RICK: CERTIFYING TRANSMISSION IN MULTI-AGENT CREDIT ASSIGNMENT VIA WORST-REGIME SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A cooperative team trained from one shared reward learns exactly what its credit assignment tells it, so the kernel that decides which agent earned a delayed outcome is itself the training signal. However, that kernel is fitted to behaviour logs whose releases and arrivals share one exogenous driver, and every quantity that certifies it, the fit residual, the perturbation schedule and the conservation percentage, is read off those same logs, so the evidence certifies itself: a kernel can leave only 0.0261 of the recorded variance unexplained while its coefficients sit 0.4541 from the truth, and no reading inside the log separates the two. We propose the Regime-Invariant Credit Kernel (RICK), which selects the kernel on intervention regimes the team manufactures and pays for. Regime-invariant kernel selection pins the kernel by its worst per-regime discrepancy instead of its aggregate fit. A regime scheduler commits the perturbation family under a bound on the return it costs, and a stop-gradient keeps team return out of the selection. Delayed backtracking and share accounting states the arrived mass a bounded window cannot reach as an unassigned share. Together, the three modules replace each self-certifying readout with one that can fail: a kernel propped up by the shared driver leaves a manufactured regime poorly explained, and the mass a window misses stays on the books. Scored against exact ground truth over ten draws per cell on three delayed environments, RICK recovers the generating coefficients on the fan-in environments to a mean of 0.0935, against 0.2186 for the pooled aggregate fit at the same recording budget. On a structurally distinct held-out family it keeps that ordering, at 0.0539 against 0.1151 for the pooled aggregate fit.

1 INTRODUCTION

A cooperative team trained from one shared reward learns exactly what its credit assignment tells it (Pignatelli et al., 2024; Foerster et al., 2018). The kernel that performs it decides which agent, and which of that agent's earlier moves, earned an outcome arriving many steps later along tangled lines of influence. Because credit is the training signal rather than a diagnostic beside it, a kernel that attributes an arrival to the wrong emitter does not merely misreport: it trains the team to repeat the behaviour that did not earn the reward. The field therefore treats the kernel as an object to be estimated well, and measures how well by how closely it reproduces the propagation its own logs display (Zhou et al., 2020; Su et al., 2020; Dou et al., 2022). What a kernel is worth then rests on whether the quantity judging it can separate a correct kernel from a confident one.

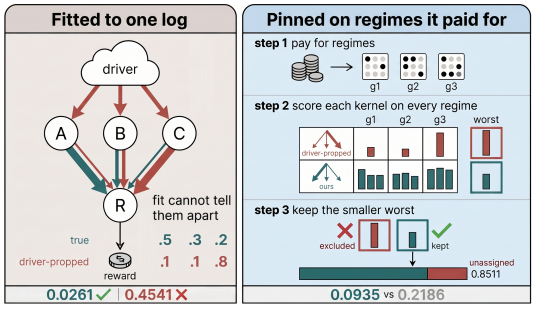


Figure 1: **Left:** fitted to one log, a kernel cannot tell transmission from a shared driver: 0.0261 unexplained, coefficients 0.4541 from the truth. **Right:** RICK pins the kernel on regimes it paid for, reaching 0.0935 against 0.2186.

Three lines of work supply that judgment, and each has made the kernel better. Value decomposition factors the shared return into per-agent terms fitted on the recorded joint play, and graph-structured

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.

and attention-based mixers recover propagation wherever the recorded signal is informative (Zhou et al., 2020; Dou et al., 2022; Naderializadeh et al., 2020; Zohar et al., 2022; Zhao et al., 2023). Counterfactual and game-theoretic estimators perturb the agents instead, scoring each by the change a replaced action induces, with model-free baselines and Shapley shares making that counterfactual computable at scale (Foerster et al., 2018; Mesnard et al., 2021; Su et al., 2020; Li et al., 2021; Ji et al., 2025). Temporal credit assignment redistributes a delayed return over the timesteps that produced it, using sequence models, inference networks and agent-time attention (Pignatelli et al., 2024; Liu et al., 2019; Ausin et al., 2021; She et al., 2022; Chen et al., 2023). Between them they cover the three axes on which a delayed team must be scored: who transmitted, how much, and when. Each, however, reads its answer from the record the agents already produced.

That record is also the record that hides the answer. Upstream releases and downstream arrivals are driven by the same autocorrelated exogenous injections, leaving the parents of a receiving point collinear in any observational log (Juliani et al., 2026; Li et al., 2025a). A kernel fitted there accounts for all but 0.0261 of the variance and is still 0.4541 from the coefficients that produced that log (Figure 1, left). Breaking that collinearity requires interventions the team applies to itself and pays for out of its own return, and what they buy at a given cost is decided by which direction they excite. Tracing the arrival back to the timestep that earned it exceeds the shortest window, and the accounting then divides by the mass the window reached and prints full conservation while 0.8239 was never reached. The three are faces of one failure: every quantity that certifies a credit kernel is computed from the record the kernel was fitted to, and each reports success by construction.

To address these three challenges, we propose RICK, a credit kernel selected on interventions the learner manufactures and pays for rather than on a fit to the logs it was handed. The insight is to replace every self-certifying readout with one that can fail: the kernel is pinned by the regime under which it loses most rather than by the aggregate fit it minimises on average, the schedule is charged against the return it costs rather than fixed in advance, and the conservation denominator is the mass an unbounded expansion carries rather than the mass the window happened to reach. A kernel propped up by the shared driver then leaves a manufactured regime poorly explained, and the return a bounded window does not reach is stated as an unassigned share rather than normalised away. Three coupled modules implement this over a training batch, one per challenge: regime-invariant kernel selection, a regime scheduler under a return-cost bound, and delayed backtracking and share accounting. Our contributions are as follows:

- We name the evidence certifying a credit kernel self-certifying: the fit residual, the perturbation schedule and the conservation percentage are all computed from the record the kernel was fitted to, and each reports success by construction.
- We propose RICK, which pins the kernel by its worst per-regime discrepancy under a return-cost-bounded perturbation family, and states the arrived mass a bounded window cannot reach as an explicit unassigned share.
- RICK recovers the generating per-edge transmission coefficients to a mean of 0.0935 on the environments that carry fan-in competition, against 0.2186 for the pooled aggregate fit at the same recording budget, establishing that a denser fit to recorded logs is not evidence that influence was transmitted.

2 MOTIVATION

To understand why a kernel that fits behaviour logs certifies no transmission, we examine fit, excitation and reach on held-out records, revealing three challenges for delayed credit assignment.

Challenge #1: Confounded fit. Credit kernels are fitted to behaviour logs on the assumption that reproducing recorded propagation recovers how influence travels. We fit a kernel to an observational log alone and read it twice, against itself and against the transmission that produced the log. Figure 2 (left) reads both readings across the driver's autocorrelation: the fit leaves at most 0.0261 of the variance unexplained while the coefficients it recovered sit 0.4087 to 0.5039 from the truth. The parents of a receiving point are nearly collinear there, leaving observational fit alone unable to exclude the driver-propped kernel. *How can a kernel certify that the influence it credits was transmitted rather than merely shared?*

Challenge #2: Unpriced identification. The recourse is to perturb the agents and read how propagation responds; the team pays for those perturbations, yet what a budget buys is assumed rather

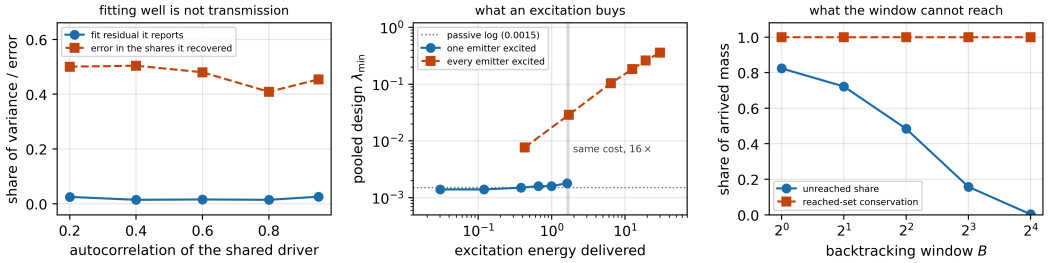


Figure 2: (**Left**) a good fit recovers wrong coefficients. (**Middle**) placement, not size, buys identification. (**Right**) a reached-set denominator hides the gap. Details in Appendix A.2.

than measured. We swept two excitations over seven strengths, one reaching a single emitter and one reaching every emitter; Figure 2 (middle) reads what each bought: reaching one emitter stays within 0.0003 of the identification the log already carries at an endpoint delivering 1.6194, while reaching all emitters reads $16\times$ higher at the comparable energy of 1.7119; identification is bought along the weakest direction of the pooled parent design while cost is charged for the whole excitation. *How can perturbations be chosen to buy separation with every unit of return they cost?*

Challenge #3: Truncated conservation. The arrival must also be traced back along directed paths to the timestep that emitted it, and with sparse rewards and state-dependent edge delays the lag before arrival is neither fixed nor small. We compressed the backtracking budget below the longest edge lag and expanded with the environment's kernel, with nothing fitted; Figure 2 (right) traces the mass no path inside the budget reaches. At the shortest window 0.8239 of that mass lies outside, while a percentage over the reached set alone prints full conservation at every rung. That denominator turns an unavoidable shortfall into a report no reading inside the window can contradict. *How can the mass the budget cannot reach be surfaced rather than normalised away?*

3 Problem Formulation

To address the three challenges of Section 2, we formulate delayed multi-agent credit assignment as a bilevel program over policies, self-applied intervention schedules and credit kernels:

$$\max_{\pi,\ \sigma\in\mathcal{S}} J(\pi;\Psi_\sigma) \qquad \text{s.t.} \qquad \Psi_\sigma \in \arg\min_{\Psi\in\mathcal{K}} \max_{g\in\sigma} D_g(\Psi) \tag{1}$$

where π is the execution policy, σ a schedule of intervention regimes the team applies to itself, J the team return, $\mathcal{S}$ the schedules whose return cost stays inside the bound the deployer affords, $\mathcal{K}$ the kernels whose reach is confined to a bounded ancestry and which carry the mass that reach misses as an explicit unassigned share, and $D_g(\Psi)$ the discrepancy between the propagation Ψ implies and the propagation regime g records.

The inner level addresses Challenge #1: read on the worst regime of a family that excites the collinear direction at every receiving point, it admits in the population limit no kernel a shared driver props up. The class $\mathcal{S}$ addresses Challenge #2: what a schedule identifies is bounded by the deployer's spend. The class $\mathcal{K}$ addresses Challenge #3: unreached mass stays in the denominator.

4 Method

4.1 Overall Framework

RICK, shown in Figure 3, instantiates the bilevel program of Eq. 1 by turning propagation recorded under a self-applied perturbation family into a per-agent, per-timestep credit map through three modules: the regime scheduler commits a family σ under the cost bound the deployer affords, regime-invariant kernel selection fixes Ψ_σ on the worst regime rather than the pooled record, and delayed backtracking and share accounting carry the mass a bounded reach misses as unassigned.

The order is forced rather than chosen: a kernel can only be certified against regimes that have already been paid for, and credit can only be read through a kernel that has already been certified; each module removes the failure the next would otherwise inherit. Because selection runs on the family the scheduler committed, the family a kernel is judged on is the family it is later spent on.

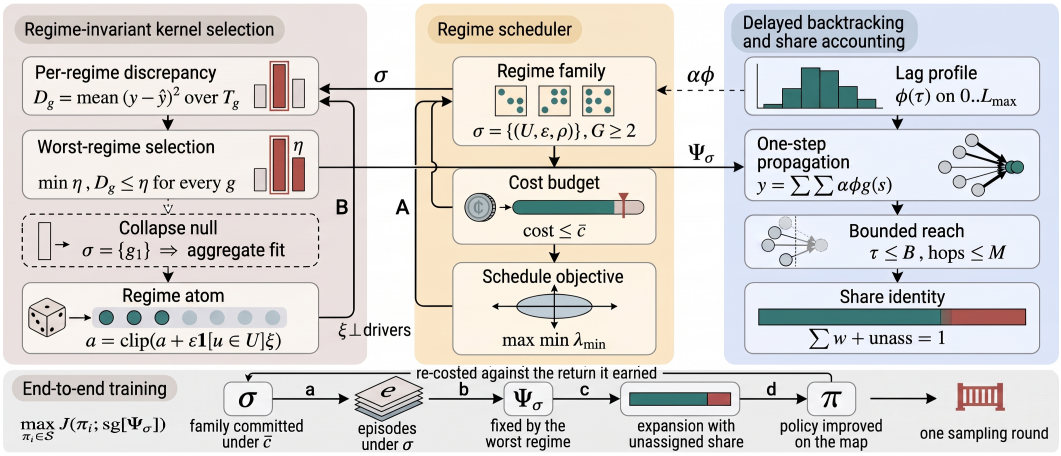


Figure 3: (**Left**) regime-invariant kernel selection. (**Middle**) regime scheduler. (**Right**) delayed backtracking and share accounting. (**Bottom**) end-to-end training.

4.2 Regime-invariant kernel selection

To address Challenge #1 of Section 2, regime-invariant kernel selection (Figure 3, left) fixes Ψ_σ on what each regime records rather than on the pooled log, through the per-regime discrepancy, a kernel pinned on the worst regime, the collapse a single regime induces, and the regime atom.

Per-regime discrepancy. The discrepancy of a regime is defined as the mean squared gap between the kernel's one-step prediction and the recorded receiving values on that regime's own points, building on the counterfactual credit-assignment view of Foerster et al. (2018).

$$D_g(\Psi) \;=\; \frac{1}{|T_g|} \sum_{(v,t)\in T_g} \big(\hat{y}_{vt}(\Psi) - y_{vt}\big)^2 \tag{2}$$

where T_g is the set of receiving points recorded under regime g, y_{vt} the value recorded at (v, t), and $\hat{y}_{vt}(\Psi)$ the propagation Ψ implies there.

Worst-regime selection. The inner problem is resolved in its epigraph form: a minimised level η upper-bounds every regime's discrepancy, which ties each admitted kernel to the hardest regime rather than the easiest.

$$(\Psi_\sigma, \eta_\sigma) \in \underset{\Psi\in\mathcal{K},\ \eta\in\mathbb{R}}{\arg\min}\ \eta \qquad \text{s.t.} \qquad D_g(\Psi) \le \eta \ \text{ for every } g \in \sigma \tag{3}$$

where η is the worst-regime discrepancy level and $D_g(\Psi)$ is evaluated per regime over its own receiving points T_g. A kernel that merely tracks the shared driver leaves a manufactured regime poorly explained, raising its η above that of a kernel that has captured transmission. The exclusion holds in the linear one-step population limit for a realisable kernel, orthogonal per-regime residuals and a common noise floor, on a family exciting the collinear direction at every receiving point.

Collapse null. The minimisation runs over a family the scheduler keeps distinct, because a single regime returns the criterion to the single-regime fit the logs already carry.

$$\sigma = \{g_1\}: \qquad \min_{\Psi\in\mathcal{K}}\ \max_{g\in\sigma} D_g(\Psi) \;=\; \min_{\Psi\in\mathcal{K}} D_{g_1}(\Psi) \tag{4}$$

where the right-hand side is the single-regime fit of the one available regime.

Regime atom. Each regime is a node mask paired with a perturbation strength, and the perturbation is drawn from a separate noise stream, which makes a masked action a genuine intervention rather than a reweighting of recorded behaviour.

$$a_u^t = \text{clip}\big(\tilde{a}_u^t + \varepsilon_g\, \mathbb{I}[u \in U_g]\, \xi_u^t\big), \qquad \xi_u^t \sim \mathcal{N}(0,1), \qquad \xi \perp \text{drivers} \tag{5}$$

where $\tilde{a}_u^t$ is the base action, U_g the mask and ε_g the strength of regime g, clip the projection onto the admissible action range, and $\perp$ records that the perturbation draw is independent of the drivers. This independence lets the criterion attribute a residual to transmission (see Appendix A.3).

4.3 REGIME SCHEDULER

To address Challenge #2, the regime scheduler (Figure 3, middle) commits σ under a return-cost bound: atoms compose into a family, the family is priced against the bound the deployer affords, and Eq. 9 measures the identification a family buys.

Regime family. A family of regimes is selected rather than a single perturbation, because one regime collapses the inner selection back to the single-regime fit; the family is the finite set of mask, strength and weight triples committed for a sampling round.

$$\sigma = \{(U_g, \varepsilon_g, \rho_g)\}_{g=1}^{G}, \qquad G \geq 2 \tag{6}$$

where U_g is the mask, ε_g the strength, ρ_g the weight regime g carries in the cost of Eq. 7 and in the pooled fit, and G the number of regimes; more than one regime is what makes the worst-case criterion of Eq. 3 informative.

Cost budget. Each family is charged the return it costs. The cost of a schedule is the return its perturbed regimes lose, weighted by ρ_g.

$$\mathrm{cost}(\sigma) = \frac{1}{|R(\text{unperturbed})|} \sum_{g=1}^{G} \rho_g \big(R(\text{unperturbed}) - R(g)\big) \tag{7}$$

A schedule is admissible when that cost stays inside the bound the deployer affords and the family it commits keeps at least two masks that differ, giving the feasible class

$$\mathcal{S} = \Big\{\sigma :\ \mathrm{cost}(\sigma) \leq \bar{c},\ G \geq 2,\ \exists\, g, g' :\ U_g \triangle U_{g'} \neq \emptyset\Big\} \tag{8}$$

where $R(g)$ is the paired unperturbed sink value net of the squared departure regime g causes from it, $\bar{c}$ the return-cost bound as a fraction of $|R(\text{unperturbed})|$, $\triangle$ the symmetric difference between masks, and $\mathcal{S}$ the class the outer level of Eq. 1 optimises over.

Schedule objective. The feasible class fixes what a family may cost but not which family to draw from it, and the scheduler carries an objective over $\mathcal{S}$: identification is bought along the direction in which a receiving point's parents are least separated, and the family is therefore the one that raises the smallest eigenvalue of the pooled parent design the excitation induces.

$$\sigma^{\star} \in \arg\max_{\sigma \in \mathcal{S}} \ \min_{v}\ \lambda_{\min}\big(\bar{\Sigma}_v(\sigma)\big), \qquad \bar{\Sigma}_v(\sigma) = \sum_{g=1}^{G} \rho_g\, \mathbb{E}_g\big[x_v x_v^{\top}\big] \tag{9}$$

where x_v collects the parent releases of a receiving point v. The main tables deal the masks by index round-robin, and a greedy search against a standardised proxy for this design is read beside it; a shared strength is bisected on the family's return cost and rescaled per mask; the strength therefore follows the behaviour in force rather than an outside setting (see Appendix A.4).

4.4 DELAYED BACKTRACKING AND SHARE ACCOUNTING

To address Challenge #3, that a bounded budget cannot reach every emitter, delayed backtracking and share accounting (Figure 3, right) expand the kernel under it, through a per-edge lag profile, one-step weights, a truncation to bounded reach, and the rest accounted as unassigned.

Lag profile. Each edge's lag profile is a normalised increment of the Gamma distribution function rather than its density, which keeps the profile finite at every integer lag the accounting visits, including the zero lag that synchronous edges take.

$$\phi_e^t(\tau) = \frac{F(\tau + 1; k_e^t, \vartheta_e^t) - F(\tau; k_e^t, \vartheta_e^t)}{\sum_{\tau'=0}^{L_e^{\max}} \big(F(\tau' + 1; k_e^t, \vartheta_e^t) - F(\tau'; k_e^t, \vartheta_e^t)\big)} \tag{10}$$

The profile says when a release arrives; how much of it arrives is a separate quantity, carried by the transmission share

$$\alpha_e^t = c_e\, \mathrm{sigmoid}\big(f_{\alpha}(z_e, h_u^t)\big) \tag{11}$$

where F is the Gamma distribution function on the integer lag grid, k_e^t and ϑ_e^t its shape and scale from f_ϕ, $L_e^{\max}$ the lag bound, and α_e^t the transmission share, whose state-dependent shape comes from f_α over edge features z_e and the state h_u^t of the emitter u of e, and whose free parameter is the transmission coefficient $c_e \in [0, 3]$, one per edge; the share is not bounded by one (see Appendix A.5).

One-step propagation. The one-step propagation the kernel predicts spends exactly this share and profile over the parents and admissible lags.

$$\hat{y}_{vt}(\Psi) = \sum_{u \in \mathrm{Pa}(v)} \sum_{\tau=0}^{L_{u\to v}^{\max}} \alpha_{u\to v}^{t-\tau}\, \phi_{u\to v}^{t-\tau}(\tau)\, g_\psi(h_u^{t-\tau}) + b_\psi(h_v^t) \tag{12}$$

where $\mathrm{Pa}(v)$ is the parent set of v, g_ψ the identically positive release, and b_ψ the state-dependent baseline; the same α and ϕ are spent here and in the path expansion, which selects the kernel on the quantity the accounting later spends.

Bounded reach. The reachable emitters are those inside the backtracking window and ancestor depth, and the credit map therefore reports mass only where the budget can carry it.

$$\mathcal{R}(j, s) = \big\{(i, s - \tau') : 0 \le \tau' \le B,\ \mathrm{hops}(i \to j) \le M\big\} \tag{13}$$

where B is the backtracking window, M the ancestor depth, τ' the lag to (j, s), and $\mathrm{hops}(i \to j)$ the directed hops from i to j.

Share identity. A forward topological expansion of Eq. 12 leaves every non-ancestor term identically zero by construction, and the unassigned share is taken against the reference-budget denominator, which yields an exact conservation identity, where a truncated-set denominator would instead report full conservation at every budget.

$$\sum_{(i, s-\tau')} w_{i,\, s-\tau' \to j, s} \;+\; \omega_\emptyset^{j,s} \;=\; 1 \qquad \text{for every receiving point } (j, s) \tag{14}$$

where $w_{i,\, s-\tau' \to j,s}$ is the weight carried from $(i, s - \tau')$ to (j, s) inside the budget, $\omega_\emptyset^{j,s}$ the unassigned share, and the sum runs over the reachable emitters. Because g_ψ is identically positive, the prefix weights are nonnegative, and a budget that under-reaches therefore raises $\omega_\emptyset$ instead of renormalising the shortfall away.

4.5 End-to-End Integration

End-to-end training. The three modules run as one loop per sampling round (Figure 3, bottom): the scheduler commits a feasible family under the cost bound $\bar{c}$, episodes are collected under it, Ψ_σ is fixed by the worst-regime criterion of Eq. 3, and the credit map comes from the expansion stage of Algorithm 1 with its unassigned share. The policy is then improved on that map while the schedule is re-costed against the return it earned, closing the loop of Eq. 1. Details in Appendix A.6.

Return decoupling. Selecting the kernel by the return it produces would reward piling credit where return is already high. The two levels are held apart by a stop-gradient between team return and the discrepancy-selected kernel, refining the outer level of Eq. 1 to $\max_{\pi,\, \sigma \in \mathcal{S}} J(\pi; \mathrm{sg}[\Psi_\sigma])$, where $\mathrm{sg}[\cdot]$ holds the selection of Ψ_σ fixed with respect to the outer gradient of J; raising return therefore never leaks into kernel selection. More details are provided in Appendix A.7.

5 Experiments

5.1 Implementation Details

Environments. Estimators are scored against exact ground truth on three controlled delayed environments, with a supply-tree family held out (Jia et al., 2024). Details in Appendix A.8.

Baselines. We contrast five runnable estimators from the temporal credit-assignment literature (Pignatelli et al., 2024): an aggregate-fit kernel, a uniform per-agent split, a concurrent-return counterfactual difference (Foerster et al., 2018), a time-decayed return redistribution (Chen et al., 2023), and the regime-collapsed $G{=}1$ null, all scored against the counterfactual-replay ground truth (Mesnard et al., 2021) on the same three environments and three edge-lag scales. Details in Appendix A.9.

Table 1: Coefficient error ↓, agreement with the reference kernel at the deployer budget ↑, unassigned share ↓. Dashes mark quantities undefined without a fitted kernel. Bold marks this work.

System	Coef. err. ↓	Agr. / true kernel at budget ↑	$\omega_\emptyset$ ↓
Cascaded production line			
Time-decayed return redistribution	–	0.8379	–
Concurrent-return counterfactual difference	–	0.5509	–
Uniform per-agent split	–	0.7193	–
Regime-collapsed (G=1) null	0.2195	0.9489	0.1692
Aggregate-fit kernel	0.2136	0.9247	0.1690
RICK (ours)	**0.0929**	**0.9994**	**0.1696**
Tidal road traffic			
Time-decayed return redistribution	–	0.1670	–
Concurrent-return counterfactual difference	–	0.0779	–
Uniform per-agent split	–	0.1237	–
Regime-collapsed (G=1) null	0.0274	0.9986	0.1317
Aggregate-fit kernel	0.0475	0.9975	0.1315
RICK (ours)	**0.0258**	**0.9992**	**0.1316**
Multi-level cache			
Time-decayed return redistribution	–	0.5329	–
Concurrent-return counterfactual difference	–	0.4686	–
Uniform per-agent split	–	0.4488	–
Regime-collapsed (G=1) null	0.1858	0.9391	0.0931
Aggregate-fit kernel	0.2236	0.9391	0.0928
RICK (ours)	**0.0942**	**0.9729**	**0.0928**

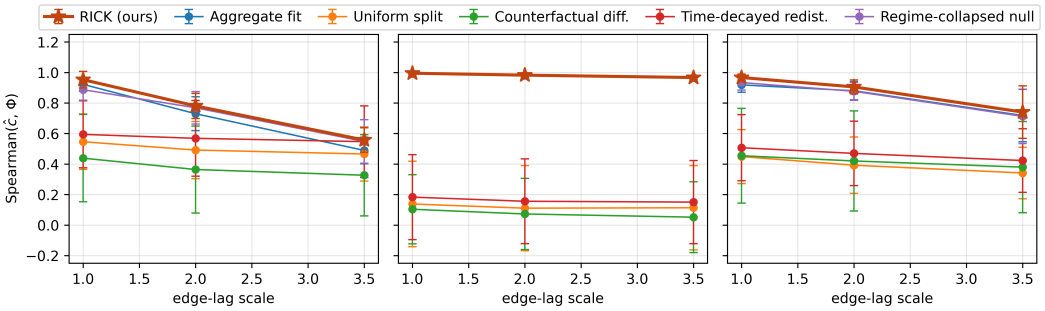


Figure 4: (**Left**) cascaded production line. (**Middle**) tidal road traffic. (**Right**) multi-level cache. Credit fidelity against the edge-lag scale for six estimators, with across-draw error bars.

Evaluation metrics. Estimators are scored on the transmission-coefficient error ↓, credit fidelity $\rho(\hat{c}, \Phi)$ ↑ and the unassigned share $\omega_\emptyset$ ↓. The coefficient error is the primary reading, the quantity the confounding acts on directly; credit fidelity is a rank correlation and compresses what that error resolves; the unassigned share reads what the window leaves unreached. Details in Appendix A.10.

Budgets. The kernel is expanded under a deployer-afforded window and ancestor depth, and the schedule is charged against a return-cost bound. The unassigned share is taken against the unbounded expansion on the same environment, not the reached set. Details in Appendix A.11.

Compute. Reported wall clock is the expansion time of one credit map, measured over the full ancestor set, which makes the reference mass exact, with the window and ancestor depth applied to the result, bounding the credit reported rather than the work done. Details in Appendix A.12.

Hyperparameters. By default the family carries $G = 4$ regimes at mixing weights $0.90/0.10$, the recording budget is 16 episodes, the conservation tolerance is 0.05, and $c_e \in [0, 3]$. Main cells average ten draws, ablation rows one draw, return-path rows five seeds. Details in Appendix A.13.

5.2 Main Results

Coefficient recovery. Table 1 reports the six estimators on the quantity the confounding acts on directly, the per-edge transmission coefficients that generated the log. Over the two environments that carry fan-in competition the worst-regime criterion recovers them to a mean of 0.0935, against 0.2186 for the aggregate-fit kernel and 0.2026 for the regime-collapsed null read at the same sixteen episodes. Both comparators carry more than twice this work's coefficient error on average, and tidal road traffic keeps the ordering. On the fan-in environments the collapsed null runs the same criterion

Table 2: Credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ on the held-out branching supply-tree family, three edge-lag scales, ten draws per cell. Dashes mark quantities undefined without a fitted kernel. Bold marks this work.

System	Short	Medium	Long	Avg	Coef. err. ↓	Agr. / true kernel ↑
Time-decayed return redistribution	0.2311	0.2294	0.2287	0.2297	–	0.2309
Concurrent-return counterfactual difference	0.1655	0.1499	0.1530	0.1561	–	0.1569
Uniform per-agent split	0.2322	0.2345	0.2325	0.2331	–	0.2344
Regime-collapsed (G=1) null	0.9785	0.9786	0.9773	0.9781	0.0783	0.9833
Aggregate-fit kernel	0.9666	0.9614	0.9585	0.9622	0.1151	0.9673
RICK (ours)	**0.9856**	**0.9825**	**0.9813**	**0.9831**	**0.0539**	**0.9884**

on the unperturbed regime alone and lands beside the aggregate fit, so there the recovery comes from the manufactured regimes: a driver-propped kernel explains one regime and misses another, and the worst-regime reading exposes what the pooled record averages away. A denser fit to the recorded log is therefore not evidence that the influence it credits was transmitted. Details in Appendix A.15.

Credit fidelity. Figure 4 presents the same six estimators on credit fidelity against the exact counterfactual-replay ground truth, one panel per environment at ten draws per cell. Averaged over the six fan-in cells the criterion reaches 0.8175 against 0.7765 for the aggregate fit and 0.7881 for the null, while the three estimators that assign credit without fitting a kernel trail between 0.3974 and 0.5183; they carry no transmission model. The kernel-fitting estimators lose fidelity as the edge-lag scale grows, and the criterion stays above the aggregate fit in every one of the nine cells. The fan-in averages preserve the coefficient-error ordering. Details in Appendix A.16.

Held-out transfer. Table 2 carries the six estimators onto a held-out supply-tree family whose points have at most four parents. The coefficient ordering holds off the family it was developed on, at 0.0539 against 0.1151 for the aggregate fit and 0.0783 for the null, and credit fidelity orders the three the same way. Over thirty paired readings this work leads the aggregate fit on credit fidelity by 0.0210, in twenty-seven of them and at each of the three edge-lag scales. On the supply trees, held apart from every development environment, the estimators without a kernel average between 0.1561 and 0.2331, while every kernel-fitting cell holds above 0.95. Details in Appendix A.17.

5.3 Ablation Study and Analysis

Regime count. Table 3 isolates one mechanism at a time, and the regime-family count is swept at a fixed recording budget of sixteen episodes, each rung dividing the same episodes over a different number of regimes. Eight regimes read above one at all three edge-lag scales, and the separation is bought from four regimes upward. Eight regimes read above four on the average fidelity column, 0.8345 against 0.8226. What the family buys is read on the coefficient-error column of Table 1, where the criterion separates from both the aggregate fit and the collapse null on the fan-in environments. More details about the ablation table are provided in Appendix A.18.

Backtracking budget. Figure 5 presents the budget ladder. Compressing the window from sixteen to one drives the unassigned share from 0.0038 to 0.8511 and credit fidelity from 0.9500 to 0.1143; the share rises as fidelity falls at every rung, because the mass a window cannot reach is exactly the mass no estimator inside that window can place. The time-decayed redistribution is flat at 0.7714 from $B = 4$; this work passes it at B=8 and B=16, at 0.8750 and 0.9500. The share of receiving points whose own unassigned share exceeds the conservation tolerance of 0.05 falls from 1.0000 to 0.0000 over the same ladder, which the pooled ratio does not show. Details in Appendix A.19.

Return decoupling. Figure 6 (right) places each seed by the credit it piles on arriving return; the leak the stop-gradient blocks scales each coefficient by the return arriving at its emitter. At the medium edge-lag scale, over five seeds, letting it through carries credit fidelity from 0.9043 to 0.3486 while the agreement between assigned credit and that arriving return rises from 0.0686 to 0.6900, in the same direction on every seed. That is the return-driven degeneracy the stop-gradient exists to prevent. The mean peak error rises by at most 0.19 steps. Details in Appendix A.20.

Lag and kernel variants. Table 3 also carries the two variants that touch the delayed expansion. Freezing the edge kernels a step before the receiving point is worse on both columns at all three scales, lowering credit fidelity and moving the credited peak further from the ground-truth peak. Collapsing the per-edge lag index to a single shared lag moves credit fidelity by at most 0.0250 at

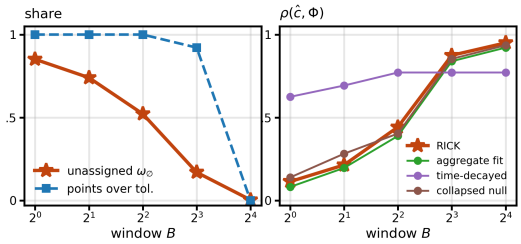


Figure 5: (**Left**) unassigned share and points over tolerance; (**Right**) credit fidelity, against the backtracking window.

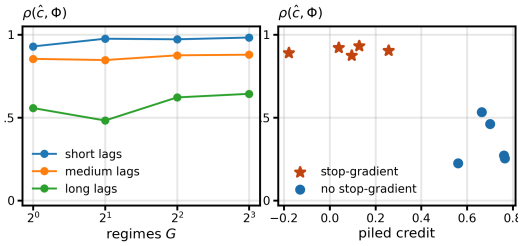


Figure 6: (**Left**) fidelity against the regime count. (**Right**) fidelity against credit piled on arriving return.

Table 3: Per-component ablations on the primary environment, read as credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ at three edge-lag scales. Bold marks the default setting.

Ablated mechanism	Short	Medium	Long	Avg	Degen. ↓	Peak err. (steps) ↓
Regime count $G = 1$	0.9286	0.8536	0.5571	0.7798	−0.3750 / 0.0071 / 0.4750	1.3333 / 4.2000 / 5.4667
Regime count $G = 2$	0.9750	0.8464	0.4821	0.7678	−0.3179 / 0.0821 / 0.4821	1.3333 / 4.2000 / 6.0000
Regime count $G = 4$	**0.9714**	**0.8750**	**0.6214**	**0.8226**	**−0.2500 / 0.0964 / 0.4393**	**1.3333 / 4.2000 / 6.7333**
Regime count $G = 8$	0.9821	0.8786	0.6429	0.8345	−0.2393 / 0.0893 / 0.4286	1.3333 / 4.2667 / 6.4667
Lag: per-edge lag index	**0.9714**	**0.8750**	**0.6214**	**0.8226**	**−0.2500 / 0.0964 / 0.4393**	**1.3333 / 4.2000 / 6.7333**
Lag: single shared lag (collapsed)	0.9786	0.9000	0.6071	0.8286	−0.2821 / 0.0357 / 0.4357	0.0000 / 5.0000 / 5.2000
Kernel: time-varying kernel	**0.9714**	**0.8750**	**0.6214**	**0.8226**	**−0.2500 / 0.0964 / 0.4393**	**1.3333 / 4.2000 / 6.7333**
Kernel: static (frozen before receiving step)	0.9643	0.8536	0.5321	0.7833	−0.2393 / 0.1286 / 0.4821	2.9333 / 6.5333 / 9.2000
Return path: with stop-gradient	**0.9764**	**0.9043**	**0.6800**	**0.8536**	**−0.2436 / 0.0686 / 0.4457**	**0.4400 / 1.1600 / 5.2400**
Return path: no stop-gradient	0.4707	0.3486	0.3029	0.3741	0.5393 / 0.6900 / 0.7736	0.5733 / 1.3467 / 5.2933

every scale. Freezing the kernel lowers average credit fidelity by 0.0393 and long-scale fidelity by 0.0893, and raises its long-scale peak error by 2.47 steps. Details in Appendix A.21.

6 Related Work

Value Decomposition for Cooperative Control. Value decomposition factors the team return into per-agent terms fitted to recorded joint play (Zhou et al., 2020; Zohar et al., 2022; Dou et al., 2022), graph-based mixers learn the topology from the same logs (Naderializadeh et al., 2020), and later designs add marginal contributions, agent types and fusion to that fit (Liu et al., 2023; Fu et al., 2025; Wang et al., 2025a). Every term is identified from variation the logs already contain, and a shared driver and true transmission are admitted alike. RICK fixes the kernel on regimes it manufactures.

Counterfactual and Game-Theoretic Credit. Counterfactual estimators score an agent by what a replaced action changes (Foerster et al., 2018; Mesnard et al., 2021; Su et al., 2020), and game-theoretic ones divide the outcome by a Shapley, nucleolus or core share (Li et al., 2021; 2025c; Ji et al., 2025; Ding et al., 2025). Neither is priced: the return a perturbation costs never enters the estimator, and which perturbations are worth buying is never asked. RICK bounds the return its perturbation family may cost and reads credit on what that family separates.

Temporal and Delayed Credit Assignment. Temporal credit assignment spreads a delayed return over the steps that earned it, through sequence models, inference networks, agent-time attention and spatial-temporal decomposition (Pignatelli et al., 2024; Liu et al., 2019; Ausin et al., 2021; She et al., 2022; Chen et al., 2023). What a bounded reach misses is renormalised away, and truncation reports full conservation. RICK states the unreachable mass as an unassigned share.

7 Conclusion

We identified self-certifying evidence in multi-agent credit assignment: the fit residual, perturbation schedule and conservation percentage that certify a credit kernel all read the record it was fitted to. We presented RICK, which couples regime-invariant kernel selection, a regime scheduler, and delayed backtracking and share accounting so that each certifying readout can fail. It recovers the coefficients on the fan-in environments to a mean of 0.0935 against 0.2186 for the pooled aggregate fit and keeps that ordering off-family, offering a practical recipe for certifying delayed credit.

REFERENCES

Markel Sanz Ausin, Hamoon Azizsoltani, Song Ju, Yeo Jin Kim, and Min Chi. InferNet for delayed reinforcement tasks: Addressing the temporal credit assignment problem. In *IEEE BigData*, 2021.

João Barretto-Bittar, Anna Levina, Emmanouil Giannakakis, and Roxana Zeraati. Diffusion of neuromodulators for temporal credit assignment. *Proceedings of the National Academy of Sciences*, 123(38):e2608831123, 2026.

Sirui Chen, Zhaowei Zhang, Yaodong Yang, and Yali Du. STAS: Spatial-temporal return decomposition for multi-agent reinforcement learning, 2023. arXiv preprint.

Shripad Deshmukh, Jayakumar Subramanian, Raghavendra Addanki, and Nikos Vlassis. COSAC: Counterfactual credit assignment in sequential cooperative teams. arXiv preprint arXiv:2604.17693, 2026.

Ao Ding, Licheng Sun, Yongjie Hou, Huaqing Zhang, and Hongbin Ma. A historical interaction-enhanced shapley policy gradient algorithm for multi-agent credit assignment, 2025. arXiv preprint.

Fei Ding, Yongkang Zhang, Youwei Wang, and Zijian Zeng. Reducing credit assignment variance via counterfactual reasoning paths, 2026. arXiv preprint.

Zehao Dou, Jakub Grudzien Kuba, and Yaodong Yang. Understanding value decomposition algorithms in deep cooperative multi-agent reinforcement learning, 2022. arXiv preprint.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Songchen Fu, Shaojing Zhao, Ta Li, and YongHong Yan. QTypeMix: Enhancing multi-agent cooperative strategies through heterogeneous and homogeneous value decomposition. *Neural Networks*, 2025.

Shuai Han, Mehdi Dastani, and Shihan Wang. Credit assignment and efficient exploration based on influence scope in multi-agent reinforcement learning, 2025. arXiv preprint.

Diyi Hu and Bhaskar Krishnamachari. Wireless communication enhanced value decomposition for multi-agent reinforcement learning, 2026. arXiv preprint.

Mengda Ji, Genjiu Xu, Keke Jia, Zekun Duan, Yong Qiu, Jianjun Ge, and Mingqiang Li. Cooperative game-theoretic credit assignment for multi-agent policy gradients via the core, 2025. arXiv preprint.

Tianrui Jia, Haoyang Li, Cheng Yang, Tao Tao, and Chuan Shi. Graph invariant learning with subgraph co-mixup for out-of-distribution generalization. In *AAAI*, 2024.

Mateo Juliani, Mingxuan Li, and Elias Bareinboim. Confounding robust continuous control via automatic reward shaping. In *AAMAS*, 2026.

Aditya Kapoor, Sushant Swamy, Kale ab Tessera, Mayank Baranwal, Mingfei Sun, Harshad Khadilkar, and Stefano V. Albrecht. Agent-temporal credit assignment for optimal policy preservation in sparse multi-agent reinforcement learning, 2024. arXiv preprint.

Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Beyond uniform credit: Causal credit assignment for policy optimization, 2026. arXiv preprint.

Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Fei Wu, and Jun Xiao. Shapley counterfactual credits for multi-agent reinforcement learning. In *KDD*, 2021.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Automatic reward shaping from confounded offline data. In *ICML*, 2025a.

Mingxuan Li, Kai-Zhan Lee, and Elias Bareinboim. Counterfactual shapley credit assignment. arXiv preprint arXiv:2607.16999, 2026.

Yuanjun Li, Zhouyang Jiang, Bin Zhang, Mingchao Zhang, Junhao Zhao, and Zhiwei Xu. QLLM: Do we really need a mixing network for credit assignment in multi-agent reinforcement learning? arXiv preprint arXiv:2504.12961, 2025b.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. Nucleolus credit assignment for effective coalitions in multi-agent reinforcement learning. In *AAMAS*, 2025c.

Yongheng Liang, Hejun Wu, Haitao Wang, and Hao Cai. Asynchronous credit assignment for multi-agent reinforcement learning. In *IJCAI*, 2025.

Shanqi Liu, Yujing Hu, Runze Wu, Dong Xing, Yu Xiong, Changjie Fan, Kun Kuang, and Yong Liu. Adaptive value decomposition with greedy marginal contribution computation for cooperative multi-agent reinforcement learning. In *AAMAS*, 2023.

Yang Liu, Yunan Luo, Yuanyi Zhong, Xi Chen, Qiang Liu, and Jian Peng. Sequence modeling of temporal credit assignment for episodic reinforcement learning, 2019. arXiv preprint.

Yuhan Helena Liu, Guangyu Robert Yang, and Christopher J. Cueva. Can biologically plausible temporal credit assignment rules match BPTT for neural similarity? E-prop as an example. In *ICML*, 2025.

Zibin Meng and Kani Chen. CRAFT: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning, 2026. arXiv preprint.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Navid Naderializadeh, Fan H. Hung, Sean Soleyman, and Deepak Khosla. Graph convolutional value decomposition in multi-agent reinforcement learning, 2020. arXiv preprint.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bNtr6SLgZf. Survey Certification.

Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative Q learning for offline multi-agent reinforcement learning. In *NeurIPS*, 2023.

Jennifer She, Jayesh K. Gupta, and Mykel J. Kochenderfer. Agent-time attention for sparse rewards multi-agent reinforcement learning. In *AAMAS*, 2022.

Jianyu Su, Stephen Adams, and Peter A. Beling. Counterfactual multi-agent reinforcement learning with graph convolution communication, 2020. arXiv preprint.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents, 2026. arXiv preprint.

Viktor Veselý, Aleksandar Todorov, Erwan Escudie, and Matthia Sabatelli. Trace-mediated peak bias: Bridging temporal credit assignment and cognitive heuristics in deep reinforcement learning, 2026. arXiv preprint.

Siying Wang, Yang Zhou, Zhitong Zhao, Ruoning Zhang, Jinliang Shao, Wenyu Chen, and Yuhua Cheng. Heterogeneous value decomposition policy fusion for multi-agent cooperation, 2025a. arXiv preprint.

Ziyan Wang, Yali Du, Yudi Zhang, Meng Fang, and Biwei Huang. MACCA: Offline multi-agent reinforcement learning with causal credit assignment. *Trans. Mach. Learn. Res.*, 2025b.

Baicen Xiao, Bhaskar Ramasubramanian, and Radha Poovendran. Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning. In *AAMAS*, 2022.

Baidi Xiao, Rongpeng Li, Fei Wang, Chenghui Peng, Jianjun Wu, Zhifeng Zhao, and Honggang Zhang. Stochastic graph neural network-based value decomposition for MARL in internet of vehicles. *IEEE Trans. Veh. Technol.*, 2024.

Zhiwei Xu, Bin Zhang, Dapeng Li, Guangchong Zhou, Zeren Zhang, and Guoliang Fan. Dual self-awareness value decomposition framework without individual global max for cooperative MARL. In *Advances in Neural Information Processing Systems 36*, 2023. doi: 10.52202/075280-3233.

Haohan Yu, Jinmiao Cong, Shengzhi Wang, Lu Wang, and Chanjuan Liu. MAGIC: Multi-step advantage-gated causal influence for multi-agent reinforcement learning, 2026. arXiv preprint.

Wenjing Zhang and Wei Zhang. ToMacVF: Temporal macro-action value factorization for asynchronous multi-agent reinforcement learning. In *ECAI*, 2025.

Qingpeng Zhao, Yuanyang Zhu, Zichuan Liu, Zhi Wang, and Chunlin Chen. Boosting value decomposition via unit-wise attentive state representation for cooperative multi-agent reinforcement learning, 2023. arXiv preprint.

Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, and Yuk Ying Chung. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. In *NeurIPS*, 2020.

Roy Zohar, Shie Mannor, and Guy Tennenholtz. Locality matters: A scalable value decomposition approach for cooperative multi-agent reinforcement learning. In *AAAI*, 2022.

AI USE STATEMENT

We used generative AI tools to translate the text (Claude Opus 5.0), to check grammar and spelling and polish the wording (GPT-5.6), to guide the drawing of the figures (Gemini), and to check the content (Grok). We did not use generative AI tools to modify code. We have reviewed all AI-assisted work: we checked every reference for hallucinated citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The details needed to reproduce the results of this paper, including the environments, the estimator configuration, the evaluation metrics, the budgets, schedule and kernel class, the compute protocol and the hyperparameters, are disclosed in this appendix.

A APPENDIX

Table 4 lists the contents of each part of the appendix.

A.0 LIMITATIONS

Scaling with environment size. The evaluation covers three environments, each at a single fixed size, and the paper lacks a study that grows the number of nodes and edges at a matched per-node backtracking budget. Such a study matters to a reader who wants to apply the method to a graph of a different size, because it would show how credit fidelity and the expansion time of one credit map change as the graph grows.

Calibration of the credit-fidelity readout. Every credit-fidelity value in the paper comes from one readout, the per-agent Spearman rank correlation against the counterfactual-replay ground truth, and the paper lacks a calibration of that readout. Such a calibration matters because a reader needs to know how large a departure from the ground truth the readout itself produces: applying the per-agent Spearman correlation, a per-agent-per-step Spearman correlation and a top-third overlap both to the estimator and to the exact replay at the same deployer budget would show the gap between the two under each readout.

Table 4: What each part of the appendix holds.

Section	Contents	Section	Contents
Appendix A.0	Limitations	Appendix A.13	Hyperparameters
Appendix A.1	Extended related work	Appendix A.14	Credit-fidelity table
Appendix A.2	Motivation diagnostics	Appendix A.15	Coefficient-error table
Appendix A.3	Worst-regime criterion	Appendix A.16	Credit-fidelity figure
Appendix A.4	Schedule search	Appendix A.17	Held-out family table
Appendix A.5	Lag and share parameterisation	Appendix A.18	Ablation table
Appendix A.6	Credit expansion procedure	Appendix A.19	Budget-ladder table
Appendix A.7	Return decoupling	Appendix A.20	Ablation and ladder figures
Appendix A.8	Environments	Appendix A.21	Lag and kernel variants
Appendix A.9	Estimator configuration	Appendix A.22	Robustness sweeps
Appendix A.10	Evaluation metrics	Appendix A.23	Cost model
Appendix A.11	Budgets, schedule and kernel class	Appendix A.24	Discussion
Appendix A.12	Compute protocol		

Measured structural identities. The credit expansion rests on four structural identities: the forward topological expansion zeroes every non-ancestor term, the lag profile stays positive on the whole grid, the weights and the unassigned share sum to one at every receiving point, and the per-regime discrepancies coincide at the true kernel when the family covers every emitter. The paper lacks a direct numerical measurement of each identity. Such measurements matter because the credit expansion is correct only when its implementation satisfies these identities, and a measured value for each identity lets a reader check that it does.

A.1 Extended Related Work

Value decomposition. Cooperative value decomposition factors a shared team return into per-agent terms fitted on the recorded joint play, either implicitly through a learned assignment or explicitly through a mixing network (Zhou et al., 2020; Dou et al., 2022). Later designs question the mixer itself, removing the reliance on an individual global maximum and asking whether a mixing network is needed at all (Xu et al., 2023; Li et al., 2025b). What these share is a criterion defined on the logged joint play, and a factorisation is accepted when it reproduces that play and for no other reason.

Structured and heterogeneous mixers. A second line makes the mixer structured. Graph-based variants parameterise the mixing over an interaction topology and learn it from the same logs (Naderializadeh et al., 2020; Zohar et al., 2022; Xiao et al., 2024), while attention-based and marginal-contribution designs sharpen which unit receives which share (Zhao et al., 2023; Liu et al., 2023). Heterogeneous fusion, agent-type decomposition and macro-action factorisation extend the same criterion to teams whose members differ in kind or in timescale (Wang et al., 2025a; Fu et al., 2025; Zhang & Zhang, 2025). Structure raises what the family can express without changing what selects among its members, which is still agreement with the recorded propagation.

Counterfactual estimators. Counterfactual credit assignment scores an agent by the change a replaced action induces, introduced for cooperative policy gradients and developed into model-free baselines and communication-aware variants (Foerster et al., 2018; Mesnard et al., 2021; Su et al., 2020). Offline and causal formulations recover the same quantity from logged play under conservative value estimates (Shao et al., 2023; Wang et al., 2025b), and recent work reduces its variance using sibling rollouts, counterfactual reasoning paths, gated causal influence and sequential team decompositions (Meng & Chen, 2026; Ding et al., 2026; Yu et al., 2026; Deshmukh et al., 2026). In all of them the replacement is evaluated inside the trajectory distribution the agents already induce, and the return such a perturbation would cost a deployed team is not part of the estimator.

Game-theoretic shares. A parallel family divides the arrived outcome by a solution concept of the recorded play, using Shapley values, nucleolus shares, core-based allocations and counterfactual-Shapley hybrids (Li et al., 2021; 2025c; Ji et al., 2025; Ding et al., 2025; Li et al., 2026). The axioms fix how a fixed characteristic function should be divided; they say nothing about how that function

is estimated, and a characteristic function a shared driver already explains is divided as faithfully as a correct one.

Return redistribution over time. Temporal credit assignment redistributes a delayed or episodic return over the timesteps that produced it, as surveyed for deep reinforcement learning (Pignatelli et al., 2024). Sequence models and inference networks learn to reallocate a sparse terminal reward to earlier decisions (Liu et al., 2019; Ausin et al., 2021), hindsight reallocation carries the idea to long-horizon agents (Tan et al., 2026), and asynchronous schemes relax the assumption that agents act on a common clock (Liang et al., 2025). These methods move mass backwards along the time axis without committing to a lag between a particular release and the arrival it caused.

Agent-time attention and propagation models. Attention-based redistribution spreads an episodic reward jointly across agents and steps, through agent-time and spatial-temporal decompositions and through policy-preserving agent-temporal schemes (She et al., 2022; Xiao et al., 2022; Chen et al., 2023; Kapoor et al., 2024). Influence-scope and biologically motivated propagation models instead carry credit along a learned or diffusive reach (Han et al., 2025; Barretto-Bittar et al., 2026; Liu et al., 2025), and recent analyses examine the bias an eligibility trace inherits and what lies beyond it (Veselý et al., 2026; Khandoga et al., 2026). The weight these designs place on a distant step is fitted on the same episodes it is read from, and whatever lies outside the reach they can represent is normalised over the part that was reached rather than reported.

A.2 Motivation diagnostics

This section enlarges each panel of the motivation figure (Figure 7) and reports the measurements behind it. Every reading here is taken on records held apart from the results reported in the body, and the kernels read are the observational ones this paper argues against, together with the generator's own kernel where nothing is being fitted.

Left panel: a shared driver behind an observational fit. A kernel is fitted to an observational log alone on the cascaded production line, the primary environment, and read twice: once by the residual it reports about itself, and once by the distance of the transmission coefficients it recovered from the ones the log was produced under. Table 5 lists both readings at five autocorrelation levels of the shared exogenous injection. At the highest level the fit accounts for all but 0.0261 of the variance it is explaining, and the coefficients it recovered are still 0.4541 away from the truth on average, because the releases of any node's parents move together at a median correlation of 0.9865. The distance holds across the five levels: it stays between 0.4087 and 0.5039; no co-parent correlation between 0.9670 and 0.9865 brings the recovered coefficients near the truth.

Middle panel: what an excitation buys. Two excitations are fixed in advance on the same environment, one reaching a single emitter and one reaching every emitter, each swept over the same seven strengths. Identification is the smallest eigenvalue any receiving point's parent design carries, each design scaled to unit diagonal, keeping the weakest direction from being chosen by the units the releases happen to carry; energy is the squared departure of the excited releases from the paired unexcited ones on the same exogenous draws. Table 6 reports both settings. Exciting one emitter moves that eigenvalue from 0.0015 to 0.0018 across the whole sweep while delivering 1.6194 of energy, and the identification standing at the end stays within 0.0003 of what the passive log already carried. Exciting every emitter carries it to 0.3595, two orders of magnitude above where it started, and at an energy of 1.7119, close to the single emitter's entire sweep, it already reads 0.0288. What separates the two readings is not how much was spent but where. Both axes of the panel are logarithmic, and the dotted horizontal line marks the identification the unperturbed log already carries. The shaded vertical band marks the matched-cost comparison the challenge rests on, where 1.6194 against 1.7119 of energy separates 0.0018 from 0.0288, a factor of 16; by the end of its own sweep the every-emitter setting is a factor of 200 above the single-emitter one, and 10.8 times more efficient per unit of energy delivered. The zero-strength rung of each setting is omitted because a logarithmic axis cannot carry it; both settings read 0.0015 there, which is the dotted line itself.

Right panel: what a bounded window cannot reach. The backtracking budget is compressed below the longest edge lag and the expansion is run with the environment's own kernel, with nothing

Table 5: Challenge #1. A kernel fitted to an observational log alone, on the primary environment, the cascaded production line, at five autocorrelation levels of the shared exogenous injection. The residual the fit reports about itself ↓ is small at every level while the transmission coefficients it recovered stay between 0.4087 and 0.5039 away ↓ from the ones the log was produced under, a distance that does not close as the releases of any node's parents move more tightly together ↑.

Driver autocorr.	Fit residual (share of var.)	Coefficient error	Co-parent collinearity
0.20	0.0253	0.5003	0.9670
0.40	0.0147	0.5039	0.9711
0.60	0.0160	0.4800	0.9756
0.80	0.0147	0.4087	0.9840
0.95	0.0261	0.4541	0.9865

fitted and the reading a property of the budget rather than of any estimator under test. Table 7 traces the unreached share as the budget is compressed. At the shortest window 0.8239 of the arrived mass lies outside it, the credited peak is misplaced by 21.5333 timesteps, and only 24.650 receiving points' worth of mass sits inside the budget; at a window of 16 the unreached share falls to 0.0030, the peak error to 0.0000, and the mass inside the budget rises to 139.580. A percentage taken over only the reached set prints 1.0000 at every one of the five windows, which is the flat upper series in the panel: it is not a measurement but the value that convention returns by construction, whatever the window failed to reach.

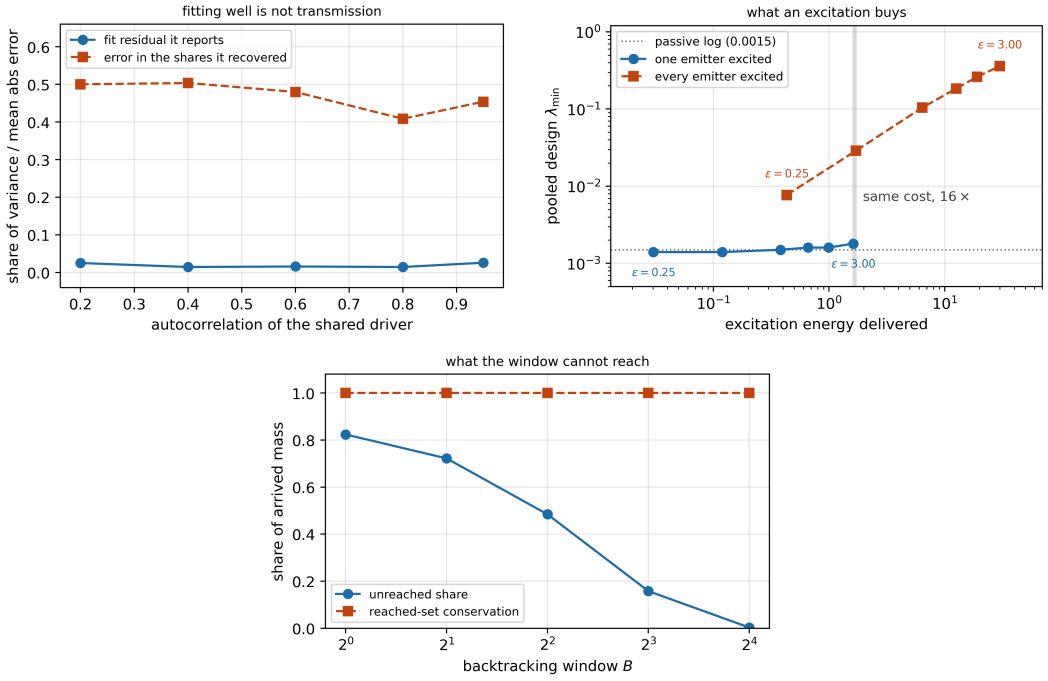


Figure 7: (Top left) the fit residual against the distance of the coefficients recovered. (Top right) identification against excitation energy. (Bottom) unreached mass against the reached-set convention.

A.3 Worst-regime criterion

For the pooled design $\bar{\Sigma}_v(\sigma)$ of Eq. 9, with positive weights ρ_g summing to one and x_v read as the features that multiply the coefficients in Eq. 12, extended by the baseline feature, suppose the true kernel $\Psi^\star$ lies in $\mathcal{K}$, the per-regime residuals are orthogonal to x_v with a common noise floor, every regime records each of the N_{r} receiving nodes equally often, and the

Table 6: Challenge #2. What an excitation buys on the primary environment, over seven strengths and two excitations fixed before the sweep. Identification ↑ is the smallest eigenvalue any receiving point's parent design carries, each design scaled to unit diagonal, keeping the weakest direction from being chosen by the units the releases happen to carry. Energy ↓ is the squared departure of the excited releases from the paired unexcited ones on the same exogenous draws. Reaching one emitter keeps identification within 0.0003 of the passive log while spending up to 1.6194 of energy; reaching every emitter carries it to 0.3595.

	One emitter		Every emitter	
ε	Identification $\lambda_{\min}$ ↑	Energy delivered ↓	Identification $\lambda_{\min}$ ↑	Energy delivered ↓
0.00	0.0015	0.0000	0.0015	0.0000
0.25	0.0014	0.0302	0.0077	0.4317
0.50	0.0014	0.1195	0.0288	1.7119
1.00	0.0015	0.3822	0.1041	6.4075
1.50	0.0016	0.6605	0.1838	12.5851
2.00	0.0016	0.9916	0.2606	19.0884
3.00	0.0018	1.6194	0.3595	30.0198

Table 7: Challenge #3. The arrived mass a bounded window cannot reach on the primary environment, over five windows, with the environment's own kernel, nothing being fitted. The last column is the mass inside the budget expressed in receiving points. The unreached share runs from 0.8239 at B=1 to 0.0030 at B=16 and the credited peak moves from 21.5333 timesteps of error to 0.0000, while a percentage taken over the reached set alone prints 1.0000 at every one of the five rungs.

B	Unreached share ↓	Peak err. ↓	Mass inside the budget, in receiving points ↑
1	0.8239	21.5333	24.650
2	0.7222	14.2667	38.890
4	0.4843	0.6000	72.190
8	0.1578	0.0667	117.910
16	0.0030	0.0000	139.580

smallest eigenvalue $\lambda_{\min}(\bar{\Sigma}_v(\sigma))$ is positive at every v. Then $\max_g D_g(\Psi) - \max_g D_g(\Psi^\star) \geq \frac{1}{N_r} \min_v \lambda_{\min}(\bar{\Sigma}_v(\sigma)) \, \|\Delta\|^2$ for the displacement Δ of the coefficients of Ψ from those of $\Psi^\star$, and the driver-propped kernel is excluded in the population limit. If the family leaves the collinear direction unexcited at some receiving point, $\min_v \lambda_{\min}(\bar{\Sigma}_v(\sigma)) = 0$, the bound is vacuous, and the criterion excludes nothing the pooled fit does not already exclude. At finite sample the exclusion holds only where the gap exceeds the estimation and solver error.

The reading that answers Challenge #1 is the coefficient-error column of Table 1; the family-count sweep against the collapsed criterion of Eq. 4 is reported in Table 3.

A.4 SCHEDULE SEARCH

The departure a regime causes from its paired unperturbed rollout is taken on the same exogenous draws and squared, each departure measured in units of the corresponding sink's receiving-value scale, and the cost of Eq. 7 is comparable across regimes of different size.

The greedy does not evaluate $\bar{\Sigma}_v(\sigma)$ itself: it reads one unperturbed rollout, standardises the parent second moment to a correlation matrix, and adds the diagonal increment an excitation of the budgeted strength would contribute if it were independent of the releases already there. It is that proxy the search maximises, and no regime weighting enters it. The excited set K is grown greedily while the objective still rises, the budget being divided equally in cost across its members, giving $\varepsilon_u^2 = \bar{c}/(|K|\kappa_u)$ for the return sensitivity κ_u each emitter is probed for; the set is then split into $G - 1$ disjoint non-empty masks. Reading the same design by its log determinant instead returns a different family. The scheduler resolves ε_g by bisecting a shared strength on the return cost the family inflicts on the behaviour in force and rescaling it per mask to that mask's measured cost, and

the strength follows that behaviour's own sensitivity rather than an outside setting; it is resolved afresh whenever a family is drawn.

A.5 Lag and share parameterisation

The per-edge transmission coefficient ranges over $c_e \in [0, 3]$ and equals one under the generator. The coefficient error reported throughout is the mean distance of the recovered c_e from that value; the share is not bounded by one.

A.6 Credit expansion procedure

Algorithm 1 gives one sampling round: the scheduler prices and commits an intervention family it can afford, episodes are recorded under that family against paired unperturbed rollouts, the kernel is fixed on the family's worst regime and expanded, the mass no bounded reach recovers is booked as unassigned, and the policy is improved while the family is re-costed against the return it earned. The cost of the expansion is given in Appendix A.23. Three orderings are load-bearing, and reversing any of them breaks a claim. The strength is bisected before the family is admitted, and feasibility is decided on the cost the behaviour in force actually incurs rather than on a nominal setting. A kernel is fixed only after the family it is judged on has been committed and paid for, and the regimes it is scored against are the ones the round actually bought. And the reference-budget denominator is formed before the window and the ancestor depth restrict what is reported, since normalising after truncation would erase exactly the mass the unassigned share exists to record. The block states the programs it solves and leaves their solvers to the prose around it: the mask search and the numerical minimisation of Eq. 3.

RICK answers each challenge with the module the formulation assigns to it: worst-regime selection over the scheduled family separates transmission from sharing, the cost-bounded scheduler makes the separating interventions payable, and the bounded backtracking surfaces the arrived mass its window cannot reach. The three compose into the single selection rule of Eq. 1.

A.7 Return decoupling

The stop-gradient separates the two levels of Eq. 1: the outer gradient moves π, while the kernel stays the one the discrepancy criterion selected under the committed family. Without it the outer maximisation could raise return by moving the kernel itself, which is the degeneracy of piling credit where return is already high rather than where it was transmitted. The separation is what the paired return-path rows of Table 3 read.

A.8 Environments

Nodes are numbered in topological order, every edge runs from a lower to a higher index, and the forward expansion stage of Algorithm 1 is a single sweep over that order. A node's release is its own base emission plus a nonnegative share of what it received; releases are positive everywhere, and the prefix weights the accounting accumulates are nonnegative by construction.

The three training environments are a cascaded production line built from layered fan-in with skip edges, a tidal road-traffic network built from a directed grid, and a multi-level cache built from layered fan-in with sparse skips. They differ in fan-in, which is the axis the confounding of Challenge #1 depends on: the road-traffic grid gives every receiving point at most two parents. The held-out family is a branching supply tree with heavy-tailed edge lags, built convergent with suppliers feeding assemblies; it is held apart from the training environments and shares no record with them.

Every node is driven by one shared autocorrelated exogenous injection, which is what makes the releases of any node's parents move together in an observational log. A regime perturbation is drawn from a noise stream independent of that driver and clipped to the admissible action range; a masked action is an intervention rather than a reweighting of recorded behaviour.

Algorithm 1 One sampling round of RICK

Require: policy π, kernel class $\mathcal{K}$, return-cost bound $\bar{c}$, regime count $G \geq 2$, mixing weights $\{\rho_g\}_{g=1}^{G}$, recording budget in episodes, window B, ancestor depth M, lag bound $L_e^{\max}$

Ensure: a policy improved on a credit map $(w, \omega_\emptyset)$ that conserves at every receiving point, with the family and kernel that produced it

deal the excited emitters into $G-1$ disjoint non-empty masks and leave the remaining regime unperturbed ▷ the index round-robin of Appendix A.11 by default, the proxy search of Appendix A.4 otherwise

bisect a shared strength on the return cost of Eq. 7 the family inflicts on the behaviour in force and rescale it per mask to ε_g ▷ resolved afresh for every family drawn, and strength tracks the current policy

commit the feasible family σ of Eq. 8 ▷ a single regime returns the criterion to the single-regime fit of Eq. 4

collect the recording budget under π and σ, each regime paired with an unperturbed rollout on the same exogenous draws ▷ the stream of Eq. 5 is independent of the drivers; a masked action is an intervention

for each regime $g \in \sigma$ **do**

    $D_g(\Psi) \leftarrow$ the per-regime discrepancy of Eq. 2 on that regime's own points T_g

end for

$(\Psi_\sigma, \eta_\sigma) \leftarrow$ the kernel and level returned by minimising the epigraph program of Eq. 3 numerically over $\mathcal{K}$ ▷ the family scored here is the one already committed and paid for

form each edge's lag profile ϕ_e^t of Eq. 10 and transmission share α_e^t of Eq. 11 from Ψ_σ

for each receiving point (j, s) in forward topological order **do**

    expand Eq. 12 over the full ancestor set and take the reference-budget denominator from it ▷ formed before any truncation, every non-ancestor term being identically zero

    $w_{i,\, s-\tau' \to j,s} \leftarrow$ the prefix weight carried from $(i, s-\tau')$ to (j, s) over that denominator, for every emitter in $\mathcal{R}(j, s)$ of Eq. 13 ▷ the reach bounds what is reported, not what is computed

    $\omega_\emptyset^{j,s} \leftarrow 1 - \sum_{(i,s-\tau') \in \mathcal{R}(j,s)} w_{i,\, s-\tau' \to j,s}$ by Eq. 14 ▷ g_ψ identically positive keeps the weights nonnegative, and an under-reaching budget raises $\omega_\emptyset$ instead of renormalising the shortfall away

end for

improve π on $(w, \omega_\emptyset)$ under $\max_{\pi,\, \sigma \in \mathcal{S}} J(\pi; \mathrm{sg}[\Psi_\sigma])$ ▷ the stop-gradient holds return out of the selection

re-cost σ by Eq. 7 against the return this round earned and carry that cost into the next search

return the improved policy with its credit map, and the family and kernel that produced it

A.9 Estimator configuration

The collapse null is fitted on an unperturbed collection carrying the same 16 episodes rather than on one regime's share of them. Any reading taken across G is therefore taken at a fixed number of recorded episodes, and a difference across G is a difference between family structures rather than between sample sizes.

All six estimators read the same recording budget and expand under the same window and ancestor depth. The three kernel-fitting rules differ in the criterion by which the kernel is selected: the worst-regime criterion of Eq. 3, solved on its dual by exponentiated gradient over the regime-weight simplex; a pooled fit over the log, weighted by the mixing weights ρ_g and pooled counts; and the collapse null of Eq. 4, which sees only the unperturbed regime. The uniform per-agent split spreads each arrival equally over the emitters the budget reaches. The concurrent-return counterfactual difference places the whole arrival on the same timestep it arrived. The time-decayed return redistribution spreads it backwards with one shared decay, in proportion to the recorded activity of each emitter.

Every estimator lays its credit density over the same reference support before the deployer budget truncates it. The three kernel-fitting rows divide by the reference mass the unbounded expansion carries and report an unassigned share. The uniform split, the concurrent-return difference and the time-decayed redistribution do not: each normalises over the emitters and timesteps its own window

and ancestor depth reached, which is the truncated-set denominator this paper argues against, and none of them carries a reference mass against which a shortfall could be stated. Their share column is therefore empty.

A.10 Evaluation metrics

Credit fidelity is the Spearman rank correlation across agents between the credit an estimator assigns and the counterfactual-replay ground truth of Mesnard et al. (2021),

$$\rho(\hat{c}, \Phi) \;=\; \mathrm{corr}_{v \in V}\big(\mathrm{rank}(\hat{c}_v),\; \mathrm{rank}(\Phi_v)\big), \tag{15}$$

where $\hat{c}_v$ is the credit assigned to agent v summed over timesteps, Φ_v the replay ground truth summed the same way, and V the agents the ground truth credits; tables write it ρ. The transmission-coefficient error is the mean distance of the recovered per-edge coefficients c_e from the ones $c_e^\star$ that generated the log,

$$\mathrm{err} \;=\; \frac{1}{|E|} \sum_{e \in E} \big| c_e - c_e^\star \big|. \tag{16}$$

A.11 Budgets, schedule and kernel class

The deployer-afforded window is $B = 8$ and the ancestor depth $M = 4$, held fixed everywhere except in the ladders of Table 7, Table 9 and Table 12, which sweep the window over the rungs 1, 2, 4, 8 and 16. The regime family carries $G = 4$ regimes outside the sweep of Table 3, which runs G over 1, 2, 4 and 8, and the default return-cost bound the schedule is charged against is 0.002 of team return. At the default $G = 4$ every regime is sampled equally, 4 episodes of 48 timesteps each, and the sweep over G holds that total of 16 at every rung; the weights ρ_g that enter the cost and the pooled fit put 0.90 on the unperturbed regime and divide 0.10 over the perturbed ones. The edge-lag grid is bounded at 8, and every cell of Table 8 is read over 10 independent draws. The perturbation strengths of Table 15 are 0.00, 0.50, 1.50, 3.50 and 7.00. The unassigned share is always taken against the reference budget, the unbounded expansion on the same environment, and never against the set the budget reached.

The schedule commits a family of disjoint non-empty masks over the emitting nodes, and the mask-separation clause of Eq. 8 holds by construction. In the main tables and in every perturbed ablation variant those masks are an index round-robin over the family of Eq. 6, which deals the emitting nodes over the $G - 1$ perturbed regimes in index order and covers every one of them; the greedy of Eq. 9 is evaluated only in the schedule-objective comparison of Tables 13 and 14, against that round-robin. In the greedy an unperturbed roll supplies each receiving point's parent design and one probe roll per emitter supplies its return sensitivity, and the excited set grows while the pooled smallest eigenvalue still rises. The bound is charged for the perturbed episodes the committed family collects; the probe rolls that size the strengths are not counted against it. The strength is resolved by bisection against the return-cost bound rather than set by hand, then aligned per regime against the policy's sensitivity to that mask, and every perturbed regime is rescaled toward the same share of the budget; Table 15 reads what a constant strength does instead. The unperturbed regime keeps 4 of the 16 episodes, and the weight above is a cost and pooled-fit weight rather than a sampling share.

The feasible kernel class carries one nonnegative capped transmission coefficient per edge, one baseline load per receiving node, and a lag-scale multiplier on a small grid. The three edge-lag scales read throughout the paper are multipliers on the environment's base edge lags, 1.0 for short, 2.0 for medium and 3.5 for long. Selection is a ridge-regularised least squares per receiving node, the ridge being 10^{-6} times the mean diagonal entry of the normal matrix, under the same conditioning for every rule, projected onto the class, and no rule is advantaged by a different regulariser.

Wall clock is the expansion time of one credit map, reported only against this paper's own variants.

A.12 Compute protocol

The expansion is run over the full ancestor set, which makes the reference mass exact, and the window B and the ancestor depth M are applied to the result. What they bound is therefore the credit reported, not the work done.

Table 8: Credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ on three directed delayed environments at three edge-lag scales, ten draws per cell. Bold marks this work.

System	Short	Medium	Long	Avg
Cascaded production line				
RICK (ours)	**0.9532 (0.0540)**	**0.7811 (0.0822)**	**0.5575 (0.0805)**	**0.7639**
Aggregate-fit kernel	0.9225 (0.0453)	0.7300 (0.1105)	0.4893 (0.0871)	0.7139
Uniform per-agent split	0.5461 (0.1818)	0.4918 (0.1875)	0.4659 (0.1773)	0.5013
Concurrent-return counterfactual difference	0.4386 (0.2862)	0.3643 (0.2851)	0.3268 (0.2665)	0.3766
Time-decayed return redistribution	0.5947 (0.2196)	0.5686 (0.2479)	0.5468 (0.2348)	0.5700
Regime-collapsed (G=1) null	0.8875 (0.0686)	0.7689 (0.1063)	0.5482 (0.1431)	0.7349
Tidal road traffic				
RICK (ours)	**0.9953 (0.0032)**	**0.9827 (0.0094)**	**0.9671 (0.0121)**	**0.9817**
Aggregate-fit kernel	0.9927 (0.0062)	0.9812 (0.0097)	0.9662 (0.0117)	0.9800
Uniform per-agent split	0.1383 (0.2802)	0.1109 (0.2787)	0.1137 (0.2761)	0.1210
Concurrent-return counterfactual difference	0.1038 (0.2265)	0.0729 (0.2328)	0.0518 (0.2325)	0.0762
Time-decayed return redistribution	0.1832 (0.2781)	0.1562 (0.2780)	0.1506 (0.2721)	0.1633
Regime-collapsed (G=1) null	0.9947 (0.0027)	0.9812 (0.0110)	0.9673 (0.0119)	0.9811
Multi-level cache				
RICK (ours)	**0.9671 (0.0228)**	**0.9056 (0.0458)**	**0.7406 (0.1723)**	**0.8711**
Aggregate-fit kernel	0.9189 (0.0481)	0.8804 (0.0598)	0.7182 (0.1715)	0.8392
Uniform per-agent split	0.4489 (0.1767)	0.3925 (0.1849)	0.3415 (0.1686)	0.3943
Concurrent-return counterfactual difference	0.4545 (0.3108)	0.4203 (0.3279)	0.3797 (0.2986)	0.4182
Time-decayed return redistribution	0.5070 (0.2167)	0.4699 (0.2116)	0.4231 (0.2085)	0.4667
Regime-collapsed (G=1) null	0.9343 (0.0490)	0.8769 (0.0587)	0.7126 (0.1782)	0.8413

A.13 Hyperparameters

The regime count is swept over $G \in \{1, 2, 4, 8\}$ with $G = 4$ the default; the backtracking window is swept over $B \in \{1, 2, 4, 8, 16\}$. The mixing weights ρ_g put 0.90 on the unperturbed regime and divide 0.10 over the perturbed ones. The recording budget is 16 episodes for every estimator and every rung. Conservation is read at a tolerance of 0.05. The per-edge transmission coefficient ranges over $[0, 3]$ and equals one under the generator. Main-table cells are means over ten independent draws; the return-path rows are means over five seeds.

A.14 Credit-fidelity table

This section provides the full credit-fidelity table and the reading conventions behind it, extending the main results. Each cell is the mean over ten independent draws with the across-draw standard deviation in parentheses, and Avg is the mean over the three lag scales, higher being better. Every estimator lays its credit density over the same reference support and is truncated by the same window and ancestor depth. The three kernel-fitting rows are scored against the reference-mass denominator; the three that fit no kernel normalise over the set their own window and ancestor depth reached and carry no unassigned share. Bold marks this work's rows. The tidal road-traffic environment has at most two parents per receiving point; it is printed here and left out of the cross-environment averages quoted in the text.

A.15 Coefficient-error table

This section provides the column conventions of the coefficient-error table, extending the main results. The transmission-coefficient error is the distance of the recovered per-edge coefficients from the ones that generated the log; the estimators that fit no kernel recover no coefficients and report none. The middle column is each estimator's own credit fidelity divided by that of the reference kernel held to the same deployer budget, and it is a relative reading against a reference that is itself truncated. It is a scale and not a ceiling: the reference kernel reads 1.0000 against the ground truth at the reference budget, and truncating it to the deployer budget costs it credit fidelity that an estimator's own errors can offset, and a cell can read above 1.0000. The last column is the unassigned share under the shared reference-mass denominator. The three estimators that fit no kernel normalise over

the set their own budget reached and carry no reference mass; they have no share to report and the cell is left empty.

Read at the deployer budget and averaged over the two fan-in environments, the worst-regime criterion sits at 0.9861 on that middle column, the aggregate-fit kernel at 0.9319 and the regime-collapsed null at 0.9440, and two of the fifty-four underlying cells read above 1.0000. On the coefficient error the separations are 0.1251 over the aggregate fit and 0.1091 over the null.

A.16 Credit-fidelity figure

This section provides the reading of the credit-fidelity figure, extending the main results. The figure carries one panel per controlled environment and one series per estimator, six in each panel, and every panel repeats the full comparison on a different propagation topology. The vertical axis is credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ against the counterfactual-replay ground truth and the horizontal axis carries the lag-scale multiplier, 1.0 for short, 2.0 for medium and 3.5 for long. Each point is the mean over ten independent draws and each error bar is the across-draw standard deviation; the figure reads the rows of Table 8.

Averaged over the six fan-in cells the three estimators that assign credit without fitting a kernel read 0.5183, 0.4478 and 0.3974. The margins this work holds there are 0.0410 over the aggregate fit and 0.0294 over the null, which this work trails by 0.0002 at the long scale of tidal road traffic. That environment gives every receiving point at most two parents, and its three kernel-fitting averages lie 0.0017 apart; it is left out of the averages above.

A.17 Held-out family table

This section provides the setting of the held-out table, extending the main results. The held-out family is structurally distinct, carries heavy-tailed edge lags and is held apart from every other block; its receiving points carry four parents at most and 3.17 on average, and fan-in competition is present there too. The last two columns are the transmission-coefficient error and the reading against the generating kernel at the same budget, both averaged over the three scales and the ten draws. Averaged over the three scales within each draw, the ten draw-level credit-fidelity differences over the aggregate fit have mean 0.0210 and standard error 0.0075, and nine of the ten are positive.

A.18 Ablation table

This section provides the column conventions of the ablation table, extending the ablation study. Each variant changes exactly one mechanism and leaves the others in place. The default configuration is the reference of every block; the four-regime row, the per-edge lag row and the time-varying kernel row carry one reading printed three times: all three are the same setting, four regimes expanded with the per-edge lag index and a kernel that is not frozen. Avg is the mean over the lag scales, on every variant. The degeneracy column is the rank agreement between the credit the kernel assigns and the return already arriving at that emitter. The peak-error column is the distance of the credited peak from the ground-truth peak in timesteps. Both are given as the short, medium and long readings in the order of the columns beside them, and both are carried on every variant. Every reading is a single draw on the base seed, except the two return-path rows, which are means over five seeds at each scale.

A.19 Budget-ladder table

This section provides the column conventions of the budget ladder, extending the ablation study. The points-over-tolerance column of each system is the distributional reading beside the pooled one: the share of individual receiving points whose own unassigned share exceeds the conservation tolerance. The share is defined for the three estimators that carry reference-mass accounting; the time-decayed redistribution carries none and reports neither column. Every reading in this table is a single draw on the base seed.

The time-decayed redistribution reads a higher credit fidelity than this work at the three shortest windows, 0.6250, 0.6929 and 0.7714 against 0.1143, 0.2143 and 0.4429, and is flat at 0.7714 from $B{=}4$ upward; this work passes it at $B{=}8$ and $B{=}16$, at 0.8750 and 0.9500.

Table 9: Unassigned share $\omega_\emptyset$, credit fidelity ρ and points over tolerance (Pts.) across five window rungs, one draw on the base seed. Bold marks this work.

	RICK (ours)			Aggregate-fit kernel			Time-decayed	Regime-collapsed (G=1) null		
B	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Pts.↓	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Pts.↓	$\rho \uparrow$	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Pts.↓
1	**0.8511**	**0.1143**	**1.0000**	0.8514	0.0821	1.0000	0.6250	0.8525	0.1393	1.0000
2	**0.7403**	**0.2143**	**1.0000**	0.7406	0.1964	1.0000	0.6929	0.7409	0.2821	1.0000
4	**0.5217**	**0.4429**	**1.0000**	0.5218	0.3893	1.0000	0.7714	0.5200	0.4036	1.0000
8	**0.1700**	**0.8750**	**0.9214**	0.1703	0.8393	0.9143	0.7714	0.1670	0.8536	0.9143
16	**0.0038**	**0.9500**	**0.0000**	0.0038	0.9214	0.0000	0.7714	0.0036	0.9357	0.0000

A.20 Ablation and ladder figures

This section provides the reading of the two ablation figures, extending the ablation study; Figure 8 and Figure 9 carry the same four panels at full size. The axes of the four body panels are abbreviated there and read as follows: share is the pooled unassigned share beside the share of receiving points over the conservation tolerance, B is the backtracking window, $\rho(\hat{c}, \Phi)$ is credit fidelity against the counterfactual-replay ground truth, G is the number of regimes in the committed family, and piled credit is the agreement between the credit a kernel assigns and the return already arriving at that emitter. In the per-component figure, panel (a) sweeps the regime count at three edge-lag scales and panel (b) places each seed by how much credit it piles where team return already arrives against the fidelity it attains, with and without the return decoupling. The ladder figure reads the backtracking window on a log axis. Its left panel carries two readings of the same run: the pooled unassigned share, which is the summed unreached mass over the summed arrived mass, and the share of individual receiving points whose own unassigned share exceeds the conservation tolerance of 0.05. The point-level reading first falls below one at a window of eight, where it reads 0.9214, and the pooled ratio understates how widely the shortfall is spread: at a window of four it reports that less than half the mass was reached, on a run in which every receiving point is still short. The other two kernel-fitting systems agree with this work on the pooled share to within 0.0030 at every rung and are not drawn there, and the time-decayed redistribution carries no reference-mass accounting and has no series in that panel.

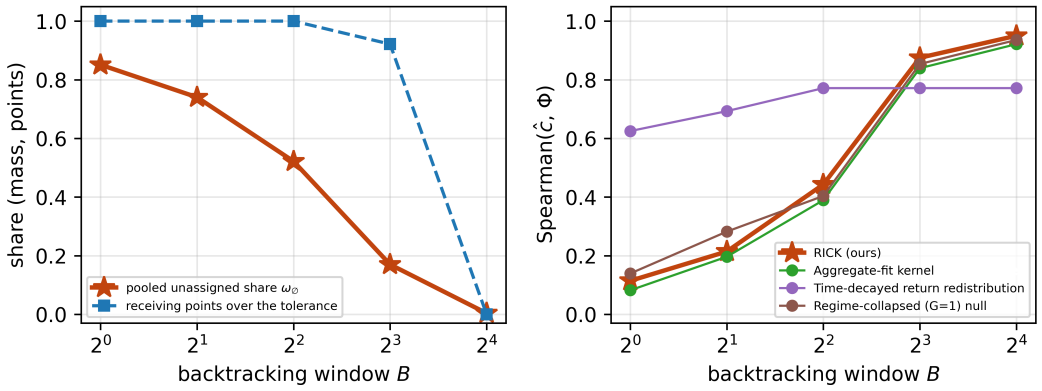


Figure 8: The budget-ladder figure, enlarged. Left: the pooled unassigned share beside the share of receiving points over the tolerance. Right: credit fidelity.

A.21 Lag and kernel variants

This section provides the two delayed-expansion variants of the ablation, extending the ablation study. The frozen-kernel variant evaluates the edge kernels once, one timestep before the receiving point, instead of at each visited step; the collapsed-lag variant replaces every edge's lag profile with the mean profile across edges. Both are read on credit fidelity and on the peak-error column, which measures the distance of the credited peak from the ground-truth peak in timesteps at the short, medium and long scales in that order.

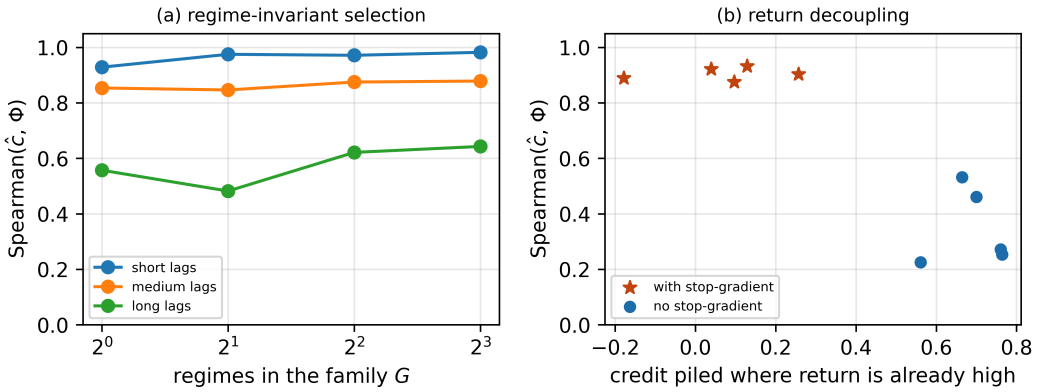


Figure 9: The per-component ablation figure, enlarged. (a) credit fidelity against the regime count. (b) credit piled where return already arrives against the fidelity attained.

The collapsed-lag variant reads above the per-edge index on fidelity at the short and medium scales by 0.0072 and 0.0250 and below it at the long scale by 0.0143; on peak placement it is closer to the ground-truth peak at the short scale by 1.3333 timesteps and at the long scale by 1.5333, and farther at the medium scale by 0.8000. The frozen-kernel variant is behind on both columns at all three scales. Its wall clock reads 602.4, 229.9 and 337.6 milliseconds against 473.9, 203.0 and 779.3 for the time-varying kernel, higher at two scales and lower at the third.

A.22 Robustness sweeps

Three sweeps sit here. Table 10 moves the autocorrelation of the shared exogenous injection over three levels at ten seeds each and reports the two kernel-fitting estimators it carries as a mean and an across-seed standard deviation. Table 12 sweeps the backtracking budget over five rungs at all three edge-lag scales and carries the unassigned share beside the fidelity; a window that under-reaches appears as an unattributed remainder instead of being renormalised away: at $B{=}1$ on the long scale it leaves 0.9190 of the reference mass unassigned at a fidelity of 0.0250, and at $B{=}16$ it leaves 0.0179 at 0.9536. Table 15 fixes the environment and moves the perturbation strength over five levels including zero, with a random mask at matched return cost at every level as the null. The return-cost column is the cost each row is charged, running from 0.00000 at zero strength to 0.03363 at $\varepsilon{=}7.00$; every matched-cost row is charged within 0.00099 of the schedule it is set against, which is what makes the fidelity columns comparable.

The autocorrelation sweep reads 0.7096, 0.7300 and 0.7197 for the aggregate fit at across-seed standard deviations of 0.1184, 0.1105 and 0.1547, against 0.7678, 0.7811 and 0.7875 for the worst-regime criterion at 0.0910, 0.0822 and 0.0873, ahead in twenty-six of the thirty paired readings.

Table 13 sits beside these three and decides what the schedule objective selects. The greedy it specifies, the index round-robin that fills the same slot and a log-determinant reading of the same design each select a different family. On the identification column, which reads the pooled parent-design eigenvalue the collected family actually carries, the objective-selected family buys less than the round-robin at every perturbed strength: 0.0083 against 0.0103, 0.0452 against 0.0647, 0.1048 against 0.1465 and 0.1350 against 0.1947. On credit fidelity it reads above the round-robin at two of the four strengths, 0.8893 against 0.8679 and 0.9036 against 0.8929, and below at the other two, 0.8893 against 0.9107 and 0.8786 against 0.8929. It is the cheapest rule at every strength, charged 0.00043, 0.00449, 0.01596 and 0.02743 against 0.00045, 0.00528, 0.01871 and 0.03264; against the random mask it reads below at three of the four and level at the fourth, and that rule is cost-matched to the round-robin family rather than to the objective-selected one, and the comparison is at equal strength and unequal cost. Over the ten draws of the primary environment the objective-selected family reads 0.9532, 0.7700 and 0.5461 across the three edge-lag scales against 0.9532, 0.7811 and 0.5575 for the round-robin family. Coverage separates them in Table 14: the round-robin partition reaches all fifteen emitting nodes, the objective-selected family leaves one unexcited and the log-determinant variant reaches six of the fifteen, counted over the same emitters.

Table 10: Credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ against the autocorrelation of the shared exogenous injection, at matched update count and matched sampling budget, as the mean and the across-seed sample standard deviation, taken with divisor $n - 1$, over ten seeds.

Autocorrelation	RICK mean	RICK sd	Agg.-fit mean	Agg.-fit sd	Agg.-fit residual
low	0.7678	0.0910	0.7096	0.1184	0.0675
mid	0.7811	0.0822	0.7300	0.1105	0.0657
high	0.7875	0.0873	0.7197	0.1547	0.0692

Table 11: Paired readings behind the two paired comparisons quoted in the text. Each row pairs this work against the aggregate-fit kernel on readings that share a seed, and the difference is taken within a pair rather than between two independent means: the held-out supply-tree family over three edge-lag scales by ten draws, and the driver-autocorrelation sweep over three levels by ten seeds. The mean and the sample standard deviation, taken with divisor $n-1$, are over the paired differences themselves, and the last column counts the pairs in which this work reads above the aggregate fit.

Paired comparison	Readings	Mean difference	SD	Ahead in
Held-out supply tree, three edge-lag scales by ten draws	30	0.0210	0.0236	27
Driver autocorrelation, three levels by ten seeds	30	0.0590	0.0873	26

Table 10 sweeps the driver autocorrelation over three levels at ten seeds. The aggregate-fit spread across the three levels is 0.0204 against across-seed standard deviations between 0.1105 and 0.1547, while the worst-regime criterion is ahead at all three levels and ahead by 0.0590 over thirty paired readings.

A.23 Cost model

This section provides the cost model of the expansion, extending the analysis. The expansion of Eq. 12 is run back from each receiving point over the full ancestor set and every lag up to the lag bound, in one sweep over the topological order, and the window and the ancestor depth are applied to the result. Its cost per receiving point is $O(\#\text{edges} \times L^{\max} \times \#\text{timesteps})$ and does not depend on B.

A.24 Discussion

One selection principle produces the separate readings reported above. The kernel is pinned by the least favourable manufactured regime rather than by aggregate goodness of fit. That principle shows up where the confounding acts, on the recovered transmission coefficients, and on credit fidelity, where the ordering over the aggregate fit holds in all nine cells. This paper places the reading beside two lines it departs from: credit schemes that divide an arrived outcome by a game-theoretic share of the recorded play (Li et al., 2025c; Ji et al., 2025; Ding et al., 2025), and value-decomposition schemes that factor team return into per-agent terms fitted on the same logs (Fu et al., 2025; Wang et al., 2025a; Hu & Krishnamachari, 2026; Xiao et al., 2024). Both read whatever variation the logs already carry, and neither separates transmission from a driver the logs never varied. The separation follows fan-in rather than autocorrelation: the aggregate reading does not decay as the injection autocorrelation of the environment grows, and the separation the principle buys grows instead with the fan-in of the receiving points, which is where collinear co-parents leave the coefficients unpinned.

The three components each close one failure the other two leave open. Taking the accounting denominator over the reached set alone hides the mass a bounded window cannot trace, and the ladder of Table 9 reads that mass where the truncated denominator reports full conservation at every rung. The modelled return leak collapses credit fidelity from 0.9043 to 0.3486, the five-seed means printed in the return-path rows of Table 3, while the return-driven degeneracy read over the same five seeds rises from 0.0686 to 0.6900. Random masks cost-matched to the index round-robin read above the scheduler's masks on credit fidelity at three of the four perturbed strengths of Table 15, and the objective-selected family carries less pooled identification than that indexing at every perturbed

Table 12: The unassigned share $\omega_\emptyset \downarrow$ and credit fidelity $\rho \uparrow$ over the five-rung backtracking-budget ladder at three edge-lag scales on the primary environment, taken against the reference-budget denominator. A window that under-reaches raises the share instead of renormalising the shortfall away. Every reading in this table is a single draw on the base seed.

	Short lags			Medium lags			Long lags		
B	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Peak err.$\downarrow$	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Peak err.$\downarrow$	$\omega_\emptyset \downarrow$	$\rho \uparrow$	Peak err.$\downarrow$
1	0.6406	0.3500	4.533	0.8511	0.1143	2.933	0.9190	0.0250	3.200
2	0.4470	0.5607	4.667	0.7403	0.2143	6.867	0.8492	0.1214	8.000
4	0.1944	0.9321	2.933	0.5217	0.4429	2.667	0.6853	0.2286	3.200
8	0.0203	0.9714	1.333	0.1700	0.8750	4.200	0.3106	0.6214	6.733
16	0.0000	0.9679	0.000	0.0038	0.9500	0.000	0.0179	0.9536	0.000

Table 13: What the schedule objective of Eq. 9 buys, on the primary environment at the medium edge-lag scale, over five strengths including zero. The objective-selected family is read against the index round-robin that fills the same slot in the main tables, a constant strength, and random masks at matched cost. At zero strength there is nothing to select or to align, and the rules that coincide there are printed as the rows the sweep produced. The identification column is the smallest eigenvalue any receiving point's parent design carries over the episodes the family collected, pooled across its regimes, and it reads what the family actually excited. The objective-selected family is the cheapest at every perturbed strength, and it reads above the round-robin family on fidelity at two of the four and below it on identification at all four. The random rule is cost-matched to the round-robin family, not to the objective-selected one, which is charged less again. Every reading in this table is a single draw on the base seed.

ε	Schedule	Fidelity $\uparrow$	Return cost $\downarrow$	Peak err. $\downarrow$	Identification $\lambda_{\min} \uparrow$
0.00	objective-selected masks	0.8321	0.00000	4.467	0.0015
0.00	random masks at matched cost	0.8321	0.00000	4.467	0.0015
0.50	objective-selected masks	0.8893	0.00043	4.467	0.0083
0.50	round-robin masks	0.9107	0.00045	4.200	0.0103
0.50	constant strength	0.9036	0.00045	4.200	0.0095
0.50	random masks at matched cost	0.9036	0.00045	4.467	0.0075
1.50	objective-selected masks	0.8893	0.00449	4.200	0.0452
1.50	round-robin masks	0.8679	0.00528	4.200	0.0647
1.50	constant strength	0.8679	0.00527	4.200	0.0616
1.50	random masks at matched cost	0.8893	0.00509	4.200	0.0459
3.50	objective-selected masks	0.8786	0.01596	4.200	0.1048
3.50	round-robin masks	0.8929	0.01871	4.200	0.1465
3.50	constant strength	0.8929	0.01892	4.200	0.1435
3.50	random masks at matched cost	0.9071	0.01886	4.200	0.1140
7.00	objective-selected masks	0.9036	0.02743	4.200	0.1350
7.00	round-robin masks	0.8929	0.03264	4.200	0.1947
7.00	constant strength	0.8929	0.03273	4.200	0.1922
7.00	random masks at matched cost	0.9071	0.03363	4.200	0.1585

strength. What the scheduling stage contributes to the readings above is a family that covers every emitting node inside the return-cost bound, which the round-robin partition supplies and the objective does not. The composed operator reduces to none of its parts.

Every reading is a statement about directed delayed environments of known transmission structure, and the bilevel problem is evaluated at its inner level with the outer level supplying the cost bound the schedule is charged against. The stop-gradient variant of Table 3 models the return leak as a rescaling of the selected coefficients rather than training without the stop-gradient. The lag profile of Eq. 10 carries a timestep index whose array does not vary with the timestep; every reading here is taken at a profile constant in time. On credit fidelity the held-out paired difference over an aggregate fit is 0.0210 in Table 11, while on the coefficient error the criterion is ahead at 0.0539 against 0.1151.

Table 14: Emitter coverage of the three mask-selection rules on the primary environment, counted over the masks each rule selects under the same return-cost bound. An emitter is covered if some mask in the family excites it. Bold marks the round-robin family, the one whose masks partition the emitters.

Selection rule	Emitters	Covered	Left unexcited
Index round-robin	**15**	**15**	**0**
Objective-selected greedy	15	14	1
Log-determinant variant	15	6	9

Table 15: Credit fidelity $\rho(\hat{c}, \Phi) \uparrow$ against the team return the schedule cost $\downarrow$, over five perturbation strengths including zero and three schedule configurations, which between them use two mask-selection rules, on the primary environment at a single draw on the base seed. The masks here are the index round-robin, which Table 13 reads against the family the schedule objective selects. At zero strength the regime term adds nothing, and the rules that coincide are printed as the two rows the sweep produced and both read 0.8321; that level is the floor against which every separation above it is to be read. The peak-error column takes one of two values, 4.200 and 4.467, at every strength and under every way of choosing the masks including the unperturbed null. The zero-strength level and the random mask at matched cost are the null controls; the scheduler's masks read above the random masks at one of the four perturbed strengths and below at the other three.

ε	Schedule	Fidelity $\uparrow$	Return cost $\downarrow$	Peak err. $\downarrow$
0.00	unperturbed null (scheduled and constant coincide)	0.8321	0.00000	4.467
0.00	random masks at matched cost	0.8321	0.00000	4.467
0.50	round-robin masks	0.9107	0.00045	4.200
0.50	constant strength	0.9036	0.00045	4.200
0.50	random masks at matched cost	0.9036	0.00045	4.467
1.50	round-robin masks	0.8679	0.00528	4.200
1.50	constant strength	0.8679	0.00527	4.200
1.50	random masks at matched cost	0.8893	0.00509	4.200
3.50	round-robin masks	0.8929	0.01871	4.200
3.50	constant strength	0.8929	0.01892	4.200
3.50	random masks at matched cost	0.9071	0.01886	4.200
7.00	round-robin masks	0.8929	0.03264	4.200
7.00	constant strength	0.8929	0.03273	4.200
7.00	random masks at matched cost	0.9071	0.03363	4.200

Several directions follow from these readings. Fitting a digital twin to a real operational site's incident and operation logs would carry the counterfactual-replay ground truth onto an operational delayed environment and turn the agreement between operator judgment and true contribution into a site-level reading. Extending the population-limit uniqueness result beyond the linear one-step case to nonlinear and multi-step propagation would widen the regime-invariant selection account. The sensitivity-directed schedule bisects for the cost bound once per round and then reads the current policy's strength-response; a closed-form update from the measured sensitivity curve to the admissible strength would remove the inner bisection and let the schedule track a policy that shifts within a round. The unassigned-share accounting points to a budget that grows only where the share is positive: an adaptive window keyed to the per-node unassigned mass would spend expansion on the receiving points that actually lose credit rather than on a uniform window. Carrying the worst-regime criterion to an environment whose exact ground truth is only approximable would test the accounting where replay is inexact.