidence step's actual learning transition (a move, or a `HARVEST0`/`HARVEST1` of either correctness) depends on ϵ and the current policy, and is not covered by any of checks (1)-(7). A dedicated action-conditioned audit (Section 4.3) found and quantified a real asymmetry this leaves undetected: LATE's evidence-step harvests are measured at 96.7% one-sided toward the shortcut cue A versus EARLY's 48.4%, a gap this appendix's checks alone cannot see. A follow-up integrity layer (exact per-seed reconstruction of the forced action sequence, cross-checked against every run's logged actions) was added specifically for the forced-action-coverage control that tests this asymmetry's causal role, passing at 0 of 30 failures on both the visitation-script and the action-conditioned layers. We recommend any future extension of this controller family report both layers explicitly rather than treating visitation-script integrity as verifying the full learning signal.

K CORRIDOR FORAGING: FULL ARCHITECTURE AND HYPERPARAMETERS

The evidence-timing intervention places 24 evidence examples within a 2400-step Phase I, analogous in structure to the grid task's 40-in-4000. The observation vector is 6-dimensional: two dimensions one-hot encoding feature A; two dimensions (b_1, b_2) encoding feature B as $b_1 \oplus b_2$, with a fresh random bit redrawn on every observation (dedicated RNG stream, seed offset `+20`) so the network cannot memorize a fixed per-station encoding; one dimension for the current station's position, normalized to $[0, 1]$; one dimension for the current station's depletion countdown, normalized by the 5-step cooldown: a correct harvest depletes that station's reward for 5 steps, giving different task dynamics from the grid task's non-depleting cells. The agent (`CorridorQNet`) is a 3-layer MLP, $6 \to 64 \to 64 \to 4$, ReLU activations, matching the grid task's Double DQN training procedure exactly (Adam 10^{-3}, batch size 64, replay buffer capacity 20,000, warmup 500 steps, target sync every 200 steps, epsilon 1.0 to 0.02 over the first 1800 of 7200 total steps, gradient-norm clipping at 10). The fixed-dataset supervised control shares the same architecture and initialization, with $K = 16$ passes and Adam 10^{-3}, selected by a pre-confirmatory sweep over $K \in \{4, 8, 16, 32, 64\} \times \text{lr} \in \{10^{-4}, \ldots, 10^{-2}\}$ on one representative seed and condition before any EARLY/LATE comparison. Two implementation defects were found and corrected during piloting, before any confirmatory run and without reference to comparative EARLY/LATE outcomes: an initial one-hot (then 3-bin) position encoding let the network memorize a position-to-label lookup table, since B is a deterministic function of station, bypassing the intended A-versus-B competition entirely; and an initial per-station-fixed encoding bit let the network memorize a (position, bit)-keyed lookup instead of the true exclusive-or relationship. Both are fixed in the construction described above and in Section 3.3.

L HYPERPARAMETERS AND SEEDS

All experiments use seeds 0–29 for their respective paired comparisons unless otherwise noted (exact-trajectory yoking uses 0–19; the dose, threshold, and consolidation-timing experiments use 0–19 or 0–29 as stated in Appendix A). Common RL hyperparameters (grid task): grid size

6, discount $\gamma = 0.9$, Adam learning rate 10^{-3}, batch size 64, replay buffer capacity 20,000, warmup 500 steps, target-network sync every 200 environment steps, epsilon from 1.0 to 0.02 over 3000 steps, gradient-norm clipping at 10. Supervised and TD arms (grid task): Adam learning rate 10^{-4} (selected once, before the confirmatory run, by a single-seed sweep over $\{10^{-4}, 1.5 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}\} \times \{2, 4, 8, 16\}$ passes, not re-tuned afterward), $K = 4$ passes, batch size 64, gradient-norm clipping at 10. TD arm additionally: target-network sync every 200 optimizer updates, $\gamma = 0.9$, matching the RL arm's values exactly. Corridor Foraging hyperparameters are listed in Appendix K.

M EXPERIMENTAL DISCLOSURES AND RESEARCHER DEGREES OF FREEDOM

This research program was exploratory before it was confirmatory, and we disclose that order of operations rather than presenting every control as planned from the outset. The anchor result (Experiment 6) was collected first; the action-coverage, integration-time, and flat-epsilon controls, and the supervised learning-rate×pass-count grid, were each designed and run afterward, in direct response to specific audit-style objections raised against the already-collected anchor (respectively: that annealed ϵ could confound action coverage with timing; that the anchor's checkpoints overlap LATE's own evidence window; and that the reported supervised operating point could have been an unrepresentatively insensitive choice). Each of these controls is a single, fully-run, pre-specified comparison, not a post-hoc selection among many attempted analyses, and we report every one we ran, including the two that did not survive as independent mechanisms (action coverage, superseded in interpretation by the broader flat-epsilon result) and the one that remains explicitly inconclusive (Experiment 11, agreement-cell diversity). The paper's full experimental history also includes one invalidated pilot (the one-sided TD-v1 construction, Appendix H, diagnosed as a design failure before any confirmatory run and never used as evidence), two implementation defects found and corrected during Corridor Foraging piloting before any confirmatory EARLY/LATE comparison (Appendix K), a shared-RNG construction issue found and corrected in the online-RL anchor and TinyViT conditions (Appendix G), and one experimental direction (an early count-vs-rate design, Appendix A Experiment 5) superseded once evidence position was manipulated directly. Readers should weigh the original anchor result and these later, audit-motivated controls accordingly: the former is a single confirmatory measurement, the latter are targeted responses to specific alternative explanations, not an unbiased sample of everything that could have been tested.

Choice of the formation-rate gap as the primary cross-paradigm statistic. We use the formation-rate gap as the primary cross-paradigm comparison because raw continuous margins are not commensurate across paradigms: they are defined on different underlying quantities (a Bellman-bootstrapped Q-value margin vs. an unconstrained cross-entropy logit margin vs. a different-modality environment's margin), so a raw-margin standardized effect size (d_z) does not share a common scale across rows. The formation-rate gap does not have this problem: it is bounded at ± 100 percentage points and defined identically, as a difference of two proportions, regardless of paradigm, architecture, or environment. The later supervised learning-rate×pass-count sweep (Appendix E) subsequently reinforced this choice: its near-floor and near-ceiling cells (e.g. $K{=}2$, lr$= 10^{-3}$: both orderings at 6.7%) show that d_z on the raw continuous margin can be large or volatile precisely where within-condition variance is small, independent of how similar the two orderings' formation rates actually are. We report this chronology for transparency; the continuous-margin interaction test, used throughout as supporting evidence, was available and reported from the anchor result onward and its role did not change.

Anchor threshold robustness. The two arms' pre-shift margin distributions are well-separated with a wide gap: natural EARLY spans $[-4.061, -3.842]$ and natural LATE spans $[0.734, 3.261]$ (a 4.58-unit gap); forced-action EARLY spans $[-4.027, -3.946]$ and forced-action LATE spans $[0.262, 3.684]$ (a 4.208-unit gap). Any binary threshold placed anywhere between these two arm-separated intervals produces the identical 0/30-vs.-30/30 split, so the formation criterion's placement at zero is not a fragile choice near either distribution's edge.

Rule-of-three note on degenerate categorical arms. Several conditions in this paper report 0/30 or 30/30 formation. By the rule of three, an observed 0 events in 30 trials implies an approximate 95% upper confidence bound of ≈10% on the true underlying event rate (equivalently, an approximate 95% lower bound of ≈90% for an observed 30/30); we do not additionally report a two-sided Wilson

or Clopper-Pearson interval for these degenerate arms in the main text, since the point estimate together with this one-sided bound already conveys the relevant precision, and the continuous margin (never degenerate) carries the main statistical weight throughout.

N ADDITIONAL LEARNING CURVES, PER-SEED DISTRIBUTIONS, AND A SUPPLEMENTARY CONCEPTUAL FIGURE

Per-seed formation outcomes, continuous ($\mathrm{Rel}(A)-\mathrm{Rel}(B)$) margins, and full learning curves for every condition in Table 3 are included as CSV and JSON files in the supplementary material (one file per run, plus one aggregated file per experiment), alongside the exact scripts (`paper/scripts/`) used to generate every figure and table number in this paper from those files.

B-decodability in the supervised control. B-decodability from the CNN supervised control's penultimate features reaches ≈0.99 in both EARLY and LATE (Section 4.2): the ground-truth feature remains highly decodable even where behavior relies on A, so shortcut reliance need not reflect loss of the unused feature from the representation. This is a representational observation, not a mechanism result.

Formation-metric calibration at initialization. Before any training, Rel(A) and Rel(B) are both approximately zero (means 0.002 and -0.003, difference borderline at $p = 0.056$, magnitude three orders of magnitude below the ≈4-6-unit trained margins seen in every reported condition): the probe carries no meaningful intrinsic zero-point offset, so the formation criterion's threshold at zero is not biased by architecture or probe construction before any learning occurs.

Interior-window within-condition breakdown (underpowered). Within the interior-window LATE-MID condition (Section 4.3, 22/30 formed), comparing formed ($n = 22$) against unformed ($n = 8$) seeds on first-post-shift return is directionally consistent with formation carrying a behavioral cost (formed seeds do worse: difference -6.79, 95% CI $[-12.45, -1.12]$, Cohen's $d = -0.71$) but does not reach significance at this sample size (Mann-Whitney $p = 0.15$). This within-condition check is genuinely underpowered, not a null result, and should be read as directionally supportive of, not independent confirmation of, the pooled continuous-margin correlation reported in Section 4.1 (itself a predominantly between-condition association, not an independently powered within-condition test) as evidence that formation is behaviorally consequential.

Figure 4 already gives the formation-rate gap directly, by paradigm/architecture/environment, established quantitatively in Table 3; we do not additionally include a supplementary conceptual scatter of (mean formation rate, gap), since it would restate the same seed-paired numbers in a second, purely illustrative visualization without adding independent information.

Direct paired interaction tests, full statistics (former Figure 4 second panel). The continuous-margin interaction test underlying Section 4.2's supporting-evidence claim, for all four architecture/environment pairings against their own online-RL row: CNN SGD vs. RL, mean diff. -6.20, 95% CI $[-6.46, -5.94]$; CNN TD vs. RL, -6.14, $[-6.39, -5.90]$; TinyViT SGD vs. RL, -5.28, $[-5.54, -5.02]$; Corridor SGD vs. RL, -5.27, $[-5.77, -4.78]$; all $p = 1.86 \times 10^{-9}$, all 30 seeds agreeing per row.

O PPO BOUNDARY CONDITION: FULL RECORD

This appendix gives the complete record behind the PPO result summarized in Section 4.3. This is a single, compact boundary-condition check, not a new experimental axis: no PPO hyperparameter sweep, replay-augmented variant, or additional on-policy algorithm was run.

Architecture. A shared visual encoder identical in shape to the DQN `QNet` (Conv 3→16, kernel 5, stride 2; Conv 16→32, kernel 3, stride 2; FC 1152→128; no BatchNorm, no dropout), with a linear policy head (128→6) and a linear value head (128→1) both reading the same encoder output. The value head receives only the raw observation, no privileged environment state and no evidence label.

Hyperparameters (one configuration, used identically for both orderings). γ=0.99, GAE λ=0.95 (Schulman et al. (2016)), clip ϵ=0.2, entropy coefficient 0.01, value coefficient 0.5, Adam lr 3×10^{-4}, rollout length 128, minibatch size 32, 4 epochs per rollout, gradient-norm clipping at

0.5. These are conventional PPO defaults, fixed before any EARLY/LATE comparison and never tuned toward a desired result; no hyperparameter differs between EARLY and LATE.

Protocol. The identical environment, evidence-generation controller (`TemporalEvidenceController`), corrected dedicated-RNG construction, evidence windows (EARLY $[0, 400)$, LATE $[3600, 4000)$, 40 events each), Phase-I duration (4000 steps), post-shift duration (8000 steps), and formation criterion (mean of the last 5 pre-shift checkpoints, Rel(A) $>$ Rel(B), via the unmodified `compute_reliance`) as the DQN anchor; n=30 paired seeds (0–29). PPO substitutes on-policy rollout collection and a clipped actor-critic update for DQN's epsilon-greedy/replay-buffer/Bellman-target loop; no other protocol element differs.

Integrity. 60/60 runs completed, 0 errors, 0 seeds dropped or rerun for scientific reasons. Evidence content-hash equality between EARLY and LATE holds for 30/30 seeds; each condition has exactly 40 evidence events; and the full 4000-step (cell, label) sequence for both orderings is byte-identical to the DQN anchor's own construction (cross-checked against the existing forced-action-coverage forensic reconstruction, 30/30 seeds). No condition-specific hyperparameter, reward modification, or privileged input was found.

Formation. EARLY 30/30, LATE 30/30 (0.0-point gap, McNemar p=1.00, 0 discordant pairs). Continuous margin: mean paired difference -0.145, 95% CI $[-0.442, 0.152]$, Wilcoxon p=0.612, $d_z = -0.18$, 14/30 seeds negative and 16/30 positive. The binary criterion is saturated at ceiling in both orderings, but the continuous margin is also centered near zero with no consistent sign, so this is not merely a binary-ceiling artifact.

Learning-health checks. Policy entropy starts at $\approx \ln 6 = 1.79$ nats (correct for an untrained uniform policy) and declines smoothly in every seed; action distributions remain balanced across both harvest actions and all four movement actions, with no collapse to a single action; post-shift probe accuracy rises to ≈ 0.88 (EARLY) / ≈ 0.84 (LATE) by step 11200, with Reliance(B) overtaking Reliance(A) after the shift in both orderings, confirming the reliance/probe machinery remains numerically meaningful under PPO. EARLY and LATE track each other closely at every checkpoint across the full 12000-step run on every one of these diagnostics.

Interpretation. This is an absence of the early-evidence benefit seen under DQN, not evidence that PPO resists shortcut formation: PPO forms the shortcut under every ordering tested. We do not identify a causal mechanism for the boundary. One code-level difference consistent with it is that PPO discards on-policy transitions after their update, whereas DQN's replay buffer can resample earlier evidence throughout the rest of Phase I; this is a hypothesis rather than a causal explanation, and replay-augmented PPO and rollout-length controls were not tested. DQN and PPO also differ in on- versus off-policy updating and value- versus actor-critic learning, any of which could instead be responsible.

P ORGANIC, ACTION-DRIVEN VISITATION: FULL RECORD

This appendix gives the complete record behind the organic-visitation result summarized in Section 4.3. This is a distinct external-validity/generalization test, not a replication of the scripted anchor: Phase-I visitation is fully endogenous, so realized evidence exposure is an outcome of the run, not an experimental input.

Construction. `env.yoked_controller` is `None` for the entire run in every condition; the agent's chosen action determines its position at every Phase-I step, exactly as in the plain organic condition used elsewhere in this project. The same fixed pool of 9 planted cells used throughout the paper is retained; outside its condition's active window, a planted cell renders and behaves identically to any ordinary agreement cell (A=B), with no reward, transition, position, or controller difference from an unremarkable cell. EARLY's window is $[0, 400)$ and LATE's is $[3600, 4000)$, matching the anchor exactly; a matched ZERO arm never opens a window, so its planted cells are never distinguished from ordinary cells. Phase I is 4000 steps, Phase II 8000, and the formation criterion (mean of Rel(A)−Rel(B) over steps 3000–3800, strictly pre-shift) is unchanged. n=30 paired seeds (0–29), same seed-driven b-map, planted-cell identity, and initial online/target network state across all three arms.

Integrity. 90/90 runs completed, 0 failures. `env.yoked_controller is None` verified for the entire run in all 90 cases via a hard runtime assertion. Planted-cell identity (9 cells) matches exactly across EARLY/LATE/ZERO for every seed. Gated cells were independently confirmed, by direct inspection of the underlying feature map at window boundaries, to read as plain agreement immediately before their window opens and immediately after it closes, and as disagreement throughout the open window. No NaN/Inf, no anomalous training seed.

Formation. EARLY 0/30, LATE 30/30 (McNemar $p{=}1.86 \times 10^{-9}$, 30/30 discordant pairs, all in the same direction). Continuous margin: EARLY mean -3.986, LATE mean 2.378, mean paired difference 6.364, 95% CI $[6.095, 6.633]$, Wilcoxon $p{=}1.86 \times 10^{-9}$, $d_z{=}8.47$, all 30 seeds agreeing. ZERO: 30/30 formed, mean margin 2.531 (SD 0.701), a reference/sanity condition, not part of the primary EARLY-vs-LATE contrast.

Exposure (reconstructed directly from raw per-step trajectories, not from a stored running counter). EARLY: median 99.5 gated-cell visits (IQR $[81.5, 110.75]$, range $[64, 123]$) and median 33 harvests there (IQR $[23.25, 37.5]$, range $[10, 48]$), 0/30 seeds with zero exposure. LATE: median 26 visits (IQR $[1.25, 59.25]$, range $[0, 241]$) and median 23 harvests (IQR $[1, 45]$, range $[0, 210]$), 8/30 seeds with zero exposure (seeds 0, 9, 19, 21, 24, 26, 27, 28). Splitting LATE by realized exposure: the 8 zero-exposure seeds form 8/8 (margin mean 2.356, SD 0.471); the 22 nonzero-exposure seeds also form 22/22 (margin mean 2.386, SD 0.834); the two subgroups are statistically indistinguishable (Mann-Whitney $p{=}0.50$ on margins, Fisher exact $p{=}1.00$ on formation). The categorical result is therefore not driven mainly by the zero-exposure seeds, but exposure remains endogenous and unequal throughout, and all three of ZERO's margin (2.531), zero-exposure-LATE's (2.356), and nonzero-exposure-LATE's (2.386) sit close together, so this experiment does not isolate a quantitative timing-specific margin from the organic baseline.

Interpretation. The categorical dissociation survives fully organic, action-driven visitation. This is not exposure-matched: timing is the experimental factor, realized exposure is a downstream, endogenous consequence of it, and we do not claim the timing effect is exposure-independent. All 30 seeds are retained in every analysis above; none were excluded or reweighted by realized exposure.

Q REPLAY-NECESSITY ABLATION: FULL RECORD

This appendix gives the complete record behind the no-replay result summarized in Sections 4.2 and 5.

Construction. `agent.update()` – the Double-DQN target computation, Huber loss, Adam step, and gradient clip – is unmodified. The only change is what feeds it: instead of sampling a batch of 64 transitions uniformly with replacement from the full replay history (the standard construction, in which an early transition remains resamplable for nearly the entire run while a late one is not), each update is fed a batch of size 1 built solely from the transition just produced that step; `buffer.sample()` is never called for these conditions, so reuse of any past transition is impossible by construction, not merely empirically absent. Warmup (500 steps), update cadence (one update per step thereafter), target-sync schedule, epsilon schedule, evidence construction, and formation criterion are all unchanged. $n{=}30$ paired seeds.

Integrity. 60/60 runs completed, 0 failures. Every run performs exactly 11,501 updates (= total steps − warmup + 1), identical between EARLY and LATE for every seed; 345,030 unique transitions trained on in each arm (30 seeds × 11,501), one apiece, none repeated, verified both by a runtime assertion and by the update-step sequence being strictly increasing with no duplicates. No NaN/Inf in any run. Gradient-norm values above the clip threshold of 10 (1.8–24.6% of logged checkpoints, depending on seed) reflect `clip_grad_norm_` reporting the pre-clip norm under batch-size-1 gradient variance, not clipping failure; no run showed a runaway value (max across all runs 50.1) or unbounded loss (max 2.40).

Formation. EARLY 29/30, LATE 29/30, 0 discordant pairs, McNemar $p{=}1.00$. Continuous margin: EARLY $2.694{\pm}1.047$, LATE $2.490{\pm}1.166$, mean paired difference -0.204, 95% CI $[-0.542, 0.135]$, Wilcoxon $p{=}0.105$, $d_z{=}-0.22$, 19/30 seeds negative and 11/30 positive (no consistent direction). One seed (21) fails to form in both orderings, an idiosyncratic hard seed rather than a timing-driven discordance. For reference, the standard replay-based anchor on the same seeds gives EARLY 0/30, LATE 30/30.

Interpretation. Removing replay under this batch-size-1 intervention eliminates the timing split. Because batch size (64 to 1) and the associated gradient variance change simultaneously with the removal of replay, this is a replay/batch-size ablation, not a clean causal isolation of replay: we do not claim replay is uniquely or causally responsible for the anchor effect.

R LEARNING-STATE REWIND: FULL RECORD

This appendix gives the complete record behind the rewind result summarized in Section 5. We call this a *learning-state rewind*, not a plasticity reset, throughout: it is a mechanistic probe into accumulated learner state, not a demonstration of any specific biological or algorithmic plasticity construct.

Construction. At the start of global step 3600, before that step's epsilon draw, action selection, transition collection, or optimizer update: the online network is restored to that seed's exact original pre-training initialization; the target network is restored to the same initialization; the optimizer is replaced with a fresh Adam instance at the same learning rate. Preserved exactly: replay contents, pointer, size, and RNG; environment, controller, action, and evidence RNG streams; epsilon schedule and global step; update and target-sync counters. No extra environment step, optimizer update, or target sync occurs at the rewind. A 2×2 design crosses timing (EARLY/LATE) with rewind (on/off): E0/E1 (EARLY, no-rewind/rewind), L0/L1 (LATE, no-rewind/rewind), plus a matched ZERO-evidence control Z0/Z1, all at the same step-3600 rewind point. n=30 paired seeds per arm, 180 runs total.

Integrity. 180/180 runs completed, 0 integrity failures. For every run, hard runtime assertions confirm: online/target/optimizer state changes at the rewind for the rewind-enabled arms only (digest equality to the seed's original initialization, 0 optimizer parameter-state entries post-reset); all preserved state (above) is bit-identical immediately before and after the rewind for every arm, rewind-enabled or not; no extra update or target sync occurred. E0 and L0 (no-rewind) logs are independently verified byte-identical to the main DQN anchor's own logs for all 30 seeds, confirming this construction reproduces the anchor exactly when the rewind is disabled.

Formation. E0/E1 0/30, L0/L1 30/30, Z0/Z1 30/30: the rewind changes no seed's categorical outcome in any arm. Continuous margins: E0 -3.987, E1 -2.971, L0 2.185, L1 1.046, Z0 2.460, Z1 1.974. The pre-registered primary interaction, $(L1{-}L0) - (E1{-}E0)$: mean -2.155, 95% CI $[-2.270, -2.040]$, Wilcoxon p=1.86×10^{-9}, d_z= $-$ 6.71, 30/30 seeds negative.

Temporal decomposition. A forensic trajectory analysis (using the exact canonical probe at the rewind instant plus the regular 200-step checkpoint grid) shows the rewind's effect on LATE is not simply an immediate, generic disruption: at the instant of rewind, LATE's and ZERO's rewind effects are statistically indistinguishable (both a pure consequence of the weight reset, before any evidence is processed), but they diverge sharply during $[3600, 3800)$, the active part of LATE's evidence window. A direct paired contrast at step 3800, (L1$-$L0) $-$ (Z1$-$Z0): mean -3.267, 95% CI $[-3.668, -2.866]$, Wilcoxon p=1.86×10^{-9}, d_z= $-$ 2.92, 30/30 seeds negative. EARLY shows a large immediate disruption from the rewind (jump of $+3.99$) that fully washes out within ≈8000 post-shift steps, a qualitatively different, non-window-associated pattern.

Interpretation. This is consistent with accumulated learner state contributing to how late evidence is processed, beyond a generic consequence of resetting network weights. It does not establish a consolidation or plasticity mechanism: replay contents are preserved throughout, so the rewind changes network state while holding data history fixed, and cannot by itself isolate replay's role. Binary formation is entirely unaffected, so this evidence bears only on the continuous margin, not on the paper's primary categorical claim.

S CHECKPOINT-SENSITIVITY ANALYSIS: FULL RECORD

This appendix gives the full record behind the checkpoint-sensitivity check noted in Section 3.2, on the main DQN anchor's own n=30 paired seeds, recomputed directly from raw per-checkpoint logs rather than from an aggregated summary. This is an exploratory sensitivity analysis, not a replacement for the primary five-checkpoint formation criterion, and step 4000 (post-shift) is excluded from every criterion below, verified directly from each run's logged phase label.

Checkpoint set	EARLY	LATE	McN. p	Mean diff. [95% CI]	Wilc. p	d_z
Primary (3000/3200/3400/3600/3800)	0/30	30/30	1.86×10^{-9}	$6.172\ [5.931, 6.412]$	1.86×10^{-9}	9.18
Final checkpoint only (3800)	0/30	28/30	7.45×10^{-9}	$5.066\ [4.796, 5.335]$	1.86×10^{-9}	6.72
Final two (3600, 3800)	0/30	30/30	1.86×10^{-9}	$5.753\ [5.529, 5.978]$	1.86×10^{-9}	9.17

The categorical conclusion is unchanged under every alternative checkpoint definition: no seed ever reverses direction under any criterion, and every criterion gives $p < 10^{-7}$ with a large, consistent effect. The single-checkpoint (3800-only) variant is the one place the result is not numerically identical to the primary criterion: seeds 14 and 27 have a positive 5-checkpoint average margin ($+0.89$ and $+1.70$) but a negative single-checkpoint-3800 margin (-0.28 and -0.83), both borderline cases near zero, consistent with the 5-checkpoint average smoothing ordinary single-timestep noise rather than exposing a measurement artifact. The 5-checkpoint average was fixed before this sensitivity analysis and is retained as primary throughout the paper.

Table 3: Cross-paradigm and cross-environment dissociation, $n = 30$ paired seeds per row. Formation rate is the fraction of seeds meeting the formation criterion (Section 3.2). Gap is LATE minus EARLY formation rate in percentage points, a bounded (± 100), directly comparable magnitude across every row regardless of paradigm, architecture, or environment. d_z and the 95% CI, by contrast, are computed on the paired difference of $(\mathrm{Rel}(A) - \mathrm{Rel}(B))$, whose scale differs across paradigms/architectures/environments; see the footnotes for scale caveats. The last column is a direct paired interaction test against that row's own online-RL row within the same architecture/environment.

Setting	Paradigm	EARLY	LATE	Gap	McN. p	d_z	95% CI	Int. p
CNN (corrected)	online RL	0/30	30/30	100.0	1.86×10^{-9}	-9.18	$[-6.41, -5.93]^{\dagger}$	—
CNN	fixed-data SGD	17/30	18/30	3.3	1.00	0.08	$[-0.09, 0.14]$	1.86×10^{-9}
CNN	fixed-data TD	24/30	26/30	6.7	0.50	-0.27	$[-0.07, 0.01]$	1.86×10^{-9}
CNN	PPO (on-policy)[¶¶]	30/30	30/30	0.0	1.00	-0.18	$[-0.44, 0.15]$	—
TinyViT (corrected)	online RL	0/30	30/30	100.0	1.86×10^{-9}	-7.12	$[-5.52, -4.99]^{\ddagger}$	—
TinyViT	fixed-data SGD	16/30	17/30	3.3	1.00	0.30	$[-0.01, 0.07]$	1.86×10^{-9}
Corridor	online RL	3/30	30/30	90.0	1.49×10^{-8}	-4.86	$[-5.06, -4.37]^{\S}$	—
Corridor	fixed-data SGD	20/30	17/30	-10.0	0.25	0.65	$[0.25, 0.86]$	1.86×10^{-9}
CNN (20-visit)	online RL	0/30	30/30	100.0	1.86×10^{-9}	-8.59	$[-6.40, -5.89]^{\P}$	—
CNN (interior window)	online RL	0/30	22/30	73.3	4.77×10^{-7}	-5.13	$[-4.75, -4.13]^{\|}$	—
CNN (forced action-coverage)	online RL	0/30	30/30	100.0	1.86×10^{-9}	-6.48	$[-6.29, -5.63]^{\#}$	—
CNN (matched integration time)	online RL	0/30	15/30	50.0	1.86×10^{-9}	-2.80	$[-4.43, -3.43]^{*}$	—
CNN (flat exploration schedule)	online RL	3/30	30/30	90.0	1.49×10^{-8}	-2.13	$[-5.45, -3.88]^{**}$	—
CNN (no-replay, batch size 1)	online RL	29/30	29/30	0.0	1.00	-0.22	$[-0.54, 0.14]^{\dagger\dagger}$	—

[†]Corrected construction (Section 4.1, Appendix G); the originally run construction gave $d_z = -9.77$, mean paired difference -6.21, 95% CI $[-6.44, -5.99]$, and its own originally reported effect size on plain $\mathrm{Rel}(A)$ rather than this margin is $d_z = -1.12$ (final $\mathrm{Rel}(A)$) to -1.86 (AUC), Holm-corrected $p \leq 3.4 \times 10^{-8}$. All describe the same effect at different points in its correction history and should not be compared to each other as if on the same scale.

[‡]Corrected construction, independently reran with dedicated evidence/agreement RNG streams (Appendix G); the originally run construction gave $d_z = -11.05$, mean paired difference -5.44, 95% CI $[-5.61, -5.26]$. CI on the mean paired difference; not compared in magnitude to the CNN row's d_z across architectures with different Reliance scales.

[§]Corridor Foraging (Appendix K); Reliance is defined identically in construction but on a different observation modality, so its scale is not compared in magnitude to the grid task's rows.

[¶]20 rather than 40 evidence-cell visits in the same extreme windows as the CNN anchor row, same corrected construction (Section 4.3); directly comparable in scale to the CNN (corrected) row.

[‖]EARLY and LATE columns are EARLY-MID $[800, 1200)$ and LATE-MID $[2800, 3200)$, not the anchor's extreme windows (Section 4.3); directly comparable in scale to the CNN (corrected) row.

[#]Identical, exact two-sided `HARVEST0`/`HARVEST1` action sequence forced at the 40 evidence steps for both orderings (Section 4.3); a paired-by-seed comparison against the CNN (corrected) row's own mean paired difference finds a small but real reduction (mean diff. -0.22, 95% CI $[-0.33, -0.10]$, paired Wilcoxon $p = 0.001$, $\approx 3.5\%$ of that row's magnitude), driven entirely by LATE.

[*]LATE's Phase I extended to 8000 steps with the same $[3600, 4000)$ evidence window (`late_t8000_rngfix`), giving LATE 4000 post-evidence integration steps, closely approximating (not exactly matching) EARLY's 3600; corrected, dedicated-RNG construction, independently re-verified against the shared-RNG original (Appendix C). Relative to the raw anchor row's own mean paired difference, this reduces the raw contrast by ≈36%; not directly comparable in scale to rows sharing the anchor's mismatched 4000-step LATE window.

[**]Exploration schedule replaced with a flat $\epsilon = 0.1$ for the entire run (both orderings), rather than annealing 1.0→0.02 (Section 4.3, Appendix D). A paired-by-seed comparison against the natural-schedule anchor's own mean paired difference finds a real reduction (≈24% of that row's magnitude, $p = 0.00034$); this is the only condition in this table where EARLY is not at 0/30.

[¶¶]Matched on-policy PPO actor-critic, identical evidence sequence and construction, one un-tuned configuration for both orderings (Appendix O); a boundary condition, not another value-based operating point – PPO forms the shortcut under both orderings, so this row's 0.0 gap reflects an absent early-evidence benefit, not robustness against formation.

[††]Experience replay removed entirely: each update trains on the single just-produced transition (batch size 1, 11,501 updates, zero transition reuse) rather than a replay-sampled batch of 64, with the Double-DQN target and optimizer otherwise unchanged (Appendix Q). Mean paired difference -0.204, Wilcoxon $p = 0.105$; this is the only condition in this table where LATE is not categorically formed. Batch size and gradient variance change simultaneously with replay removal, so this row is a replay/batch-size ablation, not evidence that replay alone is causal.

Table 4: Full numerical record of the elimination and bounded-probe experiments summarized in Section 4.2, numbered per Appendix A. All $n = 30$ paired seeds.

Experiment	Key contrast	Result
Replay manipulation (Exp. 8)	boosted LATE (147.9 samples/item) vs. natural LATE	100% vs. 100%, McNemar Holm-$p = 1.0$
Replay manipulation (Exp. 8)	throttled EARLY (2.94 samples/item) vs. natural EARLY	33.3% vs. 0.0%, McNemar Holm-$p = 0.0059$
Environment experience vs. updates (Exp. 10)	update-matched (4804 updates), short-experience/4× density vs. long-experience/1× density	6.7% vs. 90.0%, McNemar Holm-$p = 1.25 \times 10^{-6}$, $d_z = -1.39$
Consolidation timing (Exp. 9, bounded/confounded)	duration-matched (window = 3100), PRIOR 2300/POST 400 vs. PRIOR 900/POST 1800	70.0% vs. 0.0%, McNemar Holm-$p = 7 \times 10^{-6}$
Shift-timing probe (Exp. 7)	LATE evidence, shift lead time $t = 4000$ vs. $t = 8000$, matched `zero` controls	100% vs. 53.3%, Holm-$p = 0.00049$; trend across t, Cochran's $Q = 34.38$, $p = 1.6 \times 10^{-7}$

Table 5: Post-evidence integration time and formation across the lead-time family. $\text{LATE}_{t=8000}$(orig.) is the originally run, non-rngfix construction; $\text{LATE}_{t=8000}$(rngfix) is the corrected confirmatory re-run reported as the main-text matched-integration-time result. All other LATE_t points and both ZERO_t controls use the originally run construction only.

Condition	Integration time	Phase-I length	Formation
EARLY (corrected)	3600	4000	0/30 (0%)
$\text{LATE}_{t=4000}$ (orig.)	0	4000	30/30 (100%)
$\text{LATE}_{t=4400}$ (orig.)	400	4400	29/30 (96.7%)
$\text{LATE}_{t=5200}$ (orig.)	1200	5200	19/30 (63.3%)
$\text{LATE}_{t=8000}$ (orig.)	4000	8000	16/30 (53.3%)
$\text{LATE}_{t=8000}$ (rngfix, corrected)	4000	8000	15/30 (50.0%)
$\text{ZERO}_{t=4000}$ (no evidence)	n/a	4000	30/30 (100%)
$\text{ZERO}_{t=8000}$ (no evidence)	n/a	8000	30/30 (100%)

Table 6: Full supervised $K \times$lr grid, CNN fixed-dataset control, grid task. Gap is LATE minus EARLY formation rate in percentage points.

K	lr	EARLY %	LATE %	gap (pts)	mean diff.	d_z
1	10^{-4}	100.0	100.0	+0.0	−0.012	−0.27
1	3×10^{-4}	100.0	100.0	+0.0	+0.449	1.31
1	10^{-3}	100.0	100.0	+0.0	+1.682	0.98
2	10^{-4}	100.0	100.0	+0.0	−0.074	−0.66
2	3×10^{-4}	26.7	33.3	+6.7	−0.229	−0.40
2	10^{-3}	6.7	6.7	+0.0	−4.260[†]	**−1.75**
4	10^{-4}	56.7	60.0	+3.3	+0.026	0.08
4	3×10^{-4}	0.0	0.0	+0.0	+0.282	0.56
4	10^{-3}	0.0	0.0	+0.0	−3.029	−0.87

[†]This cell's raw mean-difference magnitude (−4.260) reflects this grid's own within-cell variance and is not compared in scale to raw margins reported for other paradigms, architectures, or environments elsewhere in this paper; only the formation-rate gap and each cell's own d_z are used for cross-paradigm comparison.

Table 7: RL formation-rate gap and $|d_z|$ at each CNN grid-task RL operating point tested in this section's controls (not every RL condition in the paper – TinyViT and Corridor RL are reported separately in Table 3), against the supervised grid's own best case.

Condition	Formation-rate gap (pts)	$\lvert d_z \rvert$
RL, natural anchor	100.0	9.18
RL, forced action-coverage	100.0	6.48
RL, flat exploration schedule	90.0	2.13
RL, matched integration time	50.0	2.80
Supervised, best case over 9-cell grid	6.7	1.75

Table 8: Move-action/scripted-position mismatch rate, full $n = 30$ batch, both orderings.

	EARLY ($n = 30$)	LATE ($n = 30$, $n = 28$ at evidence steps[†])
Overall mismatch rate	97.19% (sd 0.43%)	97.19% (sd 0.43%)
Mismatch rate at evidence steps	97.89% (sd 2.60%)	98.40% (sd 3.02%)
Mismatch rate at agreement steps	97.17% (sd 0.44%)	97.18% (sd 0.43%)
Move fraction of Phase I	26.71%	27.17%

[†]Two LATE seeds had zero move actions recorded at their scheduled evidence steps (all-harvest), giving an undefined rate; excluded, not zero-filled.