A Contrastive Diagnostic of Target-Synchronization Sensitivity

Anonymous authors
Paper under double-blind review

Abstract

We study whether the learned-function effects of delaying continuation-target synchronization by one update can be anticipated under information-equivalent representations. We focus on ordering finite response contrasts between prespecified temporal-credit configurations across training realizations; absolute response prediction remains unresolved. Contrastive Synchronization Tangent Energy (CSTE) is a diagnostic that takes the difference between the configurations' squared function-space tangent norms, propagating first-order sensitivity through the complete training recursion. It is computed from current-side training artifacts and frozen before corresponding lagged responses are generated. In a delayed tracking task with 48 fresh schedule twins (pairs of independent training schedules), prospective whole-twin concordance was 0.81120, with Bonferroni-adjusted 98.333% confidence interval $[0.77474, 0.84505]$. A separately frozen external test supported learned-function contrast ordering (0.9375, separate 99% interval $[0.87495, 0.98958]$). In the external policy-endpoint assessment, the synchronization-shift interval lay wholly within the registered practical-equivalence region, so the joint endpoint replication criterion was not met.

1 Introduction

Bootstrapped value learning uses learned continuation values (Sutton, 1988; Mnih et al., 2015). Our *one-update continuation-target synchronization lag* uses previous parameters (One-Lag) instead of current ones (Current), retaining bootstrapping. Under information-equivalent representations, can the resulting finite learned-function response contrasts between prespecified temporal-credit configurations be ordered before observation?

Representation geometry and TD dynamics (Dabney & Thomas, 2014; Lyle et al., 2022), target-update stability and function-space target construction (Lee & He, 2019; Fellows et al., 2023; Asadi et al., 2024), and differentiation through training (Maclaurin et al., 2015; Franceschi et al., 2018; Xu et al., 2018) are established. We use these primitives for response-blind diagnosis, claiming no firstness.

The broader *Return Bridge* names the path from local learning signals through actions and trajectories to non-immediate terminal Game Return. We distinguish response levels within a configuration from contrasts between configurations. The earlier absolute-response test of the former remained unresolved. CSTE was defined afterward and frozen before fresh N1 responses. Our target is which independent *training realization*—data order, random events, and optimization path—has the larger finite learned-function synchronization-response contrast.

Infinitesimal synchronization contributions propagate through all later updates to form the final function's tangent (first-order sensitivity). *Contrastive Synchronization Tangent Energy* (CSTE) is the difference between two prespecified configurations' squared tangent norms: a diagnostic, not an algorithm, frozen from Current-only artifacts before corresponding One-Lag responses (Figure 1).

N1's 48 fresh twins yielded primary concordance 0.81120, Bonferroni-adjusted 98.333% CI $[0.77474, 0.84505]$ (supported; Figure 4). Six retrospective CSTE–reference differences have intervals above zero; the 1,200-update comparison remains unresolved (Section 5.3). External learned-function ordering was supported, while practical equivalence falsified the material synchronization endpoint-shift gate; the joint endpoint criterion was not met (Figure 5).

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

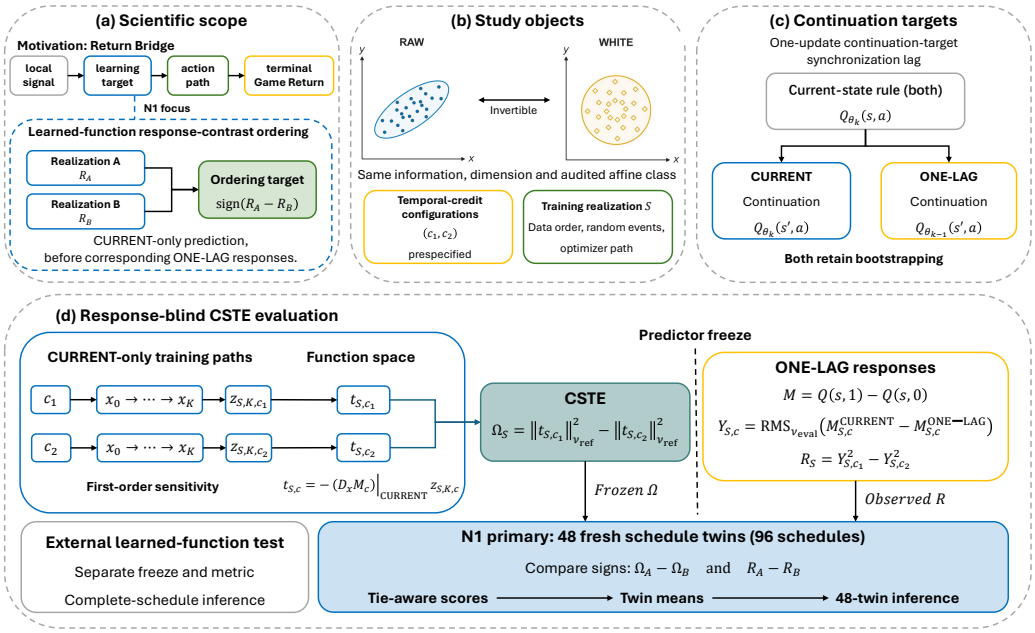

Figure 1: Response-blind learned-function ordering. (a) Return Bridge motivates A/B response-contrast ordering. (b) Prespecified temporal-credit configurations use information-equivalent RAW/WHITE coordinates. (c) Current/previous-update continuation parameters retain bootstrapping. (d) Complete CURRENT paths yield tangent-energy contrasts Ω, frozen before corresponding finite contrasts R. N1 uses disjoint reference/evaluation fixtures and tie-aware A/B concordance: 16-row means per twin, then 48-twin averaging and inference. The external test has its own metric, freeze, and complete-schedule inference. Absolute prediction remains unresolved.

Contributions. Full-recursion CSTE with prospective learned-function ordering evidence is primary. Separate structural findings establish regime-dependent geometry effects on relative temporal-credit terminal efficacy in tested Tetris and controlled realization materiality. M1 supports a bounded synchronization mean component; its log-SD shift remains unresolved. Absolute magnitude, policy utility, mediation, and remainder dominance remain separate questions.

2 Problem Setup and Claim Scope

Let $g \in \{\mathrm{RAW}, \mathrm{WHITE}\}$ denote coordinate geometry, c a prespecified temporal-credit configuration, r the representation regime, and S a complete training realization (data order, random events, optimizer path). These symbols align vocabulary, not studies: each retains its platform, estimand, outcome, and inferential unit.

Within r, RAW/WHITE are information-equivalent: their invertible, dimension-preserving mapping preserves sampled collision classes and the audited affine function class; paired learners start with identical Q functions. Only coordinate scale, correlation, and conditioning—*representation geometry*—change (audits: Appendix B).

For action-value function Q_θ, synchronization endpoints use different parameter versions for next-state continuation:

$$\textsc{Current} : Q_{\theta_k}(s', a), \qquad \textsc{One-Lag} : Q_{\theta_{k-1}}(s', a). \tag{1}$$

Both evaluate the current state with Q_{θ_k} and retain bootstrapping with the same stop-target semi-gradient semantics. This is one-update continuation-target synchronization lag, not a new target-network method. Section 4's interpolation is for analysis, not a training hyperparameter.

Let $t_{S,c}$ be the complete-recursion one-sided function tangent on frozen predictor fixture ν_{ref}. For a prespecified ordered pair,

$$\Omega_{S;c_1,c_2} = \|t_{S,c_1}\|^2_{\nu_{\text{ref}}} - \|t_{S,c_2}\|^2_{\nu_{\text{ref}}} \tag{2}$$

is response-blind CSTE. The squared tangent norm $Q^2_{S,c}$ is an energy, distinct from action-value Q_θ.

Prediction objects. $Y_{S,c}$ is the RMS of the finite CURRENT–ONE-LAG function difference on ν_{eval}; $R_{S;c_1,c_2} = Y^2_{S,c_1} - Y^2_{S,c_2}$ contrasts configurations within S. N1 primary compares signs of $\Omega_A - \Omega_B$ and $R_A - R_B$ for independent realizations A, B.

N1 uses two-action margins; external uses centered three-action scores, each with its own metric and separately frozen, disjoint $\nu_{\text{ref}}/\nu_{\text{eval}}$ (Sections 5–6). Section 4's theoretical bound instead uses a common ν.

G denotes the external normalized 128-step policy endpoint (Equation 27); GAME RETURN is Tetris's terminal outcome. CSTE diagnoses a fixed learner's contrast ordering. Absolute response, policy utility, terminal outcomes, and mediation are outside its supported scope.

N1 preview (Table 1). Primary T_{N1} averages tie-aware twin-difference concordance over 16 rows/twin, then 48 twins. Secondary H (18 versus 6) and λ (.6 versus 0) concordances compare Ω with R within realizations. These three CIs share one Bonferroni family; Section 5 gives the scoring rule. Throughout, adjusted confidence levels denote nominal coverage per interval, including those printed in the figures; the corresponding Bonferroni families target 95% familywise coverage. Percentile-bootstrap coverage is approximate.

3 STRUCTURAL SENSITIVITY AND CONTROLLED PERTURBATION

H2, M0b, and M1 estimate distinct structural effects (Appendix Table 7).

Information-equivalent geometry. H2 crossed exact eight-feature Tetris afterstates (R0) and their componentwise bin-width-32 coarsening (R3) with RAW/WHITE and TD(0)/TD($\lambda = .9$) on 512 fresh paired reset blocks (Appendix B). Within each representation r, the estimand is the RAW-to-WHITE change in the terminal GAME RETURN contrast TD($\lambda = .9$)–TD(0). With Bonferroni-adjusted 98.333% CIs, the R0 interaction was -1980.45 $[-2599.68, -1361.22]$ (SUPPORTED); R3 was 0.0195 $[-0.0162, 0.0552]$ (NOT ESTABLISHED). R3 establishes neither zero nor equivalence.

Both R0 means rose under WHITE, so the negative interaction does not mean harm to TD($\lambda = .9$). R3-minus-R0, 1980.47 $[1361.24, 2599.70]$ (SUPPORTED), establishes regime dependence of this causal effect on relative temporal-credit efficacy, not a universal direction.

Complete training realization. BASE is a delayed binary-choice task: the initial choice determines a terminal reward of 0 or 1 after a fixed delay. With AdamW, $M_H(S)$ averages the registered geometry–credit–horizon contrast in this return across information regimes (Figure 3; Appendix C.2). M0b's 16 fresh schedules gave a schedule-component SD of 0.13423, 95% CI $[0.08715, 0.16693]$, with lower bound above the registered 0.02 threshold (ESTABLISHED). Inference resampled schedules and shared evaluation blocks (Appendix C). The earlier five-schedule screen remained unresolved, motivating precision only.

Bounded synchronization component. M1 paired CURRENT and ONE-LAG within 48 fresh complete schedules on BASE/AdamW. The schedule-averaged $M_H(\text{CURRENT}, S) - M_H(\text{ONE-LAG}, S)$ was 0.05551, frozen simultaneous CI $[0.02143, 0.09063]$ (SUPPORTED). The log-SD shift, -0.00351 $[-0.25889, 0.26249]$, remained UNRESOLVED, establishing neither equal variance nor attenuation. The frozen secondary mean family supported LOW and HIGH-minus-LOW, with HIGH unresolved: information conditioning, not a LOW mechanism or universal CURRENT advantage.

We next study learned-function sensitivity (Section 4); Section 6 separately assesses external policy endpoints.

Table 1: Frozen N1 learned-function ordering: 48 complete twins, 20,000 whole-twin bootstrap replicates, and Bonferroni-adjusted CIs: nominal 98.333% per interval, targeting 95% familywise coverage. The 0.5 population boundary does not classify individual twins; 768 internal rows form 48 twin statistics, not independent observations.

Role	Estimand	Estimate	Adjusted 98.333% CI	Status
Primary	T_{N1}	0.81120	$[0.77474,\ 0.84505]$	SUPPORTED
Secondary	Horizon	0.89714	$[0.86849,\ 0.92448]$	SUPPORTED
Secondary	λ	0.88411	$[0.84766,\ 0.91667]$	SUPPORTED

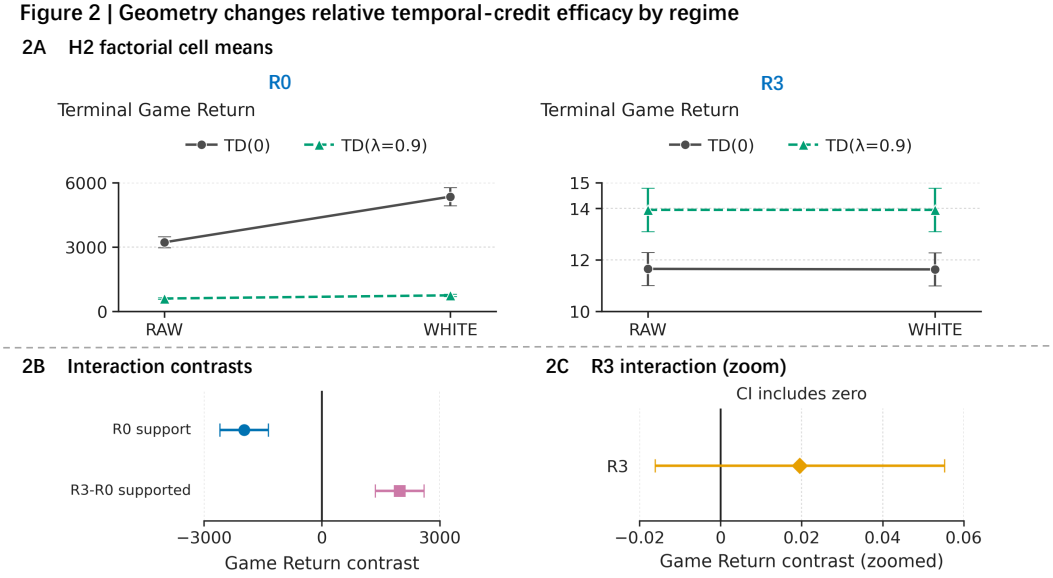

Figure 2: Geometry effects on relative temporal-credit efficacy in H2 Tetris. (A) Eight terminal GAME RETURN means with descriptive 95% CIs from 512 paired reset blocks ($n = 512$/arm), on separate regime scales. RAW/WHITE preserve information, dimensionality, and the audited affine class. (B) R0 interaction and R3-minus-R0 contrast. Interactions measure the WHITE-minus-RAW change in TD($\lambda = .9$)-minus-TD(0) return. (C) R3 uses a separate zoomed scale. B–C labels give nominal 98.333% coverage per interval; the three Bonferroni-adjusted intervals target 95% familywise coverage. R0's negative interaction coexists with higher means for both learners under WHITE. R3 NOT ESTABLISHED implies neither zero nor equivalence.

4 FULL-RECURSION CONTRASTIVE SYNCHRONIZATION SENSITIVITY

Local forcing and absolute-response prediction. For schedule S, update k, and configuration c, let $C_{S,k,c}$ be the frozen eligibility–next-feature coupling, and $\Delta\theta_{S,k,c} = \theta_{S,k,c} - \theta_{S,k-1,c}$. At a common state and trajectory, the frozen affine learner obeys

$$g^{\text{CURRENT}}_{S,k,c} - g^{\text{ONE-LAG}}_{S,k,c} = -C_{S,k,c}\Delta\theta_{S,k,c}, \qquad b_{S,k,c} = C_{S,k,c}\Delta\theta_{S,k,c}. \tag{3}$$

Optimizer states and parameters can diverge across training; this local identity alone predicts neither finite response nor return. M2-V's prospective absolute P1 remains UNRESOLVED. Its auxiliary P2-H had 384/384 positive signs and limited anti-triviality; mixed-sign P2-λ was more discriminative. M2-V tested a local predictor. CSTE was defined afterward, without fitting M2-V responses, and frozen before fresh N1 responses.

Complete learner recursion. Set $\theta^{\text{tgt}}_k(\eta) = (1-\eta)\theta_k(\eta) + \eta\theta_{k-1}(\eta)$ at every update, so $\eta = 0$ is CURRENT and $\eta = 1$ is ONE-LAG. Let $x = (\theta_k, \theta_{k-1}, m_k, v_k)$ contain the state needed by the next AdamW update and $z_{S,k,c} = \mathrm{d}x_{S,k,c}/\mathrm{d}\eta|_{0^+}$. Where the complete update admits a joint linear

Figure 3 | Training realization is material; synchronization-lag mean component is bounded

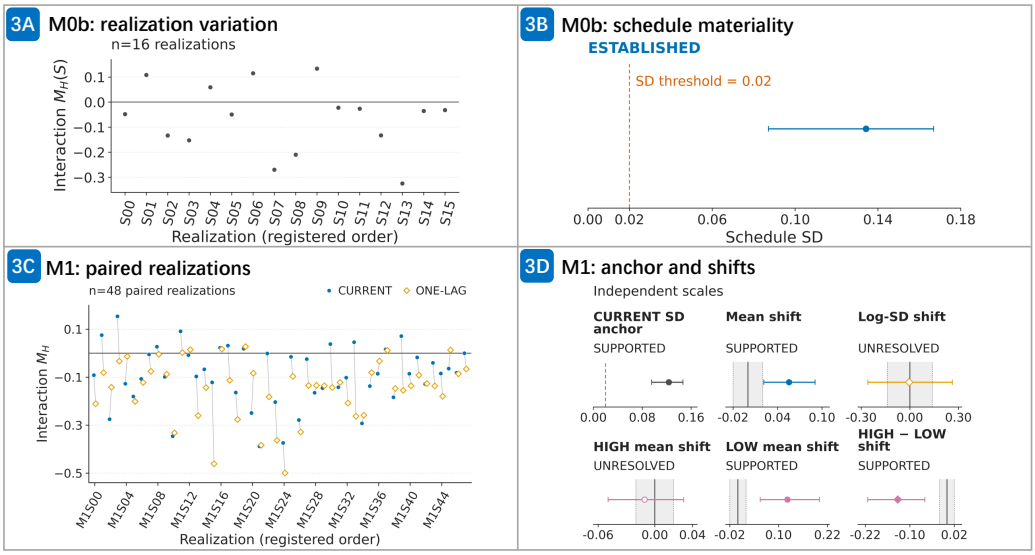

Figure 3: Realization heterogeneity and synchronization on BASE/AdamW. M_H is the HIGH/LOW-averaged geometry–credit–horizon interaction (Appendix C.2). (A–B) M0b: descriptive realization values and schedule variance-component SD. (C) M1: within-realization CURRENT/ONE-LAG pairs, not trajectories; both bootstrap, with a one-update continuation-target lag. (D) CURRENT SD anchors materiality at 0.02, as in B; it is not an intervention effect. Mean shifts are schedule averages of CURRENT-minus-ONE-LAG interaction differences, pooled over HIGH/LOW or information-conditional; HIGH-minus-LOW directly compares conditional shifts. Log-SD is the regularized CURRENT/ONE-LAG schedule-SD log ratio. Gray bands mark registered practical-equivalence regions: ± 0.02 for means, ± 0.13889 for log-SD. CIs: 95% for B/SD anchor; 97.5% for the two co-primary shifts; 99.5% for the three displayed conditional shifts from the ten-member secondary family (Bonferroni familywise 95% per family). Crossed bootstrap resamples whole realizations and shared evaluation blocks, preserving M1 pairing. Unresolved log-SD establishes neither zero nor equivalence; information conditioning does not identify a LOW mechanism.

first-order differential on its admissible state domain, Az propagates inherited sensitivity, Jb injects local forcing, and Φ transports it through later updates along the CURRENT path:

$$z_{S,k+1,c} = A_{S,k,c} z_{S,k,c} + J_{S,k,c} b_{S,k,c}, \quad z_{S,0,c} = 0,$$
$$z_{S,K,c} = \sum_{k=0}^{K-1} \Phi_{S,K\leftarrow k+1,c} J_{S,k,c} b_{S,k,c}. \tag{4}$$

A and J linearize state and gradient input. At clipping kinks, the implementation uses the combined direction; the matrix claim requires the stated linearity (Appendix A). The derivative is local at $\eta \to 0^+$; finite $\eta = 1$ response requires separate control. Section 5.3 compares full recursion with fixed windows.

CSTE and the finite boundary. For learned action margin $M_c(x;\chi)$ and predeclared function-space measure ν,

$$t_{S,c}(\chi) = -D_x M_c(x_{S,K,c}(0);\chi) z_{S,K,c}, \quad Q_{S,c} = \|t_{S,c}\|_\nu, \quad \Omega_{S;c_1,c_2} = Q^2_{S,c_1} - Q^2_{S,c_2}. \tag{5}$$

CSTE is a response-blind theory and diagnostic predictor, not an algorithm or absolute predictor. If the finite learned-function difference is $d_{S,c}(1) = t_{S,c} + \epsilon_{S,c}$ with $\|\epsilon_{S,c}\|_\nu \le \rho_{S,c}$, define $R = \|d_{S,c_1}(1)\|^2_\nu - \|d_{S,c_2}(1)\|^2_\nu$ and $E_{S,c} = 2Q_{S,c}\rho_{S,c} + \rho^2_{S,c}$. Then

$$|R - \Omega_{S;c_1,c_2}| \le E_{S,c_1} + E_{S,c_2}, \qquad |\Omega_{S;c_1,c_2}| > E_{S,c_1} + E_{S,c_2} \Rightarrow \operatorname{sign} R = \operatorname{sign} \Omega_{S;c_1,c_2}. \tag{6}$$

This condition is sufficient, not necessary; remainder dominance is unverified. Ordering support establishes neither this condition nor policy/return prediction.

Response-blind freeze. Operational values on ν_{ref}, directions, and SHA-256 identity were frozen from CURRENT-only artifacts while zero corresponding ONE-LAG learner, response, or inference artifacts existed. Only after this barrier were ONE-LAG responses generated on separately frozen, disjoint ν_{eval}. The prospective test evaluates this ordering claim, not the shared-measure sufficient condition.

5 N1 VALIDATION AND RETROSPECTIVE COMPARISONS

5.1 EXPERIMENTAL SETUP

N1 uses a purpose-built delayed tracking task. Each episode draws $\xi \sim \text{Uniform}[-1, 1]$ and holds an assigned action $a \in \{0, 1\}$ for H transitions. Starting at $q_0 = 0$, the state follows $q_{t+1} = .75q_t + .25(2a - 1)$. Rewards are zero until the terminal transition, where $r_H = 1 - .25(q_H - \xi)^2$. HIGH observes ξ; LOW observes its five-bin midpoint quantization.

Eight fixed features plus a bias feed two affine action heads (18 parameters). We cross HIGH/LOW, $H = 6/18$, RAW/invertibly WHITE, $\lambda = 0/.6$, and CURRENT/ONE-LAG: 32 cells per schedule. RAW/WHITE preserve information within each observation regime. From a zero function, each learner performs 12,000 episode TD(λ) updates with $\gamma = .99$, AdamW learning rate 3×10^{-4}, and native gradient-norm clipping at 1. Parameters remain fixed within each episode. Appendix G.1 gives the complete configuration.

The 48 fresh twins (A_i, B_i) contain 96 independent schedules, each with 6,000 assignments per action and identical cue/action order across its cells (3,072 learners total). Complete twins are the inferential units. Predictions from 1,536 CURRENT learners were frozen before training the corresponding 1,536 ONE-LAG learners.

5.2 PROSPECTIVE VALIDATION

After the response-blind barrier in Section 4, finite responses were generated on disjoint ν_{eval}:

$$Y_{S,c} = \text{RMS}_{\nu_{\text{eval}}}[M^{\text{CURRENT}}_{S,c} - M^{\text{ONE-LAG}}_{S,c}], \qquad R_{S;c_1,c_2} = Y^2_{S,c_1} - Y^2_{S,c_2}. \tag{7}$$

$M = Q(s, 1) - Q(s, 0)$ is the action margin and R a learned-function contrast; Appendix G details chronology and fixtures.

For twin i and its 16 registered rows j, define tie-aware concordance by $\kappa(a, b) = 1$ when $ab > 0$, 0.5 when $ab = 0$, and 0 otherwise. The primary whole-twin estimand is

$$D_i = \frac{1}{16}\sum_{j=1}^{16} \kappa(\Omega_{A_ij} - \Omega_{B_ij}, R_{A_ij} - R_{B_ij}), \qquad T_{\text{N1}} = \frac{1}{48}\sum_{i=1}^{48} D_i. \tag{8}$$

Frozen inference used 20,000 whole-twin bootstrap replicates for three two-sided Bonferroni-adjusted 98.333% intervals (nominal per interval; 95% familywise target). The population decision boundary is 0.5; the figure's per-twin 0.5 line is descriptive.

All three clear the boundary (Table 1; Figure 4): the primary supports prospective ordering, with prespecified secondary H/λ contrasts. These differ from M2-V P2-H/P2-λ and retain their own criteria; temporal directions are configuration-specific.

Fresh independent-schedule differences make a purely response-side paired- cancellation account insufficient, without proving cancellation absent or mediation. Support concerns learned-function ordering, not magnitude or utility.

5.3 RETROSPECTIVE BASELINE COMPARISONS

Seven diagnostic references reuse these 48 twins (1,536 CURRENT learners). Constructed from CURRENT-only artifacts, predictors were fixed before scoring existing responses; prior N1 knowledge makes this *retrospective exploratory* (Table 2; Appendix J).

Six intervals lie above zero. B3b, using the last 1,200 of 12,000 tangent updates, has concordance 0.79948, versus CSTE's frozen 0.81120. The paired difference 0.01172 has interval

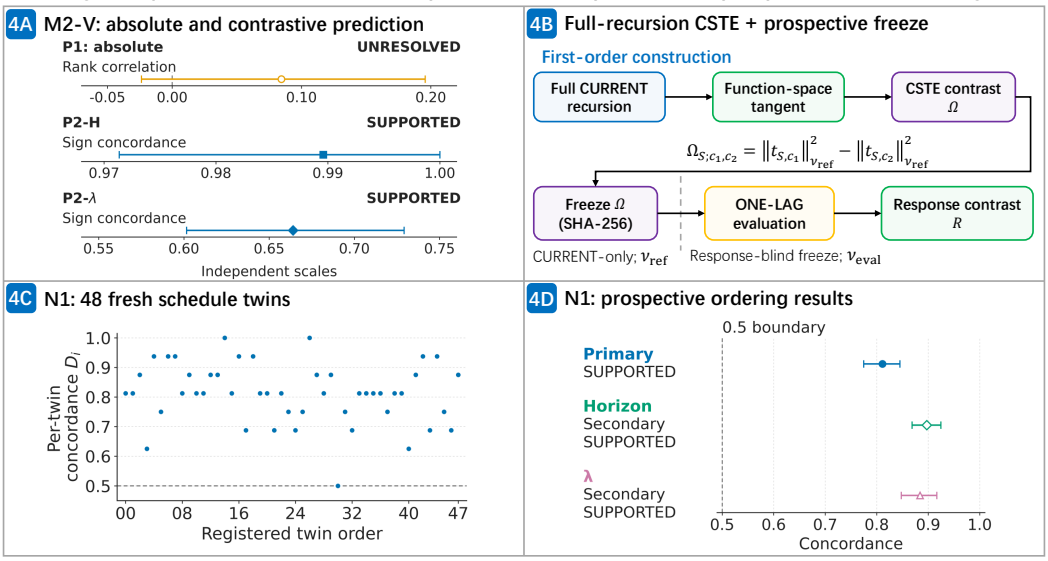

Figure 4: From absolute prediction to contrast ordering. (A) M2-V local predictor: primary P1 rank correlation is unresolved; auxiliary P2 sign concordances do not establish absolute prediction or mechanism. References are 0 for P1 and 0.5 for P2 (outside its zoomed axes). The 99% CIs retain the original five-test family over 48 schedules; P3/P4 remain unresolved (Appendix F.2). (B) CSTE Ω contrasts squared complete-recursion function-space tangent norms for configurations c_1, c_2 within realization S. CURRENT-only predictors on ν_{ref} are frozen before any corresponding ONE-LAG artifacts. Subsequent finite responses use disjoint ν_{eval}; R contrasts squared RMS action-margin differences (Eq. 7). Local tangents are not exact finite responses. (C) Each fresh twin's D_i averages 16 nested sign-concordance scores for $\Omega_A - \Omega_B$ versus $R_A - R_B$: 1 for agreement, $1/2$ for either zero, 0 otherwise. The 0.5 line is descriptive. (D) Primary averages 48 twins (96 schedules). Separate within-realization H/λ estimands are not Primary subsets (Table 1). All three adjusted CIs (nominal 98.333% each; 95% familywise target) have lower bounds above 0.5. A/D use 20,000 percentile bootstrap replicates over whole schedules/twins, respectively. Ordering support establishes neither policy utility nor the shared-measure remainder condition.

Table 2: Retrospective exploratory N1 comparisons: Δ is frozen CSTE minus baseline concordance. Intervals use 20,000 paired bootstrap replicates over 48 whole twins (16 rows/twin). This separate seven-comparison Bonferroni family ($\alpha = .05$) gives nominal 99.285714% intervals via approximate percentile inference. "Positive" means wholly above zero; "Unresolved" is not equivalence. Definitions: Appendix J.

ID	Diagnostic reference	Δ	Paired interval	Status
B1	Mean local forcing energy	0.32943	$[0.26823,\ 0.39323]$	Positive
B2	Last parameter update	0.28906	$[0.22266,\ 0.35286]$	Positive
B3a	Tangent, last 1 update	0.30729	$[0.24479,\ 0.36979]$	Positive
B3b	Tangent, last 1,200 updates	0.01172	$[-0.02344,\ 0.04948]$	Unresolved
B4	Native parameter tangent	0.26953	$[0.20443,\ 0.33464]$	Positive
B5	Fixed direction (twin tie)	0.31120	$[0.27083,\ 0.35026]$	Positive
B6	Fixed random ± 1	0.30534	$[0.25260,\ 0.35417]$	Positive

$[-0.02344, 0.04948]$: UNRESOLVED, not equivalence. B5 has structural twin ties; B6 is a single fixed random reference.

Windows reset tangent state while retaining the complete parameter and optimizer trajectory. Nearby concordance establishes neither sufficient propagation length nor dispensable early training or computational savings (Section 8).

6 External Evaluation

The separately frozen external study crossed RAW/WHITE, TD(0)/TD($\lambda = .9$), and Current/One-Lag in a fixed-horizon two-dimensional discrete-action linear–quadratic domain. Its 48 independent complete schedules—not cells, episodes, states, or transitions—were the inferential units. The policy endpoint, schedule population, and learned-function metric differ from the Tetris, M1, and N1 objects; Appendix H gives the gates.

Table 3: Frozen external decisions (48 complete schedules). Three jointly required core gates use Bonferroni-adjusted CIs (nominal 98.333% each, 95% familywise target); separate learned-function CSTE uses a two-sided 99% CI and is evaluated outside that conjunction. Gate 3 fixes the core status as NOT REPLICATED.

Decision	Estimand	Estimate	CI	Registered rule	Status
Policy-endpoint empirical core: adjusted 98.333% CIs; all three required					
G1 geometry	$\mu_{I,\text{CURRENT}}$	-0.22330	$[-0.24696, -0.20072]$	wholly outside $[-.01, .01]$	SUPPORTED
G2 schedule SD	$\sigma_{I,\text{CURRENT}}$	0.06789	$[0.05277, 0.08003]$	lower $> .02$	SUPPORTED
G3 endpoint	$\Delta\mu_{\text{sync}}$	0.00371	$[0.00236, 0.00540]$	wholly inside $[-.01, .01]$	FALSIFIED
Separate learned-function result: two-sided 99% CI; outside the core					
External CSTE	concordance	0.9375	$[0.87495, 0.98958]$	lower $> .5$	SUPPORTED

The registered core required all three gates (Table 3; Figure 5A–C). Gate 1's negative sign is contrast direction, not WHITE harm; Gate 2 establishes schedule materiality.

Gate 3's interval excludes zero but lies wholly inside the registered practical-equivalence region. The material endpoint-shift claim is therefore FALSIFIED, and the joint replication criterion is not met. This is a domain-specific materiality decision; the controlled M1 result retains its own scope.

Panel D's CSTE predictor was hash-frozen from Current-only paths before corresponding One-Lag artifacts. Its tie-aware concordance clears the separate 0.5 boundary (Table 3; SUPPORTED). This independent outcome and interval family concern learned-function contrast ordering, with utility, absolute response, and mediation outside the supported scope.

7 Related Work and Bounded Novelty

Representation geometry and temporal credit. TD metrics include least-squares, preconditioned, natural-gradient, and curvature-aware updates (Bradtke & Barto, 1996; Boyan, 2002; Yao & Liu, 2008; Yao et al., 2009; Dabney & Thomas, 2014; Romoff et al., 2021); representation studies address interference, auxiliary prediction, and bootstrapping (Bengio et al., 2020; Lyle et al., 2021; 2022; Tang & Munos, 2023; Le Lan et al., 2023). Our coordinate intervention fixes information, dimension, and the affine class. Temporal-credit variants span traces, true-online TD(λ), Retrace, and adaptive or replay-compatible returns (Sutton, 1988; van Seijen & Sutton, 2014; Munos et al., 2016; White & White, 2016; Hernandez-Garcia & Sutton, 2019; Daley & Amato, 2019; Anand & Precup, 2021; Deb et al., 2022; Ramesh et al., 2024; Allen et al., 2024; Elelimy et al., 2025; Daley, 2026). Our study fixes the TD configurations and uses CSTE diagnostically.

Target synchronization and training derivatives. Composed target-update dynamics and matrix splitting explain conditioning and stability (Fellows et al., 2023; Wu et al., 2025); function-space matching constructs target updates (Asadi et al., 2024). Training-error sensitivity, Q-value sensitivity, and TD-error alignment also guide training interventions (Snyder et al., 2023; Ly et al., 2026; Pleiss et al., 2026). Full-training and recursive meta-gradients are established (Maclaurin et al., 2015; Xu et al., 2018). CSTE applies these foundations diagnostically to a fixed learner; Appendix K compares their objects and purposes.

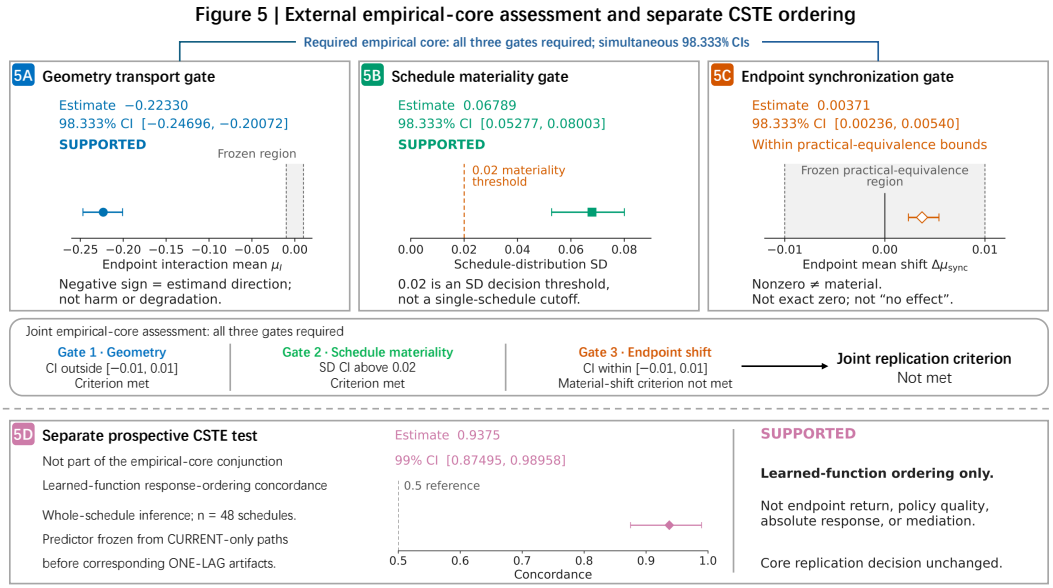

Figure 5: Separate external assessments across 48 complete schedules. Geometry and schedule materiality are SUPPORTED; the material endpoint-shift claim is FALSIFIED under practical equivalence. The core required all three gates; A–C labels give nominal 98.333% coverage per interval (Bonferroni-adjusted; 95% familywise target). Panel D assesses learned-function CSTE ordering with a separate freeze and interval family; its SUPPORTED result leaves the core decision unchanged. Learned-function ordering and policy-endpoint effects have different scopes.

Training realization and inference. RL reproducibility work documents implementation, hyperparameter, environment, seed, and test sensitivity (Islam et al., 2017; Henderson et al., 2018; Colas et al., 2018; Engstrom et al., 2020); later work treats uncertainty, tuning seeds, repeated measures, and design freedom (Agarwal et al., 2021; Eimer et al., 2023; Patterson et al., 2024). Here complete realizations define blocking, effect modification, transport, and inference.

Bounded conjunction. Our conjunction combines information-equivalent coordinates, exact CURRENT-to-ONE-LAG interpolation, full-recursion CSTE, and response-blind prospective ordering. This non-exhaustive comparison locates it among established foundations without a priority claim.

8 Discussion and Conclusion

Response-blind CSTE prospectively ordered finite learned-function response contrasts in fresh twins and a separately frozen external test. Full recursion defines all-update sensitivity, but its ranking advantage over 1,200 updates remains unresolved; equivalence and computational savings were not tested. The external joint endpoint criterion was not met (Section 6). Diagnostic support is bounded to learned-function ordering in the tested populations; absolute response, policy utility, mediation, and empirical remainder dominance remain unvalidated.

AI USE STATEMENT

Generative AI tools, including Codex, assisted throughout the research and writing workflow. Uses included exploring research questions and related literature; developing conceptual and mathematical formulations and proof drafts; proposing and reviewing experimental designs; implementing and debugging the synthetic-task, training, and analysis code; organizing experiment execution and evidence; interpreting results; and drafting, translating, and revising text, tables, and figure materials. Quantitative results came from the recorded computational experiments. Verification included source and artifact checks, numerical conformance and replay checks, and reviews by separate AI agents. These automated reviews are distinguished from human review. The authors take responsibility for the final text, claims, code, and other artifacts, including material developed with AI assistance.

REPRODUCIBILITY STATEMENT

Section 5.1 and Appendix G.1 specify the N1 task, features, optimizer, training schedules, and evaluation fixtures. Appendix A gives the derivative assumptions and derivations; Appendices C, G, and H describe the study protocols, predictor freezes, outcome definitions, and inference rules. Appendix J specifies the retrospective references, numerical runtime, and paired uncertainty procedure. Canonical artifact identifiers and SHA-256 bindings are listed in Appendix I. These records distinguish the original prospective tests from the retrospective supplement and document the provenance of reported estimates.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Cameron Allen, Aaron Kirtland, Ruo Yu Tao, Sam Lobel, Daniel Scott, Nicholas Petrocelli, Omer Gottesman, Ronald Parr, Michael L. Littman, and George Konidaris. Mitigating partial observability in sequential decision processes via the lambda discrepancy. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62988–63028, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/73073ccb3bc559fd001e66b9079d6d5e-Abstract-Conference.html.

Nishanth Anand and Doina Precup. Preferential temporal difference learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 286–296, 2021. URL https://proceedings.mlr.press/v139/anand21a.html.

Kavosh Asadi, Yao Liu, Shoham Sabach, Ming Yin, and Rasool Fakoor. Learning the target network in function space. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1902–1923, 2024. URL https://proceedings.mlr.press/v235/asadi24a.html.

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 767–777, 2020. URL https://proceedings.mlr.press/v119/bengio20a.html.

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Correcting momentum in temporal difference learning, 2021. URL https://arxiv.org/abs/2106.03955. arXiv preprint arXiv:2106.03955.

Justin A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2–3):233–246, 2002. doi: 10.1023/A:1017936530646. URL https://doi.org/10.1023/A:1017936530646.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1–3):33–57, 1996. doi: 10.1007/BF00114723. URL https://doi.org/10.1007/BF00114723.

Fengdi Che, Chenjun Xiao, Jincheng Mei, Bo Dai, Ramki Gummadi, Oscar A. Ramirez, Christopher K. Harris, A. Rupam Mahmood, and Dale Schuurmans. Target networks and over-parameterization stabilize off-policy bootstrapping with function approximation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6372–6396, 2024. URL https://proceedings.mlr.press/v235/che24a.html.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments, 2018. URL https://arxiv.org/abs/1806.08295. arXiv preprint arXiv:1806.08295.

William Dabney and Philip S. Thomas. Natural temporal difference learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1767–1773, 2014. doi: 10.1609/aaai.v28i1.9018. URL https://ojs.aaai.org/index.php/AAAI/article/view/9018.

Brett Daley. Gated Q-learning: Add off-policy bias to taste. *Reinforcement Learning Journal*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper141.html. RLC 2026 pre-proceedings; pagination forthcoming.

Brett Daley and Christopher Amato. Reconciling λ-returns with experience replay. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/9f396fe44e7c05c16873b05ec425cbad-Abstract.html.

Rohan Deb, Meet Gandhi, and Shalabh Bhatnagar. Schedule based temporal difference algorithms. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1–6, 2022. doi: 10.1109/Allerton49937.2022.9929388. URL https://arxiv.org/abs/2111.11768.

Theresa Eimer, Marius Lindauer, and Roberta Raileanu. Hyperparameters in reinforcement learning and how to tune them. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9104–9149, 2023. URL https://proceedings.mlr.press/v202/eimer23a.html.

Esraa Elelimy, Brett Daley, Andrew Patterson, Marlos C. Machado, Adam White, and Martha White. Deep reinforcement learning with gradient eligibility traces. *Reinforcement Learning Journal*, 6: 2530–2550, 2025. URL https://rlj.cs.umass.edu/2025/papers/Paper302.html.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO, 2020. URL https://arxiv.org/abs/2005.12729. arXiv preprint arXiv:2005.12729.

Mattie Fellows, Matthew J. A. Smith, and Shimon Whiteson. Why target networks stabilise temporal difference methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9886–9909, 2023. URL https://proceedings.mlr.press/v202/fellows23a.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Ahmed Hendawy, Henrik Metternich, Théo Vincent, Mahdi Kallel, Jan Peters, and Carlo D'Eramo. Use the online network if you can: Towards fast and stable reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ffbc863edf0a708ee331540517abe0fc-Paper-Conference.pdf.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

J. Fernando Hernandez-Garcia and Richard S. Sutton. Understanding multi-step deep reinforcement learning: A systematic study of the DQN target, 2019. URL https://arxiv.org/abs/1901.07510. arXiv preprint arXiv:1901.07510.

Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control, 2017. URL https://arxiv.org/abs/1708.04133. arXiv preprint arXiv:1708.04133.

Charline Le Lan, Stephen Tu, Mark Rowland, Anna Harutyunyan, Rishabh Agarwal, Marc G. Bellemare, and Will Dabney. Bootstrapped representations in reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18686–18713, 2023. URL https://proceedings.mlr.press/v202/le-lan23a.html.

Donghwan Lee. Target updates may stabilize linear Q-learning: Periodic and soft dynamics, 2026. URL https://arxiv.org/abs/2606.02645. arXiv preprint arXiv:2606.02645.

Donghwan Lee and Niao He. Target-based temporal-difference learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3713–3722, 2019. URL https://proceedings.mlr.press/v97/lee19a.html.

Adrian Ly, Richard Dazeley, Peter Vamplew, Sunil Aryal, and Francisco Cruz. Memory merge DQN: Sensitivity weighted target updates for stable value learning, 2026. URL https://arxiv.org/abs/2607.19397. arXiv preprint arXiv:2607.19397.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1–9, 2021. URL https://proceedings.mlr.press/v130/lyle21a.html.

Clare Lyle, Mark Rowland, Will Dabney, Marta Kwiatkowska, and Yarin Gal. Learning dynamics and generalization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14560–14581, 2022. URL https://proceedings.mlr.press/v162/lyle22a.html.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2113–2122, 2015. URL https://proceedings.mlr.press/v37/maclaurin15.html.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado van Hasselt, Satinder Singh, and David Silver. Discovering reinforcement learning algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1060–1070, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0b96d81f0494fde5428c7aea243c9157-Abstract.html.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024. URL https://www.jmlr.org/papers/v25/23-0183.html.

Leonard S. Pleiss, James Harrison, and Maximilian Schiffer. Target-aligned reinforcement learning, 2026. URL https://arxiv.org/abs/2603.29501. arXiv preprint arXiv:2603.29501.

Aditya Ramesh, Kenny John Young, Louis Kirsch, and Jürgen Schmidhuber. Sequence compression speeds up credit assignment in reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42104–42126, 2024. URL https://proceedings.mlr.press/v235/ramesh24b.html.

Joshua Romoff, Peter Henderson, David Kanaa, Emmanuel Bengio, Ahmed Touati, Pierre-Luc Bacon, and Joelle Pineau. TDprop: Does adaptive optimization with Jacobi preconditioning help temporal difference learning? In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1082–1090, 2021. URL https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1082.pdf.

Braham Snyder, Amy Zhang, and Yuke Zhu. Target rate optimization: Avoiding iterative error exploitation. In *NeurIPS Workshop on Foundation Models for Decision Making*, 2023. URL https://neurips.cc/virtual/2023/82744. Workshop paper.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Yunhao Tang and Remi Munos. Towards a better understanding of representation dynamics under TD-learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33720–33738, 2023. URL https://proceedings.mlr.press/v202/tang23g.html.

Harm van Seijen and Richard S. Sutton. True online TD(λ). In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 692–700, 2014. URL https://proceedings.mlr.press/v32/seijen14.html.

Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Janarthanan Rajendran, Richard L. Lewis, Junhyuk Oh, Hado van Hasselt, David Silver, and Satinder Singh. Discovery of useful questions as auxiliary tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/10ff0b5e85e5b85cc3095d431d8c08b4-Abstract.html.

Théo Vincent, Yogesh Tripathi, Tim Faust, Abdullah Akgül, Yaniv Oren, Melih Kandemir, Jan Peters, and Carlo D'Eramo. Bridging the performance gap between target-free and target-based reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5ce377d14a21ef1fea0400049ad324b4-Abstract-Conference.html.

Martha White and Adam White. A greedy approach to adapting the trace parameter for temporal difference learning. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pp. 557–565, 2016. URL https://aamas.csc.liv.ac.uk/Proceedings/aamas2016/pdfs/p557.pdf.

Zechen Wu, Amy Greenwald, and Ronald Parr. A unifying view of linear function approximation in off-policy RL through matrix splitting and preconditioning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 26251–26328, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/25c152f350ef5f4f5878732ff2b57277-Abstract-Conference.html.

Zhongwen Xu, Hado van Hasselt, and David Silver. Meta-gradient reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2715518c875999308842e3455eda2fe3-Abstract.html.

Zhongwen Xu, Hado van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, and David Silver. Meta-gradient reinforcement learning with an objective discovered online. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15254–15264, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/ae3d525daf92cee0003a7f2d92c34ea3-Abstract.html.

Hengshuai Yao and Zhi-Qiang Liu. Preconditioned temporal difference learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1208–1215, 2008. doi: 10.1145/1390156.1390308. URL https://doi.org/10.1145/1390156.1390308.

Hengshuai Yao, Shalabh Bhatnagar, and Csaba Szepesvári. Temporal difference learning by direct preconditioning. In *Multidisciplinary Symposium on Reinforcement Learning*, 2009. URL https://repository.ias.ac.in/116709/. Workshop paper.

Tom Zahavy, Zhongwen Xu, Vivek Veeriah, Matteo Hessel, Junhyuk Oh, Hado van Hasselt, David Silver, and Satinder Singh. A self-tuning actor-critic algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20913–20924, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f02208a057804ee16ac72ff4d3cec53b-Abstract.html.

A Full Derivations

This appendix expands the mathematical statements in Section 4. It separates an exact local identity, an exact one-sided derivative, a conditional finite-response bound, and the operational prospective test. None of these objects is a learner modification.

A.1 Local synchronization forcing

For the two-action affine learner, the local equations below suppress the assigned-head subscript: $\theta_k, g_k \in \mathbb{R}^9$. For the action assigned to episode k, let $u_{k,t} = [\varphi_{k,t}; 1] \in \mathbb{R}^9$ denote the bias-augmented feature, let $\iota_t = \mathbf{1}\{t < H-1\}$, and define the accumulating trace

$$e_{k,-1} = 0, \qquad e_{k,t} = \gamma\lambda e_{k,t-1} + u_{k,t}, \qquad \gamma = .99. \tag{9}$$

Parameters remain fixed within the episode and are updated once afterward. The stop-target update vector for the assigned head is

$$g_k = -\frac{1}{H}\sum_{t=0}^{H-1} \delta_{k,t} e_{k,t}; \tag{10}$$

the unassigned head has zero direct update. Here $r_{k,t}$ is the reward on transition $t \to t+1$, so $r_{k,H-1}$ is the terminal reward denoted r_H in Section 5. Set $u_{k,H} = 0$ as a terminal-continuation convention. At a common current state and on the same trajectory, the two numerical residuals are

$$\begin{aligned} \delta_{k,t}^{\text{Current}} &= r_{k,t} + \gamma\iota_t u_{k,t+1}^\top \theta_k - u_{k,t}^\top \theta_k, \\ \delta_{k,t}^{\text{One-Lag}} &= r_{k,t} + \gamma\iota_t u_{k,t+1}^\top \theta_{k-1} - u_{k,t}^\top \theta_k. \end{aligned} \tag{11}$$

Both retain nonterminal continuation value, and neither differentiates through the continuation target. With $\Delta\theta_k = \theta_k - \theta_{k-1}$, the reward and current-state terms cancel exactly:

$$\delta_{k,t}^{\text{Current}} - \delta_{k,t}^{\text{One-Lag}} = \gamma\iota_t u_{k,t+1}^\top \Delta\theta_k. \tag{12}$$

Substitution into equation 10 yields

$$g_k^{\text{Current}} - g_k^{\text{One-Lag}} = -C_{k,\lambda,H}\Delta\theta_k, \qquad C_{k,\lambda,H} = \frac{\gamma}{H}\sum_{t=0}^{H-2} e_{k,t} u_{k,t+1}^\top. \tag{13}$$

Expanding the trace makes the cross-time geometry explicit:

$$C_{k,\lambda,H} = \frac{\gamma}{H}\sum_{t=0}^{H-2}\sum_{j=0}^{t} (\gamma\lambda)^{t-j} u_{k,j} u_{k,t+1}^\top. \tag{14}$$

For $\lambda = 0$, only adjacent-feature terms remain; for $\lambda > 0$, signed cross-time terms may align, cancel, or reverse. The block C has direct TD/LSTD operator-algebra antecedents and is not claimed as novel.

Equation 13 requires the same parameters, optimizer state, reward, action, horizon, feature sequence, trace, current-value evaluation, and trajectory. Once two training branches diverge, their later trajectories, clipping events, optimizer moments, residuals, and parameters can differ. The identity is therefore local, not an exact finite-training or terminal-outcome identity.

A.2 Complete learner-recursion derivative

For schedule S and temporal-credit configuration c, let $x_{S,k,c} = (\theta_{S,k,c}, \theta_{S,k-1,c}, m_{S,k,c}, v_{S,k,c})$ contain the parameters and AdamW moments needed for the next update. Here θ stacks all action heads. For N1, $\theta, m, v \in \mathbb{R}^{18}$ and $x, z \in \mathbb{R}^{72}$; A, Φ are 72×72 and J is 72×18. With the assigned-head selector $P_a : \mathbb{R}^{18} \to \mathbb{R}^9$, the complete coupling is $P_a^\top C_{k,\lambda,H} P_a$, written $C_{S,k,c}$ below; thus forcing is zero outside that head. Old moments and weight decay can still move the other head. The external three-head learner instead has 27 parameters. Initially $\theta_{-1} = \theta_0$ and the complete state is independent of η. The fixed analysis path uses

$$\theta^{\mathrm{tgt}}_{S,k,c}(\eta) = (1-\eta)\theta_{S,k,c}(\eta) + \eta\theta_{S,k-1,c}(\eta), \qquad \eta \in [0,1], \tag{15}$$

so $\eta = 0$ is Current and $\eta = 1$ is One-Lag. This scalar is neither trained nor swept experimentally. Writing the frozen schedule-specific update as $x_{S,k+1,c}(\eta) = F_{S,k,c}(x_{S,k,c}(\eta), \eta)$, suppose each update admits a joint linear first-order differential at its Current state on the admissible state domain. Explicitly, suppressing indices and setting $x = x_{S,k,c}(0)$, require

$$F(x+h,\eta) - F(x,0) = Ah + (Jb)\eta + o(\|h\| + |\eta|)$$

for admissible perturbations approaching $(h,\eta) = (0,0)$. Here J linearizes the next state with respect to the pre-clipping gradient, holding the incoming state fixed. Under this condition, define

$$\begin{aligned} z_{S,k,c} &= \left.\frac{\mathrm{d}x_{S,k,c}(\eta)}{\mathrm{d}\eta}\right|_{0^+}, \\ A_{S,k,c} &= D_x F_{S,k,c}(x_{S,k,c}(0), 0), \\ J_{S,k,c} b_{S,k,c} &= \partial_\eta F_{S,k,c}(x_{S,k,c}(0), 0), \end{aligned} \tag{16}$$

where $b_{S,k,c} = C_{S,k,c}\Delta\theta_{S,k,c}$. The chain rule gives

$$z_{S,0,c} = 0, \qquad z_{S,k+1,c} = A_{S,k,c} z_{S,k,c} + J_{S,k,c} b_{S,k,c}. \tag{17}$$

For $\Phi_{S,K\leftarrow k+1,c} = A_{S,K-1,c} \cdots A_{S,k+1,c}$, with the empty product equal to the identity,

$$z_{S,K,c} = \sum_{k=0}^{K-1} \Phi_{S,K\leftarrow k+1,c} J_{S,k,c} b_{S,k,c}. \tag{18}$$

Under the joint linear-differential condition, this is the exact first derivative at $\eta \to 0^+$. At a clipping kink, the implementation combines inherited sensitivity and direct forcing before selecting the one-sided branch; separate directional responses need not add. The matrix statements do not assert that superposition at arbitrary kinks.

The frozen N1 and external audits recorded positive clipping-boundary distances and positive Adam second moments on influenced coordinates. These support the evaluated directional recursions; they do not certify differentiability on the full ambient state space, including inactive zero-moment coordinates, or control the finite branch difference at $\eta = 1$.

A.3 Function-space tangent, CSTE, and the finite bound

Use the real Hilbert space $L^2(\nu)$ on a declared probability measure, with $\langle f, h\rangle_\nu = \int f(\chi)h(\chi)\,\mathrm{d}\nu(\chi)$ and its induced norm. For external centered action scores, use the same definition with the mean of the three actionwise products. The affine readouts have finite feature Gram

matrices on the frozen fixtures, hence are bounded linear maps from parameter displacements into this space; positive definiteness is not required. For action-margin function $M_c(x;\chi)$, define

$$d_{S,c}(\eta;\chi) = M_c(x_{S,K,c}(0);\chi) - M_c(x_{S,K,c}(\eta);\chi), \qquad t_{S,c}(\chi) = -D_x M_c(x_{S,K,c}(0);\chi) z_{S,K,c}. \tag{19}$$

Let $Q_{S,c} = \|t_{S,c}\|_\nu$. For a prespecified ordered pair (c_1, c_2), CSTE is

$$\Omega_{S;c_1,c_2} = Q^2_{S,c_1} - Q^2_{S,c_2}. \tag{20}$$

It compares complete-recursion sensitivities; it is not an objective or a claim that either configuration produces a better policy.

At the finite endpoint, write $d_{S,c}(1) = t_{S,c} + \epsilon_{S,c}$ and assume $\|\epsilon_{S,c}\|_\nu \le \rho_{S,c}$. Cauchy–Schwarz gives

$$\begin{aligned} \left| \|d_{S,c}(1)\|^2_\nu - Q^2_{S,c} \right| &= \left| 2\langle t_{S,c}, \epsilon_{S,c}\rangle_\nu + \|\epsilon_{S,c}\|^2_\nu \right| \\ &\le 2Q_{S,c}\rho_{S,c} + \rho^2_{S,c} =: E_{S,c}. \end{aligned} \tag{21}$$

With $Y_{S,c} = \|d_{S,c}(1)\|_\nu$ and $R_{S;c_1,c_2} = Y^2_{S,c_1} - Y^2_{S,c_2}$, the triangle inequality implies

$$|R_{S;c_1,c_2} - \Omega_{S;c_1,c_2}| \le E_{S,c_1} + E_{S,c_2}. \tag{22}$$

Consequently,

$$|\Omega_{S;c_1,c_2}| > E_{S,c_1} + E_{S,c_2} \quad \Longrightarrow \quad \operatorname{sign} R_{S;c_1,c_2} = \operatorname{sign} \Omega_{S;c_1,c_2}. \tag{23}$$

This condition is sufficient, not necessary. It requires the same function-space measure and finite-remainder dominance. The corresponding schedule-twin statement follows without changing the bound. Let $D_S = E_{S,c_1} + E_{S,c_2}$. For independent schedules A and B,

$$|\Omega_{A;c_1,c_2} - \Omega_{B;c_1,c_2}| > D_A + D_B \quad \Longrightarrow \quad \operatorname{sign}(R_{A;c_1,c_2} - R_{B;c_1,c_2}) = \operatorname{sign}(\Omega_{A;c_1,c_2} - \Omega_{B;c_1,c_2}). \tag{24}$$

The bounded affine readout carries the state derivative into $d_{S,c}(\eta)/\eta \to t_{S,c}$ in $L^2(\nu)$. Norm continuity therefore gives

$$\lim_{\eta \to 0^+} \frac{\|d_{S,c}(\eta)\|^2_\nu}{\eta^2} = Q^2_{S,c}. \tag{25}$$

Any nonzero Ω thus fixes the contrast sign for sufficiently small positive η. Extending either sign statement to the finite $\eta = 1$ endpoint requires the stated remainder bound; it is not automatic. The operational N1 and external tests instead froze predictors on ν_{ref} and observed responses on a disjoint ν_{eval}; they test a separately frozen ordering claim and do not empirically verify equation 23.

Table 4: Mathematical status of the synchronization statements.

Object	Status	Does not establish
Equation 13	Exact at a common state and trajectory	Finite branch or terminal difference
Equation 18	Exact first derivative under joint linear differentiability	Exact $\eta = 1$ difference
Equation 23	Sufficient finite-response condition	Necessity or empirical remainder dominance
Equation 24	Sufficient schedule-ordering corollary on one common measure	Direct empirical instantiation by disjoint fixtures
N1/external ordering	Frozen prospective empirical test	Absolute magnitude, calibration, utility, or mediation

B Information-Equivalence Construction

R0 is the exact eight-dimensional BCTS Tetris afterstate vector: landing height, eroded piece cells, row transitions, column transitions, holes, cumulative wells, hole depth, and rows containing holes.

R3 coarsens each coordinate to the lower edge of its width-32 bin, $\phi_{R3} = 32\lfloor \phi_{R0}/32 \rfloor$, componentwise in raw feature units. R0 and R3 are different information regimes; the information-equivalence claim applies to RAW versus WHITE within each regime.

Information equivalence is an intervention property, not a claim that RAW and WHITE follow the same optimization path. In tested Tetris, WHITE used a training-only regularized invertible ZCA map. The eight-dimensional transform preserved sampled collision classes, the audited affine function class, and the initial action-value function. Table 5 records the most decision-relevant invariant checks.

Table 5: Frozen tested-Tetris information-equivalence audit. R3's empirical covariance is rank five, while its regularized transform is rank eight.

Audit quantity	R0	R3
Representation dimension	8	8
Empirical covariance rank	8	5
Transform rank	8	8
Maximum inverse error	4.55×10^{-13}	1.99×10^{-13}
Unique RAW/WHITE rows	12,331,250 / 12,331,250	385 / 385
Collision counts RAW/WHITE	19,525,772 / 19,525,772	31,856,637 / 31,856,637
Maximum affine-score error	5.33×10^{-15}	8.88×10^{-16}
Initial-Q maximum error	0	0

The R3 distinction is essential. Three coordinates are constant on the training population, so neither RAW nor transformed empirical covariance is full rank. Regularization retains and finitely scales those coordinates rather than dropping them; the transformation itself remains an invertible rank-eight map. The equivalence claim follows from invertibility, unchanged dimensionality, collision preservation, affine-function mapping, and identical initial functions—not from a false full-rank covariance claim.

The BASE/AdamW, N1, and external studies each performed their own training-only equivalence audit. The controlled studies used four canonical full-rank transforms indexed by information and horizon and shared them across schedules and synchronization branches. N1 independently bound its transforms before the response stage. The external domain fitted one full-rank ZCA transform on 65,536 states from an independent training-only behavior population and shared it across 48 schedules, temporal levels, and synchronization branches. Its RAW and WHITE representations were two invertible eight-dimensional coordinate forms of the same features and affine action-value class.

Across these constructions, information, dimensionality, and the audited affine capacity are invariant. Coordinate scale, correlation, conditioning, native Euclidean clipping, coordinatewise AdamW moments, and the realized optimizer path are not. Thus information equivalence does not imply training-path or efficacy invariance, and no audit supplies a universal RAW-versus-WHITE direction.

Table 6: Meaning of information equivalence in the frozen interventions.

Audited as preserved	Not asserted to be preserved
Information and dimensionality; invertible coordinate map; audited affine action-value class; paired initial function; collision classes where applicable	Coordinate scale and correlation; covariance conditioning; Euclidean clipping geometry; coordinatewise optimizer moments; stochastic optimizer path; learned parameters, efficacy, or effect direction

C Experimental Protocols

The studies use different platforms and outcomes; Table 7 is their separation map, not a pooled design. Every uncertainty statement below is the original frozen family, and every random unit is the unit generated by the study design.

Table 7: Cross-study separation map. The studies use different platforms, outcomes, estimands, and inferential units; the rows are not a pooled design or a mediation chain. GR denotes terminal GAME RETURN, M_H the controlled high-order interaction, and LF a learned-function response.

Study	Setting and outcome	Inferential unit	Frozen interval family and role
H2	Tetris; R0/R3, RAW/WHITE, TD(0)/TD(λ); terminal GR	Common paired protected block (512; $n = 512$/arm)	Arm means: descriptive 95%; three adjusted 98.333% interaction intervals; bounded geometry causality
M0b	BASE/AdamW; information, horizon, geometry, and temporal credit; schedule distribution of M_H	Schedule ($n = 16$) $\times$ shared block (4096)	Frozen 95% SD interval; training-realization materiality
M1	Same controlled platform plus CURRENT/ONE-LAG; mean and log-SD contrasts of M_H	Schedule ($n = 48$) $\times$ shared block (4096)	Two co-primary 97.5% intervals; bounded synchronization component
M2-V	BASE/AdamW; $H = 4/32$, $\lambda = 0/.9$; LF absolute and contrast tests	Complete schedule ($n = 48$)	Five adjusted 99% intervals; unresolved absolute-to-contrastive pivot
N1	Fresh schedule twins; $H = 6/18$, $\lambda = 0/.6$; LF contrast ordering	Complete twin ($n = 48$)	Three adjusted 98.333% intervals; primary prospective test
External	Fixed-horizon LQR; policy endpoint and separate LF ordering	Complete schedule ($n = 48$)	Core: three adjusted 98.333%; CSTE: separate 99%; transport and falsification

The paper-facing canonical records for these rows are, respectively, `PHASE_H2_PROSPECTIVE_LONG_GAME_ENDPOINT_REPORT`, `PHASE_M0B_TRAINING_SCHEDULE_TRANSPORT_PRECISION_REPORT`, `PHASE_M1_SCHEDULE_BLOCKED_BOOTSTRAP_PATH_CAUSAL_REPORT`, `PHASE_M2V_DETAILED_PROSPECTIVE_THEORY_VALIDATION_REPORT`, `PHASE_N1_FRESH_CONTRASTIVE_THEORY_VALIDATION_REPORT`, and `EXTERNAL_CAUSAL_REPLICATION_REPORT`. Predictor-freeze and independent-validation records remain separate from these terminal reports; the terminal report names do not replace their chronology evidence.

C.1 H2 TERMINAL INTERVENTION

H2 reused all 40 Phase-H-trained members—five members in each of eight R0/R3 $\times$ RAW/WHITE $\times$ TD0/TD($\lambda = .9$) arms—without retraining, refitting whitening, or selecting models from protected outcomes. It generated a new 512-block population with zero reset-key or seed overlap with the compromised Phase H campaign. Exact resumable state snapshots allowed a live game to cross execution boundaries; scientific termination remained genuine environment game over. All $4{,}608$ required arm/control episodes terminated, and a return-blind completeness gate preceded unblinding. All arms were evaluated on the same fresh blocks. The three interaction estimands used the frozen Bonferroni-three two-sided 98.333% family; arm intervals are descriptive 95% intervals.

Phase H itself is not a null precursor. One of 512 paired blocks lacked a genuine terminal observation at its frozen execution cap, so no complete factorial estimand was formed and all returns remained sealed. H2 was a new prospective endpoint study, not a completion or pooling of Phase H.

C.2 M0B AND M1 CONTROLLED SCHEDULES

BASE draws a latent cue $z \sim \mathrm{Unif}[-1, 1]$. An initial binary action a selects a deterministic path of $H \in \{4, 32\}$ transitions. All intermediate rewards are zero; the terminal reward is $\mathbf{1}\{a = \mathbf{1}[z > 0]\}$.

Thus mean terminal return is the probability of a correct initial choice, on a 0–1 scale. HIGH observes z; LOW observes its width-0.5 bin midpoint with origin -1.125, clipped to $[-0.875, 1.125]$. At evaluation, the five members' action scores are averaged before choosing the maximizing action, with ties assigned to action 0. This terminal return is the base outcome for the contrasts.

M_H averages, over HIGH/LOW information regimes, the LONG-minus-SHORT contrast of the WHITE-minus-RAW change in TD($\lambda = 0.9$)-minus-TD(0).

M0b prospectively generated 16 fresh complete schedules. Within a schedule it crossed information, horizon, geometry, and temporal credit, with five members per cell and a common protected evaluation population. The primary object was the between-schedule SD of the prespecified geometry $\times$ temporal-credit $\times$ horizon interaction. Increasing protected blocks or learners within a schedule could not replace independent schedules.

M1 generated 48 new complete schedules and crossed the same controlled factorial with CURRENT versus ONE-LAG. Both branches used $\gamma = .99$, true-terminal zero continuation, TD(0)/TD($\lambda = .9$), horizons 4/32, the same affine class, AdamW/clipping semantics, training budget, and zero-function initialization. The sole intended intervention changed the continuation-target parameter version from θ_k to θ_{k-1}. The paired schedule mean shift and log-SD contrast were co-primary members of a familywise 95% procedure. This design identifies the bundled one-update synchronization intervention, not separate self-reference, staleness, clipping, or optimizer mediators.

Both studies estimated schedule variability from the balanced schedule $\times$ common-block design, with 4,096 protected blocks shared across cells. Writing MS_S and $MS_{S\times B}$ for the schedule and interaction mean squares, the raw schedule-variance component was $(MS_S - MS_{S\times B})/4096$. Raw components were reported; negative components were set to zero for SDs and decisions. Thus $\widehat{\sigma}_S = \sqrt{\max\{0, (MS_S - MS_{S\times B})/4096\}}$ is a variance-component estimate, rather than the sample SD of schedule means. M1 used $\log[(\widehat{\sigma}_{\text{CURRENT}} + 10^{-12})/(\widehat{\sigma}_{\text{ONE-LAG}} + 10^{-12})]$, also reporting the unregularized SDs. Each study used 5,000 crossed bootstrap replicates, resampling complete schedule slabs and common block columns while preserving shared cells and, for M1, branch pairing. M0b's primary interval was 95%; M1 used two Bonferroni 97.5% co-primary intervals and a separate family of ten 99.5% secondary intervals. These percentile intervals used linear quantile interpolation. The 16/48 schedules remain the independent training-realization counts; the block axis does not add training replications.

C.3 M2-V, N0, AND N1 CHRONOLOGY

M2-V prospectively froze its CURRENT-only local predictors before the held-out ONE-LAG responses. It used 48 schedules, $H = 4/32$, $\lambda = 0/.9$, 20,000 whole-schedule percentile-bootstrap replicates, and five Bonferroni two-sided 99% intervals. It evaluated no protected GAME RETURN. N0 subsequently derived CSTE theory-only and froze a new N1 schedule-twin prediction; it did not fit CSTE to M2-V responses.

N1 then generated 48 fresh twin blocks (96 schedules), with 16 registered temporal rows per twin. Stage 1 completed 1,536 CURRENT learners and tangents and froze all CSTE predictors while every corresponding ONE-LAG learner, response, and inference count was zero. Stage 2 completed 1,536 ONE-LAG learners. Its three estimands used 20,000 percentile-bootstrap replicates of whole twin blocks and Bonferroni-adjusted 98.333% intervals. The 768 nested row comparisons construct 48 twin statistics; they are not independent replications.

C.4 EXTERNAL FIXED-HORIZON DOMAIN

The external state lies in $[-2, 2]^2$, the action set is $\{-1, 0, +1\}$, and every policy terminates after exactly 128 transitions. The frozen componentwise-clipped dynamics are

$$\begin{aligned} x_{t+1} &= \text{clip}(Ax_t + Ba_t + w_t, -2, 2), \\ A &= \begin{bmatrix} .92 & .12 \\ -.08 & .90 \end{bmatrix}, \qquad B = \begin{bmatrix} .18 \\ .10 \end{bmatrix}, \\ w_t &\sim \text{Unif}([-.025, .025]^2), \end{aligned} \tag{26}$$

with $x_0 \sim \mathrm{Unif}([-1,1]^2)$. Each schedule froze the ordered initial states, process noise, and uniformly sampled behavior actions, and all eight cells in that schedule used the same realized transitions. Final greedy policies used the fixed tie order $(-1, 0, +1)$. With $c_t = x_{1,t}^2 + x_{2,t}^2 + 0.05a_t^2$, training reward is $-c_t/8.05$ and the continuous policy endpoint is

$$G = 1 - \frac{1}{128 \cdot 8.05} \sum_{t=0}^{127} c_t \in [0, 1]. \tag{27}$$

All eight geometry × temporal-credit × synchronization cells were paired within each of 48 complete schedules. The cell endpoint $G_{S,g,c,B}$ is the arithmetic mean of G over the 2,048 common evaluation episodes for that policy. Whole schedules were resampled with 20,000 PCG64 percentile-bootstrap replicates; the three empirical-core intervals and the separate learned- function CSTE interval retain their original families described in Appendix H.

D Origin and C06 Contraction

The research origin supplies a governance lesson rather than the final claim. The first independent terminal study passed protocol and completion checks but closed with the funnel

$$8 \text{ registered candidates} \to 2 \text{ exploratory finalists} \to 0 \text{ independently confirmed finalists.}$$

This does not show that every return-gap method is ineffective. It shows that local, ranking, selected-return, or exploratory evidence cannot substitute for an untouched terminal confirmation endpoint. Earlier SFT observations remain provenance-only because they used imitation/candidate diagnostics rather than closed-loop terminal Game Return and do not establish a common mechanism with the present RL result.

A later independently protected study confirmed C06 against B00: the paired terminal difference was 2.7832 and its one-sided 97.5% lower bound 2.2582 exceeded the frozen $+1.0$ SESOI. C04 did not cross the same rule. Efficacy and novelty were then separated. In a fresh matched audit, the C06 learner did not establish superiority over the matched uncorrected finite-horizon TD(λ)-style arm, and the formula audit classified C06 as a finite-horizon fitted Exact Retrace specialization. A large matched representation comparison additionally exposed sensitivity to representation choice. These results motivated the geometry–temporal-credit question; they do not support a C06 algorithm claim, universal multi-step benefit, or a representation novelty claim.

The paper-facing contraction is therefore:

terminal confirmation is necessary
$\longrightarrow$ C06 efficacy is bounded but its algorithm-novelty route closes
$\longrightarrow$ representation-conditioned temporal credit becomes the scientific problem.

No later CSTE evidence retroactively changes these classifications.

E Phase E–L Evidence Packages

The intermediate studies distinguish candidate explanations and their tested boundaries. Table 8 reports their dominant role; none is promoted to the primary CSTE evidence chain.

Phase I directly intervened on the coordinate norm used by global clipping. Both the native and canonical-clipping R0 interactions remained materially supported. Their direct difference was 165.93 with frozen 98.333% CI $[-154.84, 486.69]$, which crossed zero and did not reach the frozen strong- attenuation reference 990.22. Figure S1 therefore rules down this particular clipping account; it neither proves invariant effects nor identifies the remaining optimizer-path mediator.

The later J–L sequence supplies a separate transport assessment. J's balanced horizon moderator was 0.04102 with Bonferroni-adjusted 98.333% CI $[0.02258, 0.05945]$; information and reward-scale moderators were unresolved. K stopped at a return-blind completeness barrier with 96 genuine-terminal and 16 ceiling-nonterminal receipts, zero complete paired blocks, and no protected effect estimate. Within L, AdamW and SGD horizon moderators had opposite signs, and their contrast was

Table 8: Supporting mechanism and transport evidence. "Incomplete" is a protocol classification, not a null effect.

Phase	Evidence and frozen status	Boundary
E	Representation-regime moderation supported	Context/motivation, not information-equivalent geometry causality
F0	Coarsening ladder supported; unique mediator not identified	Information, geometry, scale, and control resolution co-vary
G	Fixed-credit burden supported; benefit of adaptive λ not established	No universal target-burden or adaptive-credit mechanism
H	Protected terminal geometry endpoint; causal identification compromised	Missing terminal cell; no effect, sign, or null estimate
I	Common-raw-norm clipping attenuation; strong attenuation not established	Tested clipping norm is not the sole/main mediator
J	Information/horizon/reward moderators; horizon supported, others unresolved	One controlled realization; no universal LONG direction
K	Non-Tetris goal-reaching endpoint incomplete/nonanalyzable	Zero complete paired blocks; no protected contrast
L	AdamW/SGD surface; within-L optimizer contrast supported, J anchor not reproduced	Does not establish Adam mediation of J

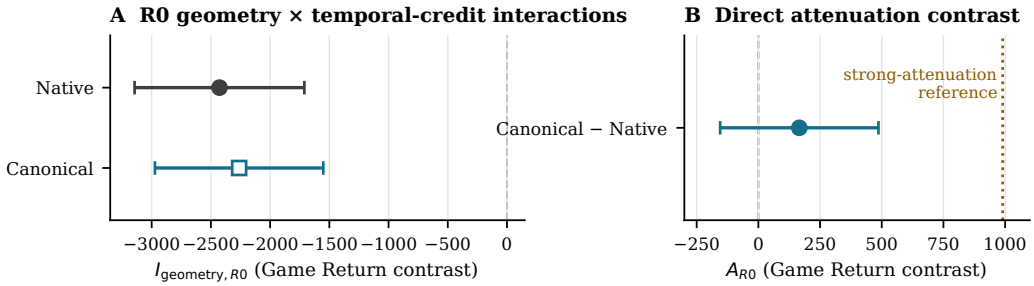

Figure S1: Phase I coordinate-clipping boundary. Panel A reproduces the frozen R0 geometry × temporal-credit interactions under native and canonical clipping with their 98.75% intervals; both are materially supported. Panel B shows the frozen canonical-minus-native attenuation contrast and adjusted 98.333% interval together with the preregistered strong-attenuation reference. Labels give nominal coverage per interval: the four-member interaction and three-member attenuation families each target 95% familywise coverage. The attenuation contrast is NOT ESTABLISHED; this is not an exact-zero or optimizer-invariance result. Panels use separate Game Return contrast scales. All values are direct transcriptions from the frozen Phase I statistics; no interval was recomputed.

precise, but the prospectively required positive AdamW anchor did not reproduce. Hence optimizer assignment changed the L surface, while the proposed Adam mediation of J remained unestablished. This sequence motivated prospective realization-level blocking; it did not itself establish M0b schedule materiality.

F M0 and M2-V Auxiliary Analyses

F.1 M0 precision screen and M0b resolution

M0 decomposed transport variability using five independent schedules and five whitening fits. Its schedule-SD estimate was 0.08009 with frozen 95% CI $[0.01853, 0.10836]$; because the lower bound did not exceed the prespecified 0.02 threshold, material schedule variability remained UNRESOLVED. The whitening and schedule–whitening components were negligible in that tested setting, but five schedules could not establish the schedule claim. Additional learners or protected blocks within those schedules would not change the independent realization count.

M0b resolved that precision question with 16 fresh schedules and shared frozen training-only transforms. The between-schedule SD of the registered high-order interaction was 0.13423, 95% CI $[0.08715, 0.16693]$, whose lower bound exceeded 0.02 (SUPPORTED). The 16 schedule values ranged from -0.32495 to 0.13354 with both signs represented. This establishes material heterogeneity under the prospective schedule generator, not a universal negative mean or a schedule mediator.

F.2 M2-V absolute and auxiliary predictions (Appendix S7)

M2-V froze five CURRENT-side predictions before generating its ONE-LAG learned-function responses. Table 9 reproduces the original five Bonferroni-adjusted intervals (nominal 99% each; 95% familywise target).

Table 9: Frozen M2-V prediction family. P1 remains unresolved under its own criterion; auxiliary P2 results predate the N0 CSTE definition.

Prediction	Estimate	CI low	CI high	Status
P1 absolute	0.08449	-0.02378	0.19566	Unresolved
P2 horizon	0.98958	0.97135	1.00000	Supported
P2 λ	0.66406	0.60156	0.72917	Supported
P3 collision	-0.17564	-0.49229	0.19363	Unresolved
P4 covariance defect	0.02970	-0.08583	0.14750	Unresolved

P2-H had positive predictor signs in all 384 rows, so it supplied limited anti-triviality against a fixed horizon direction. P2-λ had 151 positive and 233 negative predictor signs and was more discriminative. Neither proves a mechanism, establishes absolute prediction, or implies a universal horizon/λ direction. P3 did not establish a collision-conditioned information explanation. P4's algebraic covariance control passed, while its empirical defect-to-response bridge remained unresolved. N0 subsequently defined a new contrastive full-recursion object and did not test it retrospectively on M2-V outcomes.

G N1 Chronology and Complete Schedule-Twin Material

G.1 Task and training configuration

The N0 prospective design fixed the following N1 configuration before any N1 response existed. We write the latent cue as ξ to distinguish it from the learner-state tangent $z_{S,k,c}$.

Delayed tracking task. For each episode, $\xi \sim \mathrm{Uniform}[-1, 1]$, $q_0 = 0$, and the assigned action $a \in \{0, 1\}$ remains fixed. The dynamics are $q_{t+1} = .75q_t + .25(2a - 1)$, with termination after exactly $H = 6$ or 18 transitions. Intermediate rewards are zero; the terminal reward is $1 - .25(q_H - \xi)^2 \in [0, 1]$. HIGH observes $o = \xi$; LOW observes the midpoint of its bin under the five-bin quantizer on $[-1, 1]$ (bin width .4). Actions are supplied by the training schedule rather than selected by a learned actor.

Features and geometry. For $t = 0, \ldots, H-1$, set $\tau_t = 2t/(H-1) - 1$. The unscaled features are

$$[o,\ q_t,\ o^2 - 1/3,\ q_t^2,\ \tau_t,\ \tau_t^2 - 1/3,\ oq_t,\ q_t\tau_t].$$

RAW applies coordinate scales $[10^{-2}, 10, 10^2, 10^{-1}, 5\times10^{-2}, 5, 2, 2\times10^{-1}]$. An affine bias and two action heads give 18 parameters. WHITE uses four training-only full-rank ZCA transforms, one for each information–horizon regime. Each is fitted on 32,768 independent feature rows, separate from the schedules and function fixtures, and shared across schedules, trace levels, and synchronization branches in its regime. Pretraining audits required rank 8, inverse and affine-score errors at most 10^{-10}, collision equivalence, and identical zero initial functions. The transforms retain all eight coordinates. Information equivalence concerns RAW/WHITE within a regime, not HIGH versus LOW.

Learner and update. All cells start from a zero action-value function. They use $\gamma = .99$, accumulating traces with $\lambda = 0$ or .6, and episode-frozen stop-target semigradients. Parameters are fixed while the episode gradient is accumulated; it is divided by H and clipped in the native 18-coordinate norm at 1 before one AdamW update. The AdamW constants are learning rate .0003, $\beta_1 = .9$, $\beta_2 = .999$, $\epsilon = 10^{-8}$, and weight decay 10^{-5}. CURRENT uses current continuation-target parameters; ONE-LAG uses the parameters from the immediately preceding optimizer update. Both retain next-state bootstrapping and zero continuation at terminal transitions. The precise update and tangent equations are in Appendix A.

Schedules and evaluation. Forty-eight independently generated twin blocks each contain two independent schedules. Each schedule has 12,000 fresh cues and 6,000 assignments to each action in shuffled order. Its 32 factor cells share that ordered cue/action sequence, yielding 3,072 learners and 12,000 optimizer updates per learner. The protocol fixes the budget without early stopping or schedule selection. A 16,384-point midpoint grid on $[-1, 1]$ supplies even-indexed cues to ν_{ref} and odd-indexed cues to ν_{eval}: 8,192 points each, with zero latent overlap. Both evaluate the two action heads only at the common initial state $q_0 = 0$, $\tau_0 = -1$. The action margin is $Q(s,1) - Q(s,0)$. Predictor energies use ν_{ref}; finite responses use ν_{eval}. The original three-estimand inference uses PCG64 seed `0xfb5557079ece4925` for 20,000 whole-twin percentile-bootstrap replicates. The following chronology and response definitions specify how those settings produce the reported ordering statistics.

G.2 PREDICTOR FREEZE AND INFERENCE

N1's scientific protection is temporal: the complete predictor existed in immutable response-blind form before the corresponding finite responses. The dependency sequence was

1. N0 froze CSTE, the ordered contrasts, the schedule-twin estimands, the tie rule, and whole-twin inference before N1 execution.
2. N1 froze 48 twins/96 schedules, completed 1,536 CURRENT learners and full-recursion tangents, and published 1,536 CSTE contrast rows.
3. The predictor table was hash-frozen, and an independently checked barrier recorded zero ONE-LAG learners, final parameters, ν_{eval} responses, Y, R, or inference artifacts.
4. A separate Stage-2 authority completed 1,536 ONE-LAG learners, materialized the disjoint response fixture, froze the Y/R table, and only then executed the registered inference.

The raw Stage-1 predictor SHA-256 is

```
a34077d56e8130aabd6959b6348b9dc2
357404eba194d4265c44021cdbaa1f5a
```

and independent Stage-2 validation reproduced all results with recorded maximum error zero.

The predictor and response fixtures were disjoint 8,192-point sets. For a schedule S and temporal configuration c, Stage 2 used

$$Y_{S,c} = \mathrm{RMS}_{\nu_{\mathrm{eval}}}[M_{S,c}^{\mathrm{CURRENT}} - M_{S,c}^{\mathrm{ONE\text{-}LAG}}], \qquad R_{S;c_1,c_2} = Y_{S,c_1}^2 - Y_{S,c_2}^2. \tag{28}$$

For each twin i and 16 registered rows j, the primary schedule comparison used

$$D_i = \frac{1}{16}\sum_{j=1}^{16} \kappa(\Omega_{A_ij} - \Omega_{B_ij}, R_{A_ij} - R_{B_ij}), \quad \kappa(a,b) = \begin{cases} 1, & ab > 0, \\ .5, & ab = 0, \\ 0, & ab < 0. \end{cases} \tag{29}$$

The 16 rows comprise eight horizon and eight lambda contrasts across frozen information and geometry blocks. They contribute to one twin statistic rather than 16 independent observations. The primary, horizon, and lambda estimands used 20,000 percentile-bootstrap replicates of the 48 whole twins and three Bonferroni two-sided 98.333% intervals.

The primary estimate was 0.81120 with CI $[0.77474, 0.84505]$; horizon was 0.89714 with $[0.86849, 0.92448]$; lambda was 0.88411 with $[0.84766, 0.91667]$. All three were SUPPORTED because their lower bounds exceeded 0.5. These N1 Horizon/Lambda estimands differ from M2-V P2-H/P2-λ in settings, formula, and evidentiary role. The result is learned-function response ordering, not absolute response calibration, GAME RETURN, policy quality, or individual-twin classification.

H EXTERNAL GATE RULES

For schedule S, geometry g, temporal configuration c, and synchronization branch B, let $G_{S,g,c,B}$ be the fixed-horizon policy endpoint in equation 27. With TDL denoting TD($\lambda = .9$), define

$$\begin{aligned} T_{S,g,B} &= G_{S,g,\mathrm{TDL},B} - G_{S,g,\mathrm{TD0},B}, \\ I_{S,B} &= T_{S,\mathrm{WHITE},B} - T_{S,\mathrm{RAW},B}, \\ \mu_{I,\textsc{Current}} &= \mathrm{mean}_S\, I_{S,\textsc{Current}}, \quad \sigma_{I,\textsc{Current}} = \mathrm{SD}_S\, I_{S,\textsc{Current}}, \\ \Delta\mu_{\mathrm{sync}} &= \mathrm{mean}_S(I_{S,\textsc{Current}} - I_{S,\textsc{One-Lag}}). \end{aligned} \tag{30}$$

The empirical core required all three gates in main-text Table 3. It used 20,000 PCG64 percentile-bootstrap replicates of whole schedules and a Bonferroni family of three two-sided 98.333% intervals. The estimands have different scales and were never pooled.

Execution was complete and valid. The third required material endpoint claim was classified as falsified under the registered rule, fixing the conjunction as `EXTERNAL-C_EMPIRICAL_CORE_NOT_REPLICATED`. The descriptive count of two supported gates is not a partial-pass score. The interval excluding zero does not conflict with falsification: the claim was a material shift beyond 0.01, and the whole interval lies inside the registered practical-equivalence region.

The external CSTE test has its own chronology, outcome, and family. Stage 1 used 192 CURRENT paths to compute 192 tangent energies and 96 temporal contrasts on ν_{ref}. Its predictor raw SHA-256 was

```
65608fcff81ffe6ab04a191bd89e64e4
ed16a7bea9a0b3fc53351e65c9968562
```

when every ONE-LAG learner, response, and inference count was zero. Stage 2 consumed the table read-only and did not retrain CURRENT or recompute Ω.

For action order $(-1, 0, +1)$, the external metric centers action scores with $P = I_3 - \frac{1}{3}\mathbf{1}\mathbf{1}^\top$ and assigns equal weight to all actions and fixture points. The response $R_{S,g;\mathrm{TDL},\mathrm{TD0}}$ is a centered learned-function contrast, not the policy endpoint G. The paired schedule/geometry concordance estimate was 0.9375 with its separate 99% whole-schedule CI $[0.87495, 0.98958]$ (SUPPORTED). Thus the learned-function ordering and endpoint falsification are jointly coherent. CSTE support cannot alter any core gate, predict endpoint return, or establish full external generalization.

I GOVERNANCE, PROVENANCE, AND PRIOR-ART RELATIONSHIPS

The evidence chain follows a fixed authority order: Unit 10 final claim and language locks; Master claim–evidence, evidence-ledger, story, and figure-data records; then the canonical phase-level report, frozen machine statistic, independent validation, and terminal authority for each study. Display

CSVs are field projections for figures and tables, not alternative sources from which new statistics may be computed. Every negative, unresolved, falsified, or incomplete classification remains visible under this order.

Protected and prospective barriers have distinct functions. H2 used a return-blind completeness gate before terminal unblinding. M2-V, N1, and the external CSTE test froze response-blind predictor tables before corresponding ONE-LAG responses. M0b/M1 inference crossed complete schedules with common evaluation blocks (Appendix C); N1 resampled whole twins, and external inference resampled whole schedules. Nested learners and factorial cells did not supply additional independent training replications. Independent validations reconstructed the frozen quantities; they did not authorize alternative estimands.

Table 10 gives compact canonical ledger bindings for the artifacts carrying the paper-facing results. The hashes are transcribed from the Master evidence ledger; each row distinguishes a file hash from a predictor-hash role and defines neither an additional analysis nor a precedence rule.

Table 10: Canonical ledger artifact/hash bindings used by the manuscript evidence chain.

Evidence role	Ledger artifact / hash role	SHA-256
H2	FACTORIAL_CAUSAL_STATISTICS.json	cc5b0f0556541fa336cd161c4f9c67b67e4a3df7037b1fc807c361fc26b9e879
M0b	SCHEDULE_TRANSPORT_STATISTICS.json	6069eafb5c8f3cbef324ef31384b3ef7af527e966c247f4ef9b79aa2b0def9ab
M1	SCHEDULE_DISTRIBUTIONAL_STATISTICS.json	d2fb40c6d488a2cb031f4075524a7e3b4c6de6c10357fd6b6e3ff9c3d74afddb
M2-V	M2V_THEORY_VALIDATION_STATISTICS.json	f1c6865362bd2ed1c8d3ab15b0e3e44f15a29cb0e44e66a2c7b5d3209e5a908d
N0	N0_CSTE_THEORY_FREEZE.md	f7e23415696502828fcffa53ce204dad90fb7d0c8e14174da793f96401d36547
N1 predictor	N1_FROZEN_THREE_ESTIMAND_INFERENCE.json predictor SHA-256	a34077d56e8130aabd6959b6348b9dc2357404eba194d4265c44021cdbaa1f5a
N1 inference	N1_FROZEN_THREE_ESTIMAND_INFERENCE.json	d17cf8863c7aa91475c0a948a110d7c6ae796020eaad1fa4e29048f237c9f3b7
External core	EXTERNAL_FROZEN_INFERENCE.json	8c921c54eeef47b0b13bb40683a939247e914751039787afcb613edb450d744a
External predictor	EXTERNAL_FROZEN_INFERENCE.json external predictor SHA-256	65608fcff81ffe6ab04a191bd89e64e4ed16a7bea9a0b3fc53351e65c9968562

For anonymous review, author names and identifying repository paths are withheld. Ledger artifact bindings and hashes remain visible so that the frozen evidence relationships can be audited without revealing author identity; access details can be supplied only through the venue-permitted anonymous artifact channel.

Two nonblocking record conflicts remain preserved rather than silently resolved. Phase I has a dual-source identity whose canonical machine source controls the displayed statistics. N0 contains an analytical-authorization registry/projection discrepancy. Neither conflict was used to change a scientific result, predictor, permission, or claim.

The prior-art boundary is likewise hierarchical. Representation geometry, whitening/preconditioning, TD(λ), multi-step returns, target networks, target synchronization, eligibility–next-feature algebra, differentiation through training, function-space sensitivity, and generic run/seed variation are established constituent families. The bounded literature audit did not locate direct prior art for the complete conjunction of information-equivalent representations, a complete-recursion CURRENT-to-ONE-LAG tangent-energy contrast, realization-specific response-blind freezing, prospective fresh-twin learned-function ordering, and separately frozen external support. This is a bounded audit conclusion, not proof of priority. The paper claims neither primitive novelty nor firstness, and recent target-network work remains a live collision risk.

Provenance supports auditability of the tested interventions and chronology. CSTE remains a diagnostic of learned-function contrast ordering under the stated conditions. Absolute response, policy utility, nonlinear-remainder dominance, and mediation require distinct evidence. The external joint endpoint decision is determined by its registered criteria, independently of CSTE support; all claims retain their tested population scope.

J Retrospective N1 Diagnostic Comparisons

Study identity. This supplement reuses the original 48 N1 twins, 96 schedules, 1,536 Current learners, and recorded response contrasts. It does not add fresh validation data. The seven references were fixed before their new scores were computed; predictor construction accessed only the allowed Current inputs, and independent acceptance of the predictor freeze preceded response activation and scoring. These controls do not erase prior knowledge of the original results: every new comparison is retrospective exploratory. Original CSTE predictions, responses, 48 twin statistics, and three estimates, intervals, and statuses were transcribed from their frozen sources.

Reference definitions. For each energy-based reference, let $E^{(\ell)}_{S,c}$ be its configuration score and $P^{(\ell)}_{S;c_1,c_2} = E^{(\ell)}_{S,c_1} - E^{(\ell)}_{S,c_2}$ its contrast. The superscript ℓ labels a baseline, distinct from local forcing $b_{S,k,c}$. All contrasts retain N1's original directions, horizon $H18 - H6$ and TD($\lambda = .6$) minus TD(0), and the same row and twin keys. Below, $K = 12{,}000$ denotes completed parameter updates; native norms include all action heads and bias coordinates. Suppressing S, c only within definitions:

- B1 uses $E^{(\mathrm{B1})} = K^{-1} \sum_{k=0}^{K-1} \|b_k\|^2_{\mathrm{native}}$. It omits recursive propagation, the optimizer injection J, and the function-space readout, so it changes several factors rather than isolating one component. The production sum follows update order and divides once by K.
- B2 uses $E^{(\mathrm{B2})} = \|\theta_K - \theta_{K-1}\|^2_{\mathrm{native}}$, the last actual AdamW update, including weight decay and nonselected-head updates; it is not a gradient norm.
- B3a and B3b use the original final function-space squared-norm readout after propagating tangent state over the last $W = 1$ and $W = 1{,}200$ updates, respectively. W is a propagation window, distinct from temporal-credit H. Their state and readout conventions are detailed below.
- B4 uses $E^{(\mathrm{B4})} = \|z_{\theta,K}\|^2_{\mathrm{native}}$ from the full recursion. It includes only the final parameter tangent, not the previous-parameter or moment tangents, and does not transform it to RAW-equivalent coordinates.
- B5 directly sets every schedule contrast $P^{(\mathrm{B5})} = +1$. Its within-twin predictor difference is identically zero, so every row scores as a tie. This is a structural reference, not evidence that a substantive alternative has been ruled out.
- B6 directly assigns a fixed random ± 1 to each unique $(\mathrm{schedule}, \mathrm{row_key})$. One PCG64 draw with seed 2026090601 generates 1,536 bits in lexicographic key order; 0 maps to -1 and 1 to $+1$. Repeated keys reuse the value, and twin ties are retained. No seed search or inference over random-reference seeds is made.

Window state and function readout. At x_{K-W}, all four tangent components $(z_\theta, z_{\theta_{\mathrm{previous}}}, z_m, z_v)$ are set to zero. The original primal state $x = (\theta, \theta_{\mathrm{previous}}, m, v)$, update index, AdamW bias correction, and subsequent trajectory are retained. B3a starts at x_{11999}; B3b at x_{10800}. Each stream then follows the production one-sided tangent update, using its own directional clipping branch at a kink. It does not reuse the full tangent's branch or reset optimizer moments. The final parameter tangent is mapped through the same affine RAW-equivalent transformation, including bias mean correction. The negative difference of the two action-head tangents gives the margin tangent; its energy uses the original ν_{ref} Gram matrix, without division by W. This preserves the function readout, not the derivative of the original all-update perturbation path. No early/late additive decomposition or truncation-error bound is asserted.

Paired inference. Replace Ω by $P^{(\ell)}$ in equation 8 to obtain $D_i^{(\ell)}$, retaining all 16 rows and the same tie-aware κ. The reported difference is

$$\Delta_\ell = \frac{1}{48}\sum_{i=1}^{48}\big(D_i - D_i^{(\ell)}\big),$$

where D_i is the original frozen CSTE statistic, not a new estimate of it. The 48 complete twins are resampled with replacement using one shared $20{,}000 \times 48$ index matrix (PCG64 seed 2026090602). Linear percentile quantiles at $0.05/(2{\cdot}7)$ and $1-0.05/(2{\cdot}7)$ give nominal 99.285714% intervals with Bonferroni familywise $\alpha = .05$. This is approximate finite-sample inference. The seven-comparison family is separate from the original three N1 intervals and the external family; no new H/λ inferential claims are made. In the fixed B1–B6 order of Table 2, descriptive baseline concordances are 0.48177, 0.52214, 0.50391, 0.79948, 0.54167, 0.50000, and 0.50586. The B3b interval crosses zero; no equivalence or noninferiority margin was specified. Nearby concordance does not establish nearby tangent vectors, energies, individual predictions, or learned functions.

Reproducibility. The supplement used Python 3.11.15, NumPy 2.3.5, and the source-bound OpenBLAS 0.3.30 binary, with one BLAS thread, Haswell dispatch, and the frozen NumPy AVX-512 feature mask. This reconstructs a validated numerical runtime behavior, not proof of the original installation or CPU identity. The fixed eight-bundle pilot matched original final parameter/tangent and clipping records byte for byte; its newly recorded per-step states also matched an independent reference replay. Original N1 did not retain complete per-step primal states. All 384 bundles passed the prescribed coverage and numerical checks. The supplement's run ID is `20260906T090703Z`; its final package `20260908T222345Z_cloud_final` contains `protocol/S2/S2_PROTOCOL_v2.md`, `results/RESULTS.json`, predictor tables, inputs, code, runtime settings, gate reviews, and preserved diagnostic and recovery records. These support reproduction without enlarging the scientific claim or changing the original registry.

K Scope of Closely Related Work

Tables 11–12 compare research tasks, analysis objects, and uses. They do not test these works on a shared benchmark or infer missing capabilities from unreported experiments. In particular, function-space target updates, local squared sensitivity, and recursive training derivatives are prior foundations. Our response-blind condition concerns the corresponding One-Lag responses; it is not a claim that objective-driven adaptation in other work constitutes leakage.

Additional overlapping foundations. TDprop uses TD-aware Jacobi preconditioning (Romoff et al., 2021); replay-compatible λ-returns address temporal targets (Daley & Amato, 2019). TD interference and function-space learning dynamics are studied by Bengio et al. (2020); Lyle et al. (2022). DQN uses target networks for stabilization (Mnih et al., 2015); target-based TD and over-parameterized bootstrapping have dedicated stability analyses (Lee & He, 2019; Che et al., 2024). Stale momentum, online/target freshness, delayed value propagation, and linear target-update dynamics are further related objects (Bengio et al., 2021; Hendawy et al., 2026; Vincent et al., 2026; Lee, 2026). Training differentiation also supports bilevel optimization, auxiliary-question discovery, online objectives, self-tuning actor–critic methods, and learned update rules (Franceschi et al., 2018; Veeriah et al., 2019; Xu et al., 2020; Zahavy et al., 2020; Oh et al., 2020).

Limits of the comparison. This selected comparison extends the presentation of the bounded prior-art audit, not its coverage or a priority claim. CSTE is a diagnostic, not an RL algorithm. Absolute prediction remains unresolved. The external material endpoint gate was falsified under practical equivalence, leaving the joint replication criterion unmet. External CSTE assesses a separate learned-function outcome. Return prediction, policy-quality prediction, mediation, and remainder dominance require distinct evidence. The retrospective N1 reference comparisons in Appendix J are a separate experiment; their unresolved 1,200-update-window difference establishes no necessity of full recursion and no sufficiency of a shorter window.

Table 11: Target-network and sensitivity neighbors. Columns describe each work's studied object and use; they are not capability scores or a common performance benchmark. Workshop and preprint status are retained.

Work	Studied object	Use of that object
Fellows et al. (2023)	Compositions of inner TD updates and the Jacobian of the resulting target map.	Analyze conditioning and stability as the target-update interval changes.
Asadi et al. (2024)	Function-value matching between online and target networks.	Lookahead–Replicate learns target updates in function space.
Wu et al. (2025)	Linear TD, FQI, and partially fitted Q-iteration as matrix splittings and preconditioners.	Characterize their common linear system and differing convergence behavior.
Snyder et al. (2023) (workshop)	Training error as a function of the target update rate.	Optimize that rate during training.
Ly et al. (2026) (preprint)	Local squared Q-value gradients and recency weights over online-parameter history.	Merge recent parameter copies to update the target network.
Pleiss et al. (2026) (preprint)	Per-transition alignment of online and target TD errors.	Select transitions for training using the alignment signal.

Table 12: Training differentiation, temporal credit, and the present diagnostic task. Full training differentiation and interactions with target cadence are established foundations. Different purposes imply no ranking of the works.

Work	Studied object	Use of that object
Hernandez-Garcia & Sutton (2019)	Multi-step target construction and target-update frequency.	Empirically examine how these design choices interact in deep RL.
Maclaurin et al. (2015)	Derivatives through an entire SGD-with-momentum training procedure.	Optimize hyperparameters through the validation objective using reversible learning.
Xu et al. (2018)	Recursive derivatives of learned parameters with respect to return meta-parameters, with approximations discussed.	Adapt return components online using a subsequent meta-objective.
This paper	Full learner-recursion function-space tangent-energy differences for an exact one-update continuation-target synchronization perturbation.	Freeze CSTE before corresponding lagged responses; prospectively test finite learned-function contrast ordering across fresh N1 schedule twins.