Parameter-State Evaluation Graphs: Structure and Prediction in Finite-Horizon Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Learning procedures can differ not only in which updates they apply, but also in which historical parameter state each update reads. We study this choice in finite-horizon adaptation and represent it as a parameter-state evaluation graph. Common summaries describe the amount of staleness, but discard the update identity associated with each stale read. We show that this missing identity information matters. Graphs with the same updates, order, and aggregate delay statistics can still reach different endpoints because the same delays attach to different updates. An exact affine analysis explains this difference. From a single fully sequential reference trajectory, we predict the endpoints of unexecuted graphs by measuring the defect introduced by each stale read and propagating it through the remaining updates. In the primary supervised setting, this pathwise prediction improves graph ranking and captures matched-delay differences missed by aggregate summaries. Our results indicate that the parameter state read by each update should be treated as part of the finite-horizon learning algorithm, not merely as an implementation detail, and that part of its effect can be predicted from a single reference trajectory.

1 Introduction

A short adaptation trajectory is usually specified by its data, its update order, its objective and step size, and the number of updates (Finn et al., 2017). A routinely implicit degree of freedom is the parameter state that each update reads. Sequential adaptation reads the state left by the immediately preceding update, whereas batched or delayed execution may evaluate several updates from an older snapshot (Zheng et al., 2017; Woodworth et al., 2020). These choices leave the data and update order unchanged while differing only in which historical parameter state each update reads. We call the resulting dependency structure the *parameter-state evaluation graph*. We study this structure in finite-horizon adaptation, where the resulting differences can be measured directly.

Common summaries tell us how stale the computation is, but not which update receives each stale read. This missing assignment matters. On average, held-out loss increases as more reads become stale across all tested supervised checkpoints and supervised transfer datasets (Section 4). More pointedly, two schedules can run the same updates in the same order with the same aggregate delay statistics, yet reach different endpoints because the same delays are attached to different updates. Section 5 isolates this effect in a matched-delay comparison.

This raises a second question. If the read pattern changes the endpoint, can that change be predicted without executing every candidate graph? We show that it can. From a single fully sequential reference trajectory, we measure the finite defect introduced by each candidate stale read and propagate it through the remaining updates (Section 6). An exact affine analysis shows where this identity dependence enters (Section 5).

Read placement is therefore a structural design variable rather than an implementation detail. We study it first in a controlled family of short adaptation trajectories, then test how the conclusions change with initialization, step size, output calibration, composition

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.

rule, parameter scope, objective, and horizon. Across these settings, read-state assignment is an explicit, analyzable part of the finite-horizon algorithm, with regime-dependent effects.

Our contributions are threefold. First, aggregate delay statistics can miss information that matters. Matched-delay schedules with the same updates, order, and aggregate delay statistics reach different endpoints because the same delays attach to different update identities (Section 5). Second, these endpoint differences can be predicted before executing the target graph. A Fresh-reference predictor measures exact finite stale-read defects on one sequential reference trajectory and propagates them through the remaining updates. In the native-operator setting, this improves 120-order ranking from 0.569–0.616 to 0.722–0.772 (Section 6). Third, we characterize when these effects matter and when prediction is useful, separating graph sensitivity from prediction accuracy and selection utility across parameter scopes, objectives, and horizons (Sections 7 and 8).

2 Related Work and Positioning

Stale, delayed, and local updates. Prior work on stale, delayed, and local updates (Zinkevich et al., 2010; Dutta et al., 2018; Stich & Karimireddy, 2020) studies both their effects and how to correct them. DC-ASGD compensates stale gradients (Zheng et al., 2017), staleness-aware async-SGD scales updates by delay (Zhang et al., 2016), and Local-SGD studies synchronization frequency and local drift (Woodworth et al., 2020; Khaled et al., 2020; Lin et al., 2020; Luo et al., 2025). Many of these analyses summarize stale computation through aggregate quantities such as delay bounds, synchronization frequency, or local-step counts, which do not retain update identity. We ask a complementary question: when the update sequence is fixed, are such summaries sufficient to determine the resulting held-out behavior, or does it also matter which update receives each stale read (Section 5)?

Meta-learning, order sensitivity, and related update compositions. The FRESH and common-origin constructions are closely related to the serial and joint/batched updates analyzed in Reptile. Its sequential-minibatch expansion yields Hessian–gradient terms $H_i g_j$ that match our leading cross-gradient terms in the unclipped two-step limit (Nichol et al., 2018). We place these endpoints inside the full intermediate family and ask whether the corresponding finite-step, clipping-aware interactions predict endpoint differences.

Data-order effects are well established (De Sa, 2020), and federated SPFL contrasts sequential client updates with parallel aggregation (Liang et al., 2025). Recent work predicts data-source transfer order using Lie-bracket commutators of update fields and uses them to plan curricula (Sweeney, 2026a). That work varies training order, whereas our primary intervention holds the order fixed and varies the snapshot-read assignment. Our experiments use memoryless updates; extending the graph to momentum- or Adam-like optimizers would also require optimizer-state and clock variables (Sweeney, 2026b).

Common-origin delta composition also appears in task arithmetic and model merging (Ilharco et al., 2023), although there the deltas come from separate fine-tuning runs rather than a single fixed update sequence. Forward-sensitivity recursions of the form $h_t = A_t h_{t-1} + \eta_t$ likewise appear in forward-mode hypergradient methods (Franceschi et al., 2017), where the differentiated quantity is a hyperparameter rather than the parameter state read by each update. Related second-order cross-update terms also arise in finite-learning-rate SGD analyses (Smith et al., 2021), in a different optimization setting. TENT-style test-time adaptation (Wang et al., 2021; Hendrycks & Dietterich, 2019) provides a contrasting unsupervised regime in which instability has been documented (Niu et al., 2023).

Across these settings, the terminology differs, but the procedures all specify which parameter state each update reads. The evaluation graph isolates this shared structural choice.

3 Evaluation Graphs in Finite-Horizon Adaptation

Task, representation, and head. We evaluate 5-way episodes from miniImageNet (Vinyals et al., 2016) using a frozen supervised ResNet-12 (Oreshkin et al., 2018). BatchNorm remains in evaluation mode, and only a linear head is adapted. Each episode provides

25 support images per class, organized into five fixed five-shot groups $S_1, \ldots, S_5$, together with 15 queries per class. The head is initialized as the exact squared-Euclidean prototype classifier constructed from S_1 (Snell et al., 2017), so all conditions begin from the same step-zero classifier. Each comparison uses 500 shared episodes and four independently trained checkpoints: one main run and three additional runs, labeled strong-42–44. Computation uses float32, with float64 statistics and episode-bootstrap confidence intervals.

Operator. For group t with features F_t and labels y_t, the one-step operator clips the joint head-gradient norm at one and then applies weight decay,

$$\mathcal{O}_{S_t}(\theta) = \theta - \alpha\Big[g_t(\theta)\,\min\Big(1, \tfrac{1}{\|g_t(\theta)\|_2 + 10^{-6}}\Big) + \beta\theta\Big], \tag{1}$$

where g_t is the mean cross-entropy gradient of group t. Each checkpoint uses learned (α, β) values that are then held fixed (step sizes 0.473–0.500, weight decay 0.01). On transferred datasets, we reuse the main checkpoint's operator without dataset-specific retuning. We keep the operator fixed here to isolate read placement; Section 7 separately considers graph-specific step-size calibration under the same bounded search budget.

Refresh partitions. A contiguous refresh partition P selects a subset of the four boundaries between adjacent groups, yielding exactly $2^4 = 16$ schedules. Within a block $B = [a.b]$, every operator is evaluated at the block-start snapshot θ_B: $\Delta_t = \mathcal{O}_{S_t}(\theta_B) - \theta_B$ for $t \in B$, and the block endpoint is $\theta_B + \sum_{t \in B} \Delta_t$. Blocks then compose sequentially. Taking every block to have size one gives fully sequential adaptation (FRESH), while a single block of five gives common-origin aggregation (CO). Common-origin aggregation resembles gradient accumulation but is not identical to accumulation on a concatenated batch, because clipping and weight decay are applied separately to each group. In the unclipped, weight-decay-free, equal-group limit, the block-start delta sum is equivalent to one step on the mean block update with effective scaling $=|B|\alpha$. A five-update common-origin block therefore behaves like a coarse 5α step on the mean group gradient, analogous to the linear-scaling regime studied in large-batch optimization (Goyal et al., 2017). This scaling alone does not explain the placement effects studied below. All sixteen schedules use the same support union, group order, five operator evaluations, five backward passes, clipping rule, α, β, and episode bank. They differ only in the refresh pattern. We call the resulting dependency structure a *parameter-state evaluation graph*, defined by the snapshot reads and commit boundaries that determine which parameter state each update uses. The intervention changes the read-state assignment, not the update order. The same representation applies whenever updates may read different historical parameter states, including batched or delayed optimization, procedures that retain earlier model states, meta-learning, and test-time adaptation. In our controlled family, the graph structure is varied explicitly, which makes its counterfactual effects and first-order response tractable (Sections 5 and 6). Each schedule can also be viewed as a deterministic delayed-read schedule in which update t reads the state from $d_t = t - a_P(t)$ updates earlier (Appendix A). Two structural coordinates summarize a schedule. The first is the *stale-pair fraction*

$$S(P) = \tfrac{1}{\binom{5}{2}} \textstyle\sum_b \binom{m_b}{2} \in \{0,\, 0.1,\, 0.2,\, 0.3,\, 0.4,\, 0.6,\, 1\}, \tag{2}$$

the fraction of the ten ordered update pairs whose interaction term is omitted by stale evaluation. The second is the block count $B(P)$, which gives the idealized synchronization depth of the schedule.

The structure exposed by this family is also visible in a local expansion,

$$\theta_{\text{FRESH}} - \theta_P = \alpha^2 \sum_{s<t \text{ same block}} J_t u_s + O(\alpha^3)$$

(Appendix A). A schedule is therefore characterized by *which* pairwise interaction terms $J_t u_s$ are omitted by its stale blocks. Equivalently, every partition shares the same first-order term,

$$\theta_P = x_0 + \sum_t \big(\mathcal{O}_{S_t}(x_0) - x_0\big) + O(\alpha^2),$$

so graphs differ through second-order and finite-step interactions. At the operational step size $\alpha \approx 0.5$, we use this expansion only as a local approximation (Sections 5–6).

4 Evaluation Graphs Change Held-Out Behavior

The complete graph family produces structured finite-horizon variation. Held-out loss increases systematically as reads become staler. Across the four miniImageNet checkpoints, the stale-pair slope of query CE is 0.48–0.59, with confidence intervals entirely above zero. The same positive relation also appears on CUB-200, tieredImageNet, and FC100 under the same operator (0.380, 0.067, and 1.098; Table 1; Figure 1). A separate five-shot setting preserves the positive direction at a much smaller scale (0.010–0.018; Appendix A.2), showing that the effect size depends on the support budget. We quantify this relation using the within-episode OLS slope of query CE on $S(P)$, averaged over episodes with episode-bootstrap confidence intervals and Holm correction (Appendix A). Because the operator was learned on the sequential inner loop, the absolute effect size is specific to this operating regime; Sections 5 and 7 examine how it changes with operator and output scale. The relation is graded rather than strictly monotone. Across checkpoints, 12–19 of the 120 schedule pairs violate pairwise monotonicity. The positive pattern remains under level balancing, after removing the FRESH and CO endpoints, and relative to step zero (Appendix A.13). Radial logit magnitude does not account for the effect. Rescaling the CO endpoint to the FRESH radius preserves 95.4–95.6% of the CE gap, and the same direction also appears in accuracy and Brier across the primary vision experiments (Appendix A). We use raw held-out CE as the primary metric. Post-hoc output calibration removes 92.1–94.0% of the raw CE contrast while preserving every sign, leaving residual differences of 0.012–0.044 CE (Appendix A.18).

Table 1: Staleness slope: within-episode OLS slope of query CE on the stale-pair fraction $S(P)$ over all 16 contiguous refresh schedules (500 episodes per row; episode bootstrap CIs, 10,000 resamples; all raw $p < 10^{-4}$ (doubled two-sided empirical tail count, no continuity correction), Holm-rejected within the miniImageNet strong family and the transfer family). The 7-level column rebalances the schedule multiplicities; the interior column deletes the Fresh (S=0) and common-origin (S=1) endpoints entirely (post-hoc descriptive). Both the supervised miniImageNet checkpoints and the three transferred datasets (frozen operator transplant, no retuning) show positive slopes throughout.

Setting	slope $\bar{b}$	95% CI	7-level	interior
miniImageNet / main	0.530	[0.471, 0.592]	0.540	0.485
miniImageNet / strong-42	0.585	[0.525, 0.646]	0.600	0.519
miniImageNet / strong-43	0.579	[0.516, 0.642]	0.598	0.497
miniImageNet / strong-44	0.480	[0.426, 0.533]	0.493	0.423
CUB-200 (transplant)	0.380	[0.353, 0.408]	0.407	0.270
tieredImageNet (transplant)	0.067	[0.034, 0.102]	0.069	0.058
FC100 (transplant)	1.098	[1.006, 1.196]	1.112	1.020

5 Aggregate Delay Statistics Do Not Identify the Endpoint

Within the contiguous K=5 family, the stale-pair fraction uniquely determines the block-size multiset and therefore the delay multiset (Appendix B.8). The two S=0.6 schedules $[4, 1]$ and $[1, 4]$ thus share the same delay multiset $\{0, 0, 1, 2, 3\}$ and differ only in which update identities receive those delays. This rules out overall staleness as the explanation for their difference. Yet the two schedules recover very different fractions of the FRESH–CO gap (Table 2): keeping the later operators inside the stale block is substantially more damaging. The same pairwise differences also appear at the argmax decision level (post-hoc), with +1.5 to +1.7 percentage points of accuracy for $[4, 1]$ over $[1, 4]$ and +0.6 to +0.9 pp for $[3, 2]$ over $[2, 3]$; all checkpoint intervals exclude zero (Table 10). The failure therefore lies in the unlabeled summary: delay statistics retain how many stale reads occur and how stale they are, but discard which updates receive them.

An exact affine characterization. This insufficiency does not require nonlinear curvature. It already appears in an exactly solvable affine model, where matched-delay differences arise from two sources: noncommutativity between update maps and heterogeneity across update identities. Let each per-group update be $F_t(x) = A_t x + b_t$, with all updates in a block evaluated at the block-start state (Section 3). For the matched pair $P = [4, 1]$,

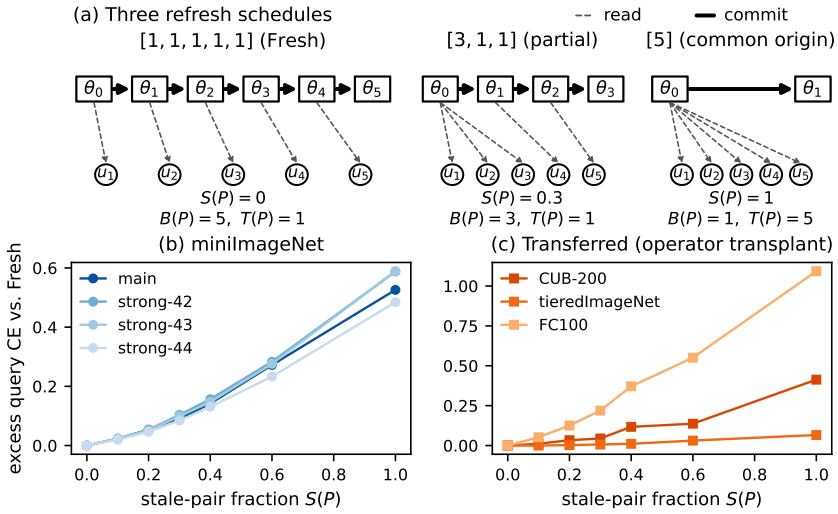


Figure 1: Evaluation graphs and their held-out effect. (a) Three refresh schedules. Dashed edges denote snapshot reads and solid edges denote commits; updates within a block all read the same block-start state. (b, c) Excess query CE relative to the fully sequential schedule, averaged within each staleness level, for the four miniImageNet checkpoints and the three transferred datasets using the same operator. The positive graded relation appears in every setting.

Table 2: Two schedules with identical aggregate delay statistics: same stale-pair fraction, same block count, same unlabeled delay multiset; only the placement of the delays across update identities differs. Recovery of the FRESH–CO gap (normalized so that FRESH= 1 and CO= 0; reference points, not bounds) is post-hoc descriptive; the paired-difference CIs exclude zero (full per-checkpoint record: Table 7).

Schedule	$S(P)$	$B(P)$	Delay multiset	Recovery of the gap
$[4,1]$	0.6	2	$\{0,0,1,2,3\}$	0.86–0.87
$[1,4]$	0.6	2	$\{0,0,1,2,3\}$	0.09–0.20

paired held-out CE difference: 0.33–0.41

$Q = [1,4]$, with $S_i = A_i - I$ and $S_{234} = S_2 + S_3 + S_4$, the endpoint difference is exactly

$$\theta_P - \theta_Q = \big(S_5 S_{234} - S_{234} S_1\big)x_0 + \big(S_5 b_{234} - S_{234} b_1\big).$$

Its linear part separates the two sources. The commutator $[S_5, S_{234}]$ captures noncommutativity, while $S_{234}(S_5 - S_1)$ captures which update identity occupies each stale-read position. Under $A_t = I + \alpha M_t$ and $b_t = \alpha c_t$, the two-block difference is exactly proportional to α^2, with no cubic remainder (Appendix B.8). If every update shares the same affine map, the block maps commute and the endpoint depends only on the block-size multiset, so matched-delay partitions coincide. Once the update maps depend on identity, equal delay multisets no longer determine the endpoint; Appendix B.8 gives the corresponding bound under bounded affine heterogeneity. The same decomposition also distinguishes read placement from an ordinary order swap. For affine maps $F_a(x) = (I + S_a)x + b_a$ and $F_b(x) = (I + S_b)x + b_b$, the linear part of an adjacent order swap is the commutator $[S_b, S_a]$, with translation term $S_b b_a - S_a b_b$. The read-placement intervention instead keeps the execution order fixed and changes which identities occupy the stale positions, adding the identity-heterogeneity term $S_{234}(S_5 - S_1)$. This invariance is specific to the homogeneous affine case; nonlinear state dependence can still produce placement effects even when the nominal update map is shared across identities. Section 6 asks whether these missing finite-step effects can also be predicted in the nonlinear system.

Support order provides a second axis on which update identity matters. Within an episode, changing only the support order spans a median 0.49–0.63 CE, with the natural order

Table 3: Matched-delay pair contrasts are prospectively predicted (500-episode bank, 120 support orders). ρ: median per-episode Spearman between the predicted and actual pair contrast, range across the four checkpoints; sign agreement likewise. Median within-episode range of the actual pair contrast in CE. Contrast scale and tie counts are descriptive; full checkpoint-wise results are reported in Appendix B.7 (Table 38).

Pair	ρ	sign agr.	range (CE)	contrast type
$[3,2]$ vs $[2,3]$	0.986–0.988	0.965–0.969	0.060–0.070	pure placement (all checkpoints)
$[4,1]$ vs $[1,4]$	0.938–0.980	0.962–0.977	0.034–0.077	jointly tuned (main)
$[3,1,1]$ vs $[1,1,3]$	0.996	0.980–0.983	0.073–0.086	pure placement (all checkpoints)

close to the order average (Figure 2, Appendix A.6). A first-order initialization-anchored interaction model ranks the harder 120-order variation poorly ($\rho = 0.40$–0.42; Appendix B), indicating that the missing component is order-dependent. A simple post-hoc pattern is that terminal-block size $T(P)$ strongly organizes the endpoints. Mean recovery decreases monotonically at every checkpoint from 0.96 at T=1 to 0.09–0.20 at T=4, so schedules that keep the final updates fresh retain much more of the sequential advantage (Table 8). A schedule-mean regression attributes 92% of between-schedule CE variation to T, compared with 67% for S alone. Exact pathwise replay gives the corresponding finite-step picture: terminal defects pass unchanged to the endpoint, while nonterminal defects attenuate to about 89% with nearly preserved direction (Appendix B).

6 Predicting Endpoint Changes

If a candidate graph had been executed, what would its changed reads have done to the endpoint? Section 5 showed that read placement changes the endpoint. We now ask whether that change can be predicted before executing the graph. At the native operator, Fresh-path transport adds substantial rank information. On an independent evaluation bank, it ranks the 120 support orders for the $[3,2]$ schedule at median Spearman 0.722–0.772, compared with 0.569–0.616 for local defects alone and 0.448–0.514 for the initialization anchor (10 distinct endpoints; Appendix B; Table 33). With joint initialization and step-size calibration, the same predictor reaches higher absolute rank fidelity on the matched-delay placement comparisons, with per-episode Spearman 0.938–0.996 (Table 3). Transport improves on the initialization anchor in all 12 pair-by-checkpoint comparisons, while its advantage over local defects is smaller and depends on the regime; on the native matched-delay pairs, the comparison is mixed (Appendix B).

Algorithm 1: Fresh-reference finite-step graph predictor.

Input: initialization x_0, support order π, ordered update fields $\{v_g\}$, candidate evaluation graph G.
Here $F_g = \mathcal{O}_{S_g}$, $v_g = F_g - \mathrm{Id}$, $A_t = DF_{\pi(t)}(x_{t-1})$, and $a_G(t)$ is the first update position of the block of G containing t.
(1) **Run** the Fresh trajectory once: $x_t = F_{\pi(t)}(x_{t-1})$;
(2) **Measure** each stale-read defect of G along the Fresh path: $\tilde{\eta}_t^G = v_{\pi(t)}(x_{a_G(t)-1}) - v_{\pi(t)}(x_{t-1})$;
(3) **Transport** each defect through the remaining Fresh updates: $\Phi_F(K{:}t) = A_K \cdots A_{t+1}$ (path-local JVPs);
(4) **Predict** the endpoint: $\hat{\theta}_G = x_K + \sum_{t \in G} \Phi_F(K{:}t)\tilde{\eta}_t^G$.
Algorithm 1 is the first-order response at λ=0 of the continuous read-state intervention defined in Appendix B.

For the selection experiments, predicted endpoints are scored using support-side validation only; query data are not used to choose the graph (Appendix B.3).

From replay to prediction. Algorithm 1 predicts unexecuted read patterns from a single Fresh reference trajectory (Section 5; Appendix B). It has no fitted graph-level coefficients and does not execute the target graph. Its field-evaluation cost grows quadratically in K, whereas the number of candidate graphs grows exponentially, so graphs sharing a support order and operator can reuse the same Fresh trajectory. Graph-specific operators require separate trajectories, and the current K=5 selection implementation is still slower

than exhaustive scoring (Appendix A.15). The predictor also extends to the complete 128-graph K=8 family (Appendix B), where full transport improves rank correlation over local defects by $\Delta\rho = +0.080$ $[0.067, 0.099]$. In a diagnostic using query scores, the predicted top-16 contains the actual best graph in 50% of episodes with near-zero regret.

With full-support initialization and graph-specific step sizes, the $[3, 2]$ order-ranking correlation reaches 0.975–0.979 across the four checkpoints, and the six tested shapes range from 0.885 to 0.993 (Appendix B.7). These higher correlations reflect the jointly calibrated setting and should be interpreted separately from the native-operator results above. Because the terminal-block size $T(P)$ is fixed across the $[3, 2]$ orders, this variation cannot be explained by terminal-block size alone.

The predictor does not transfer uniformly across adaptation regimes. Under supervised CLIP ViT-B/16 LoRA adaptation, graph sensitivity remains substantial but pooled median rank correlation falls to 0.067. In the tested entropy-minimization TTA configuration, transport adds little because local defects already capture most of the graph-wise structure (Appendices A and B).

Support-side reliability signals. We next ask whether support-side quantities can indicate when the predictor is likely to be inaccurate. We consider three pre-specified diagnostics, κ_0, A_G, and D_G (Appendix A.22). On a 500-episode reliability subset with All25 initialization and graph-specific step sizes, κ_0 and D_G correlate strongly with per-episode absolute query-CE prediction error (Spearman 0.824–0.843 and 0.790–0.826; worst-decile AUC 0.91–0.94), whereas A_G shows little useful within-regime signal (Table 32). In the CLIP-LoRA setting, the magnitudes of κ_0 and D_G are much larger than in the supervised setting (medians 39× and 103×; Holm $p < 10^{-190}$), although their association with prediction error varies across datasets (Appendix A.22). Post-hoc analyses also associate both diagnostics with order-ranking error after accounting for effect magnitude. A separate development/confirmation experiment at the native fixed operator shows the same qualitative pattern. On the confirmation bank, κ_0 discriminates episodes with larger paired prediction error at every checkpoint (ρ = 0.724–0.806). The association remains after conditioning on predicted displacement, with partial $\rho = 0.579$–0.715 (Appendix A.22).

7 Breadth and Regime Dependence

Regime transfer. Graph sensitivity persists as the adaptable parameter set expands. Moving from the linear head to block 4 plus the head (11 tensors), blocks 3–4 plus the head (20), and the full network (38) leaves the $S(P)$ slope positive throughout at a shared conservative step size (Table 11; Appendix A.7). Frozen CLIP features behave differently. At the transferred step size, the slopes are close to zero and slightly negative, and the negative direction persists across all 56 sweep settings while increasing sharply in magnitude (k^*=34: −5.12 to −7.03 CE; Table 14). Adapting the backbone restores a pooled preference for fresher reads. With LoRA (Hu et al., 2022) on the CLIP (Radford et al., 2021) visual tower, last-block adaptation favors FRESH on all four datasets. Adapting the last four blocks gives the same direction on three datasets and an approximately null effect on DTD (Table 15). A ViT-L/14 last-block replication also gives a pooled FRESH-favoring effect, although only two of the four datasets favor FRESH individually and the pooled difference is concentrated in FGVC-Aircraft (Appendix A.10).

The preferred direction also depends on the adaptation objective. In the supervised ResNet/head-adaptation settings above, fresher reads perform better. In the tested TENT-style (Wang et al., 2021) entropy-minimization setting on ImageNet-C (Hendrycks & Dietterich, 2019), the direction reverses. With an ImageNet-pretrained ResNet-50 (He et al., 2016) and only the 106 BatchNorm affine tensors adapted, FRESH has higher held-out CE than common-origin adaptation ($\Delta_{\mathrm{FC}} = +0.198$, 95% CI $[0.189, 0.207]$; stale-pair slope -0.199 $[-0.208, -0.189]$), with the same direction across the four tested corruptions (Table 12). Replacing entropy minimization with a supervised oracle-CE signal restores the supervised direction, showing that the reversal depends on the adaptation objective and supervision signal (Appendix A.9).

Table 4: Initialization and graph-specific step-size calibration both attenuate the FRESH–CO endpoint contrast. Under their joint control, the residual contrast is 0.002–0.010 CE across checkpoints (500 shared evaluation episodes each; range over main/strong-42/43/44; equal-budget graph-specific selection over the fixed bounded grid). Tuned cells compare tuned policies rather than pure placement interventions. Individual checkpoint intervals are mixed in the jointly controlled setting (Appendix A.16).

Init	LR policy	FRESH−CO CE	Reading
S_1	shared	−0.636 to −0.537	large under shared step size
S_1	graph-selected	−0.129 to −0.118	attenuated by LR calibration
All25	shared	−0.520 to −0.444	attenuated by richer initialization
All25	graph-selected	−0.010 to −0.002	strongly attenuated under joint init+step-size condition

Scale and composition. Initialization and optimization scale both moderate the effect. Full-support initialization alone leaves the FRESH–CO contrast large, while graph-specific step-size selection reduces it by roughly a factor of five under the same search budget. Combining the two attenuates the contrast further (Table 4). Same-(S, B) positional dispersion falls from 0.167–0.264 CE under a shared α to 0.009–0.013 CE under joint tuning (Table 23). The remaining identity-conditioned differences are nevertheless not eliminated. Matched-delay contrasts still span median within-episode ranges of 0.034–0.086 CE and remain predictable (Table 3). Thus, a smaller graph effect does not necessarily imply lower predictive fidelity. A first-step scale-calibrated comparison targets the FRESH first-commit displacement (Appendix A.19). At the resulting scales, the FRESH–CO contrast becomes smaller but remains substantial; the $[4, 1]$–$[1, 4]$ comparison keeps its direction, whereas the $[3, 2]$–$[2, 3]$ preference reverses (Table 30). Because the grid does not contain the exact $1/3$ and $1/5$ scalings, the outcomes, including the reversal, remain descriptive: they demonstrate strong scale dependence but do not by themselves explain the reversal (Appendix A.19). The block-mean construction probes a different intervention, because moving a block also changes which update identities receive the block-dependent scaling. Its near-zero matched-delay contrasts nevertheless show that the composition rule can strongly reduce the observed placement effect (descriptive, main checkpoint only; Appendix A.19).

Other boundaries. Graph sensitivity also remains visible as the support budget and initialization change. We vary the per-group support budget $m \in \{1, 2, 5\}$ under first-group and full-union initialization and include a disjoint-initialization condition (Figure 5; Table 25, Appendix A.17). The disjoint condition has the largest observed contrast, showing that support reuse is not required. Within the first-group initialization family, the preferred endpoint also changes with support scale under the calibration policy fixed from the least-targeted setting. The phenomenon is not limited to vision. In a non-saturated Banking77 head-adaptation setting, adaptation improves CE by 27.5% relative to step zero, while graph choice changes the held-out endpoint. Common-origin adaptation achieves 0.0128 lower CE than FRESH, whereas FRESH achieves 3.50 percentage points higher accuracy (Table 31, Appendix A.20). Thus, probabilistic loss and 0–1 accuracy can prefer different evaluation graphs. The graded effect also appears in the complete 128-graph family at eight updates (Holm $p = 0.0012$). The FRESH–CO direction persists across the 48 tested operator perturbations, and a separate plain-SGD run retains the graded effect over the full sixteen-graph family (Appendix A.3).

Beyond finite-horizon adaptation. The same structural distinction appears in other learning procedures. Local and distributed SGD read synchronized or local snapshots, while asynchronous optimization reads older iterates (Zheng et al., 2017; Stich & Karimireddy, 2020). Continual-learning procedures may retain teacher, anchor, or historical model states, and meta-learning and test-time adaptation compose inner-loop, memory, and anchor states (Finn et al., 2017; Wang et al., 2021). Such procedures can match in data, update count, and initialization while differing in which state each computation reads. These examples share the same structural distinction, although the present experiments do not establish that the measured effect sizes or predictor accuracy transfer to those settings (Figure 6, Appendix A.23).

8 From Prediction to Decision

Support losses contain useful information for selecting evaluation graphs. The paired jackknife selector (PJGS; Appendix A.12) uses paired support-side validation to choose among candidate graphs. PJGS improves over both structural anchors at the four primary checkpoints, and the same selection signal extends to longer horizons, wider parameter scopes, and additional datasets. The tieredImageNet contrast, however, is not distinguishable from zero (Table 17). The number of jackknife folds trades selection accuracy against computation, and why support-side loss predicts the held-out graph ranking remains an open question.

We next ask whether the selection gain survives output calibration. We combine full-support initialization with per-graph learning rates chosen on a disjoint development bank and compare PJGS with the best fixed graph from that bank on 800 new episodes (Appendix A.18). On raw logits, PJGS has lower held-out CE in four of the six checkpoint-by-budget comparisons and outperforms no adaptation in all six (Table 27). After policy-specific temperature calibration, its advantage over the fixed graph remains at B=2 but not at B=3. The comparison with no adaptation remains positive in point estimate but is no longer statistically distinguishable from zero. To test whether prediction can replace exhaustive candidate scoring, PG-JGS scores all ten candidate graphs from the reference trajectory and executes only the selected graph on an independent 800-episode evaluation set. It improves over the strong fixed graph at B=2. At B=3, the contrast is not distinguishable from zero, and the noninferiority criterion relative to exhaustive PJGS is not met. Output calibration further reduces the raw gains (Tables 22 and 29; Appendix A.15).

Why does approximate scoring not always preserve the exact choice? A deterministic stability condition guarantees agreement with exhaustive selection whenever twice the centered worst-case score error is smaller than the gap between the best and second-best candidates. This condition holds for 27–44% of episodes. All observed regret occurs outside this guaranteed subset, where the median best–second margin (0.002–0.007 CE) is comparable to the prediction error (Appendix A.15). The condition is diagnostic rather than available to the selector at decision time. Predictive ranking, agreement with exhaustive selection, and calibrated utility are therefore distinct questions.

9 Discussion and Limitations

Our central finding is that changing which historical parameter state each update reads can change the endpoint, even when aggregate staleness statistics are unchanged. The existence of these graph effects does not imply that the predictor transfers equally well across regimes. Algorithm 1 is the exact first-order response at the FRESH endpoint of the continuous read-state intervention, but its accuracy for finite graph changes varies substantially across settings (Section 6). The magnitude of the observed effect is also sensitive to optimization and output scale. Output calibration removes most of the raw-CE contrast, and graph-specific step sizes attenuate it further (Section 7). Banking77 further shows that graph preference can depend on the evaluation metric, with CE and accuracy favoring different endpoints.

Our strongest evidence comes from supervised vision adaptation. The language, CLIP-LoRA, and TTA experiments probe other regimes, but the support-side diagnostics are not yet reliable indicators across all of them (Section 6). Most predictor experiments use the five-update contiguous-block family. The method also extends to the complete 128-graph family at K=8, although the gain from transport is smaller there; non-contiguous schedules and graphs with optimizer state remain untested (Sections 6 and 2). Because the three additional checkpoints share the primary model's episode banks, we treat them as replications across the tested models rather than as random draws supporting population-level inference over training seeds.

10 Conclusion

Finite-horizon adaptation is specified not only by which updates are applied and when, but also by which model state each update reads. Aggregate delay summaries discard that identity information, and part of its effect can be predicted from a single reference trajectory. In the primary supervised regime, schedules that keep later updates fresh tend to retain more of the sequential advantage. More generally, read-state assignment should be treated as an explicit, analyzable part of finite-horizon adaptation rather than as an implementation detail.

Reproducibility Statement

The experimental setting is defined in Section 3: 5-way miniImageNet episodes (25 support images per class in five fixed five-shot groups, 15 queries), a frozen ResNet-12 backbone with a prototype-initialized linear head, the one-step operator with per-checkpoint learned rates, and all sixteen contiguous refresh schedules evaluated on 500 shared episodes per checkpoint with episode bootstrap confidence intervals. Code, episode manifests, model checkpoints, and analysis scripts for all reported experiments are included in the supplementary repository; the repository README maps each main-text and appendix result to its implementation and stored records. Experiment-specific statistical procedures and prospective-evaluation details are documented in Appendices A and B.

AI Use Statement

Generative AI tools were used under author direction to assist with manuscript writing and editing and with writing and debugging evaluation code. The authors designed the study, developed the methods, performed all experiments, interpreted the results, verified all reported numbers, and take full responsibility for the final paper.

References

Christopher M. De Sa. Random reshuffling is not always better. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5957–5967, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/42299f06ee419aa5d9d07798b56779e2-Abstract.html.

Sanghamitra Dutta, Gauri Joshi, Soumyadip Ghosh, Parijat Dube, and Priya Nagpurkar. Slow and stale gradients can win the race: Error-runtime trade-offs in distributed sgd. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 803–812, 2018.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1165–1173. PMLR, 2017.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4519–4529, 2020.

Haoyuan Liang, Shilei Cao, Guowen Li, Zhiyu Ye, Haohuan Fu, and Juepeng Zheng. SPFL: Sequential updates with parallel aggregation for enhanced federated learning under category and domain shifts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations (ICLR)*, 2020.

Ruichen Luo, Sebastian U. Stich, Samuel Horváth, and Martin Takáč. Revisiting LocalSGD and SCAFFOLD: Improved rates and missing analysis. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. URL https://arxiv.org/abs/1803.02999.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 719–729, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/66808e327dc79d135ba18e051673d906-Abstract.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: SGD with delayed gradients. *Journal of Machine Learning Research*, 21(237):1–36, 2020.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. *arXiv preprint arXiv:2606.29554*, 2026b. URL https://arxiv.org/abs/2606.29554.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3630–3638, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10334–10343. PMLR, 2020.

Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. Staleness-aware async-sgd for distributed deep learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2350–2356, 2016. arXiv:1511.05950.

Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, and Tie-Yan Liu. Asynchronous stochastic gradient descent with delay compensation. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4120–4129. PMLR, 2017.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alexander J Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

A Experimental Protocols, Robustness, and Scope

Roadmap. Appendix A reports experimental protocols and robustness analyses: Sections A.1–A.2 cover analysis scope and the canonical five-shot extension; A.3–A.11 structural robustness and breadth (operator and metric robustness, step-size sweep, support order and terminal-block structure, parameter scope, TTA, foundation backbones, horizon); A.12–A.15 selection and decision-level evidence; A.16–A.19 initialization, scale, calibration, and composition controls; A.20–A.21 language and cross-modality transfer boundaries; A.22 predictor reliability; A.23 a scope schematic. Appendix B reports the predictor derivation and its evaluation (B.1–B.7) and the exact affine characterization (B.8).

The local expansion of Section 3 reads

$$\theta_{\text{FRESH}} - \theta_P = \alpha^2 \sum_{\substack{s<t \\ s,t \text{ same block}}} J_t u_s + O(\alpha^3), \tag{3}$$

where, in the convention of Eq. (1), $\mathcal{O}_t(\theta) = \theta + \alpha\, u_t(\theta)$ with $u_t(\theta) = -c_t(\theta)\, g_t(\theta) - \beta\theta$ (implemented gradient clipping and weight decay, $c_t(\theta) = \min(1,\, 1/(\|g_t(\theta)\|_2 + 10^{-6}))$), and $J_t = Du_t(\theta_0)$. The expansion is local: it applies on a neighborhood of θ_0 on which the active clipping branch is unchanged, so the implemented update field is differentiable there; no Taylor claim is made exactly at a clipping-switch surface. It was verified numerically on synthetic heads over all sixteen partitions ($O(\alpha)$ relative error); at the operational step size it is only a local approximation (Section 3).

Corollary (count/depth insufficiency). Equality of stale-pair fraction and block count does not imply equality of finite-horizon endpoints. Under the local expansion, two partitions P and Q differ at second order whenever the sums of their omitted cross-update terms differ:

$$\theta_P - \theta_Q = \alpha^2 \Big[\sum_{(s,t)\in M(Q)} J_t u_s - \sum_{(s,t)\in M(P)} J_t u_s \Big] + O(\alpha^3),$$

where $M(P) = \{(s,t) : s < t \text{ and } s,t \text{ lie in the same block of } P\}$. Because $S(P)$ and $B(P)$ fix only the omitted-pair count $|M(P)|$ and the number of blocks—not which pairs are omitted, nor the values of their cross terms—schedules with equal (S, B) can realize different second-order sums. The corollary is subject to the same local scope as Eq. (3) (fixed clipping branch, a neighborhood of θ_0): it explains why scalar count/depth summaries cannot identify the endpoint, not why any particular schedule generalizes.

Delayed-read form. For the memoryless operator used throughout, the construction is equivalently a deterministic delayed-read schedule: letting $a_P(t)$ denote the first update position of the block containing t, update t reads the state left $d_t = t - a_P(t)$ updates earlier, and the stale-pair fraction satisfies $\sum_t d_t = \binom{K}{2} S(P)$. The two schedules matched at S=0.6, B=2 in Section 5 even share the same delay multiset $\{0, 0, 1, 2, 3\}$, so neither the total delay, the block count, nor the delay multiset identifies the finite-horizon endpoint—only the placement of the delays relative to the update identities does.

Analysis conventions. Confirmatory analyses were specified before outcome inspection, including their criteria, thresholds, and multiplicity families; post-hoc analyses are identified as descriptive. Throughout, *pre-specified* means fixed in writing before the corresponding outcome data were examined. Episode banks were checked for support/query disjointness, and endpoint replays and parameter-invariance checks were used to verify the reported runs.

Statistical conventions. Bootstrap p-values in the paper's confirmatory families are two-sided doubled empirical tail counts at B=10^4 (no continuity correction; zero crossings report as $p < 10^{-4}$), Holm-adjusted within each pre-specified family; the other experiment families use the plus-one convention and state it where they report finite-resolution p-values. The principal inference families are:

Analysis family	Resampling unit	p-value convention	Multiplicity
Sixteen-schedule slopes; initialization × step-size positional estimands (Tables 1, 23)	episode	doubled empirical tail, floor 10^{-4}	Holm per family
Parameter-scope ladder (Table 11)	episode cluster	directional plus-one, floor 10^{-4}	Holm–3
Label-free selector and TTA contrasts (Tables 20, 12)	slot / paired-slot cluster	plus-one two-sided	Holm–4; single primary for TTA
Selection utility and prospective-predictor confirmations, incl. foundation and horizon (Tables 17, 27, 22, 15, 16)	episode	plus-one two-sided; one-sided where the direction was fixed in advance	family-specific Holm

The two-sided plus-one families use two stated constructions: Table 20 uses the absolute-deviation form $p = (\#\{|T_b - T_{\text{obs}}| \geq |T_{\text{obs}}|\} + 1)/(B{+}1)$ with floor $1/(B{+}1)$, whereas the remaining two-sided families double one-sided tails (floor $2/(B{+}1)$); both at $B{=}10^4$.

The strict monotonicity criterion was not met (hence the graded-on-average wording of Section 4), and the initialization-anchored equal-count diagnostic is superseded by the revised position-space estimand of Appendix B.

A-I. Protocol, Metrics, and Analysis Scope

A.1 Analysis Scope and Provenance

Two implementation issues affected intermediate development records but not the results reported here. First, one development predictor bank was exposed to query labels during pipeline testing. We discarded that bank and generated a new independent bank, fixing predictions before query evaluation (Appendix B.3). Second, one stored no-adaptation array contained values from a single checkpoint; the affected statistics were recomputed separately for each checkpoint from the evaluation records. All reported stale-pair slopes use the within-episode OLS definition in Section 4. No reported result uses the discarded predictor bank.

A.2 Canonical Five-Shot Extension

Canonical-five-shot convention. To preserve the canonical-five-shot convention, the step-zero prototype head is constructed from all five support images per class; those same five images are then partitioned into five one-shot update groups for the refresh intervention. This extension evaluates the model through precomputed frozen features rather than full-model execution; a numerical comparison against the full-model implementation is 3–7 × 10^{-5}, so exact identity is not expected there.

Table 5: Canonical five-shot sixteen-schedule refresh slopes (500 episodes, five one-shot groups per class). Numerical comparison against the full-model execution: max-abs 3–7×10^{-5} (exact identity not expected; Appendix A).

Checkpoint	CE slope	95% CI	Holm adj. p
main	0.018	[0.006, 0.031]	(desc.)
strong-42	0.013	[0.005, 0.022]	0.0016
strong-43	0.017	[0.007, 0.029]	0.0012
strong-44	0.010	[0.002, 0.020]	0.0162

A-II. Structural Robustness and Breadth

A.3 Operator Robustness Across 48 Settings

The operator sensitivity analysis spans three α-scales ($0.5\times$, $1\times$, $2\times$) $\times$ clipping on/off $\times$ weight decay on/off across four checkpoints. The Fresh−CO direction is negative (Fresh-favoring) in all 48 operator settings. The `plain` cell implements β=0 and no clipping, i.e. $\theta \leftarrow \theta - \alpha g$ (plain SGD); its Fresh−CO CE at the $1\times$ scale is -4.86 (main), confirming that the direction is not an artifact of the learned clipped operator. Under the same plain-SGD setting and 500-episode bank, the sixteen-schedule CE slope is 4.92 $[4.53, 5.31]$, and the endpoint losses (Fresh 4.69, CO 9.55) are much larger than under the learned operator (Fresh 4.10, CO 4.62): the graded effect is present under plain SGD, and effect magnitude is strongly operator-dependent. The per-checkpoint (α, β) are meta-learned scalar parameters of the bilevel procedure whose inner loop performs the same sequential head adaptation used at evaluation; the weight decay moves from its 5×10^{-4} initialization to 0.01, and the learning rate converges to or near its 0.5 upper bound (strong-43 exactly at the bound).

Radial-magnitude control. Rescaling the common-origin endpoint to the FRESH endpoint's radius leaves 95.4–95.6% of the original FRESH–CO held-out-CE gap intact across the four checkpoints (paired mean FRESH–rescaled advantages -0.46 to -0.56 CE on the same 500-episode bank; Holm-significant in the strong checkpoints). Radial magnitude alone therefore does not separate the schedules.

A.4 Accuracy and Brier Robustness

Table 6: Accuracy and Brier robustness: slopes across the sixteen refresh schedules. Accuracy slope is negative in all seven cells; six of seven confidence intervals exclude zero. Brier slope is positive in all seven. Fresh−CO accuracy is the endpoint gap in percentage points.

Setting	Acc. slope	Fresh−CO acc (pp)	Brier slope
miniImageNet / main	−0.0179	1.7	0.0354
miniImageNet / strong-42	−0.0253	2.5	0.0530
miniImageNet / strong-43	−0.0260	2.6	0.0516
miniImageNet / strong-44	−0.0225	2.2	0.0445
CUB-200 (transplant)	−0.0242	2.6	0.0800
tieredImageNet (transplant)	−0.0011	0.1	0.0029
FC100 (transplant)	−0.0362	3.4	0.0774

A.5 CLIP Calibration and Step-Size Sweep

The step-size calibration uses the evaluation DTD bank's last fifty episodes (support-side displacement only; no query metrics) and measures the median first-step support-side displacement under the supervised main operator ($r_{\text{sup}} = 2.198$) and under the CLIP transplanted operator at k=1 ($r_{\text{clip}} = 0.064$), giving a ratio of 34×. The step-size multiplier k^*=34 is selected to match this displacement; no query labels are used. Table 14 reports slopes across the full seven-scale, two-encoder, four-dataset sweep; clipping is never activated (rate 0 at all scales). At the transplanted scale the slopes are near zero and slightly negative ($|\cdot| < 6.5 \times 10^{-4}$ across all 56 settings), so the supervised positive direction does not recover.

A.6 Support-Order Sensitivity and Terminal-Block Structure

Within an episode, changing only the support order spans 0.49–0.63 CE at the median, with the natural order near the order-mean (Figure 2); per-episode order effects largely cancel under averaging. The two schedules matched at S=0.6, B=2 also differ substantially in absolute terms: paired per-episode CE differences of +0.33 to +0.41 across checkpoints (Table 7, descriptive).

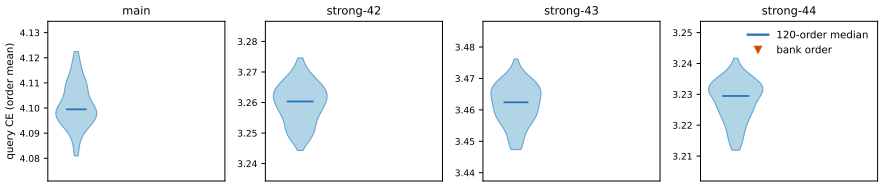


Figure 2: Support-order sensitivity within sequential composition: per-order mean query CE across all 5! = 120 permutations. The natural bank order sits at the center of the order-mean distribution; within a single episode, order alone spans 0.49–0.63 CE, while per-episode order effects largely cancel under averaging.

Table 7: Absolute paired held-out CE separation between the two schedules matched at S=0.6, B=2 (Section 5). Per-episode paired differences $\mathrm{CE}_i([1]|[2,3,4,5]) - \mathrm{CE}_i([1,2,3,4]|[5])$ on the paper's primary 500-episode bank per checkpoint (the pre-specified primary evaluation bank, the same bank as Table 2; an independent draw from the output-calibration bank behind Table 28 and from the initialization–step-size bank behind Table 23); paired mean with percentile 95% episode-bootstrap CI (10,000 resamples). Post-hoc descriptive uncertainty summary for a pair highlighted after inspection of the schedule family: no new hypothesis test, no multiplicity correction, and no confirmatory role is claimed.

Checkpoint	mean paired ΔCE	95% CI
main	+0.411	[+0.362, +0.462]
strong-42	+0.408	[+0.361, +0.457]
strong-43	+0.392	[+0.344, +0.443]
strong-44	+0.329	[+0.287, +0.374]

Terminal-block structure. A descriptive schedule-level regularity emerges from the family: the terminal-block size $T(P)$ organizes recovery of the sequential advantage, with grouped mean recovery decreasing monotonically at every checkpoint (Table 8). To test whether this regularity survives outside the discovery family, the coordinate, schedule list, and reported metrics (grouped mean recovery by T; Spearman and linear R^2 of schedule-mean CE on T; residual association after removing S) were fixed before the validation schedules were examined, and all statistics were computed from previously stored results (Table 9). The grouped trend replicates in all three transferred datasets; the canonical five-shot and plain-SGD cells do not preserve the exact grouped ordering throughout, so the terminal-block regularity remains a main-regime descriptive structure.

Table 8: Recovery of the sequential advantage by refresh schedule (miniImageNet; mean over four checkpoints, range). Fresh and common-origin are shown as endpoints; interior schedules are grouped by terminal-block size $T(P)$, ascending. Grouped mean recovery decreases monotonically with $T(P)$ at every checkpoint, although individual schedules with equal or adjacent T overlap: a small terminal block retains nearly all of the sequential advantage regardless of stale-pair count, whereas a large stale terminal block is disproportionately damaging. This post-hoc structural diagnostic was identified from the schedule family and is reported descriptively.

Schedule	T	B	S	mean recovery	range
[1]|[2]|[3]|[4]|[5]	1	5	0.0	1.00	[1.00, 1.00]
[1, 2, 3, 4]|[5]	1	2	0.6	0.87	[0.86, 0.87]
[1, 2, 3]|[4]|[5]	1	3	0.3	0.98	[0.97, 0.98]
[1]|[2, 3, 4]|[5]	1	3	0.3	0.92	[0.91, 0.92]
[1, 2]|[3, 4]|[5]	1	3	0.2	0.97	[0.97, 0.98]
[1, 2]|[3]|[4]|[5]	1	4	0.1	1.00	[0.99, 1.00]
[1]|[2, 3]|[4]|[5]	1	4	0.1	0.99	[0.99, 1.00]
[1]|[2]|[3, 4]|[5]	1	4	0.1	0.98	[0.97, 0.98]
[1, 2, 3]|[4, 5]	2	2	0.4	0.89	[0.88, 0.90]
[1, 2]|[3]|[4, 5]	2	3	0.2	0.87	[0.86, 0.88]
[1]|[2, 3]|[4, 5]	2	3	0.2	0.87	[0.86, 0.88]
[1]|[2]|[3]|[4, 5]	2	4	0.1	0.87	[0.86, 0.89]
[1, 2]|[3, 4, 5]	3	2	0.4	0.59	[0.57, 0.60]
[1]|[2]|[3, 4, 5]	3	3	0.3	0.59	[0.58, 0.59]
[1]|[2, 3, 4, 5]	4	2	0.6	0.16	[0.09, 0.20]
[1, 2, 3, 4, 5]	5	1	1.0	0.00	[0.00, 0.00]

Table 9: External-cell validation of the terminal-block coordinate $T(P)$ (descriptive only; no new model or adaptation runs). All statistics below were fixed in the protocol before outcomes and computed from the stored records. Grouped mean recovery by $T(P)$ over the 14 interior schedules; 'mono' requires the strict ordering $\bar{R}(T{=}1) > \bar{R}(T{=}2) > \bar{R}(T{=}3) > \bar{R}(T{=}4)$; ρ is the Spearman correlation of schedule-mean CE with T; R^2 is the linear coefficient of determination of schedule-mean CE on T; the residual association is the Spearman correlation of T with the CE residual after linearly removing S (all over the same interior schedules; the discovery-family R^2 in Section 4 is computed on all sixteen schedules and is a different estimand). Recovery is normalized to Fresh (= 1) and common-origin (= 0): values above 1 outperform Fresh and values below 0 are worse than common-origin; these are reference points, not bounds. The discovery family (bottom, reference) is monotone at all four checkpoints; the grouped trend replicates in the three transferred datasets but is not uniform across the regime-changed cells.

Cell	$\bar{R}(T{=}1)$	$\bar{R}(T{=}2)$	$\bar{R}(T{=}3)$	$\bar{R}(T{=}4)$	mono	ρ	R^2	resid. assoc.
CUB-200 (transplant)	0.96	0.88	0.67	0.49	yes	+0.84	0.86	+0.84
tieredImageNet (transplant)	0.99	0.95	0.71	0.16	yes	+0.74	0.75	+0.58
FC100 (transplant)	0.97	0.84	0.47	0.10	yes	+0.92	0.93	+0.79
five-shot, main	0.74	1.09	0.71	0.19	no	+0.08	0.07	−0.34
five-shot, strong-42	0.85	0.59	0.31	0.43	no	+0.88	0.71	+0.82
five-shot, strong-43	0.82	0.86	0.67	0.62	no	+0.35	0.20	+0.04
five-shot, strong-44	0.84	0.70	0.67	0.48	yes	+0.67	0.46	+0.61
plain SGD, main	0.97	0.49	−0.12	−0.10	no	+0.92	0.92	+0.91
miniImageNet, main	0.96	0.87	0.58	0.09	yes	+0.90	0.86	+0.79
miniImageNet, strong-42	0.96	0.88	0.58	0.17	yes	+0.82	0.86	+0.79
miniImageNet, strong-43	0.96	0.89	0.60	0.20	yes	+0.82	0.86	+0.79
miniImageNet, strong-44	0.96	0.87	0.58	0.18	yes	+0.82	0.87	+0.79

Table 10: Post-hoc metric-robustness analysis: paired mean query-accuracy differences in percentage points, $[4,1]-[1,4]$ and $[3,2]-[2,3]$, on the paper's primary 500-episode bank (same bank and conventions as Table 2); 95% percentile episode-bootstrap intervals (B=10,000). No p-values, no multiplicity correction; reported regardless of sign or magnitude. Computed from stored per-episode records; no model execution.

Pair	main	strong-42	strong-43	strong-44
$[4,1]$ vs $[1,4]$	1.53 (1.25, 1.82)	1.68 (1.38, 1.99)	1.61 (1.31, 1.91)	1.50 (1.23, 1.78)
$[3,2]$ vs $[2,3]$	0.55 (0.32, 0.78)	0.74 (0.47, 1.01)	0.89 (0.60, 1.17)	0.72 (0.47, 0.96)

Table 11: Parameter-scope persistence (miniImageNet, strong-family pooled, 500 episodes per rung). Adaptable set widened from the head to the full network; the $S(P)$ slope stays positive and Holm-significant at every rung (directional plus-one $p_{\text{boot}} \leq 1.0 \times 10^{-4}$ at bootstrap resolution per rung, Holm-adjusted 3.0×10^{-4}). Pre-specified i.i.d. CIs shown alongside episode-cluster CIs (the three checkpoints share one episode bank; cluster resampling carries them together). Largest absolute slope at full-network scope; no monotonic cross-rung magnitude claim is made (calibration fallback s=0.0625 for all widened rungs).

Adaptable parameters	slope	pre-spec. 95% CI	cluster 95% CI
block 4 + head (11)	0.45	[0.42, 0.49]	[0.39, 0.52]
blocks 3–4 + head (20)	1.09	[1.02, 1.17]	[0.97, 1.23]
full network (38)	1.48	[1.38, 1.57]	[1.33, 1.63]

A.7 Parameter-Scope Extension

We evaluated the pre-specified scope ladder (head → block-4+head → blocks-3–4+head → full network) with the same sixteen-graph intervention, using the pre-specified fallback calibration scale on every widened rung (s=0.0625; the plain-SGD arm was infeasible at scope). Episode-cluster bootstrap intervals preserve the positive slope at every rung: the three strong checkpoints share one bank (per-episode cross-checkpoint slope correlation 0.86–0.94), so clustered CIs are ∼1.65× wider and remain positive throughout. Because all widened rungs use the fallback calibration, cross-rung magnitude comparisons are descriptive. The positive slope therefore does not appear to be an artifact of head-only adaptation, and the relation with $T(P)$ varies across parameter scopes.

Table 12: Beyond few-shot adaptation: TENT-style entropy-minimization adaptation with frozen normalization buffers on ImageNet-C (severity 5, standard pretrained ResNet-50; four severity-5 corruptions; 125 paired base-image slots; $\eta = 0.03$ label-free calibration; seeded deterministic execution under a frozen tolerance envelope). Primary contrast $\Delta_{\text{FC}} = \text{CE}_{\text{FRESH}} - \text{CE}_{\text{CO}}$ on the disjoint held-out batch: pooled +0.198, 95% CI [0.189, 0.207] (paired-slot cluster bootstrap, B=10^4), two-sided plus-one $p_{\text{boot}} \leq 2 \times 10^{-4}$ (bootstrap resolution); within-stream $S(P)$ slope −0.199, CI [−0.208, −0.189]. Positive Δ means fully sequential adaptation is WORSE than common-origin aggregation—the direction is reversed relative to supervised few-shot adaptation. Zero collapsed slots. Per-corruption rows are descriptive; inference is conditional on the four chosen corruptions.

Corruption	Δ_{FC}	$S(P)$ slope
Gaussian noise	+0.233	−0.233
motion blur	+0.491	−0.493
fog	+0.035	−0.035
pixelate	+0.034	−0.033
Pooled (4×125)	+0.198	−0.199

A.8 Entropy TTA on ImageNet-C

We evaluate TENT-style entropy minimization on ImageNet-C with an ImageNet-pretrained ResNet-50, adapting only the 106 BatchNorm affine parameters with normalization buffers in evaluation mode. Each stream contains five unlabeled adaptation batches followed by a disjoint held-out evaluation batch; paired base-image slots (125 evaluation, 5 calibration) are shared across the four severity-5 corruptions, and the same sixteen evaluation graphs are compared at a label-free calibrated step size (η_{conf}=0.03), with all design choices, including the corruption scope, fixed before outcomes. The evaluation run executed all $125 \times 4 \times 16 = 8{,}000$ stream×corruption×graph combinations; the primary contrast is Δ_{FC} with a paired-slot cluster bootstrap and plus-one finite-resolution p-values (direction not specified in advance), and streams were never removed based on their outcomes. The preferred direction reverses relative to the supervised regime, and inference is conditional on the four chosen corruptions.

A.9 TTA Objective/Supervision and Scope Interventions

Objective intervention. This intervention held the TTA configuration fixed (ResNet-50; 106 BN-affine tensors; frozen buffers; the four severity-5 corruptions; K=5 plus a disjoint held-out batch; the same sixteen graphs) and changed the adaptation objective and its supervision signal, entropy minimization on unlabeled adaptation batches versus supervised cross-entropy on labeled batches (oracle labels as a diagnostic adaptation signal only; no deployment claim), with the two arms' predictive displacements matched by a support-only calibration. Across 500 streams (slot-cluster bootstrap; Holm-4), the entropy arm reproduces the reversed direction ($b = -0.452$ $[-0.472, -0.432]$, $d = -0.453$ $[-0.473, -0.434]$; 71.2% base-name overlap with the prior TTA bank), whereas the supervised oracle-CE arm restores a positive graph effect ($b = +0.041$ $[+0.036, +0.047]$, $d = +0.046$ $[+0.040, +0.051]$); the interaction contrasts are $+0.493$ and $+0.499$ (Holm-adjusted $p = 4.0 \times 10^{-4}$). Changing the objective and supervision signal is thus sufficient to reverse the evaluation-graph preference within this fixed TTA configuration; the restored magnitude is roughly 8% of the supervised head-only reference (descriptive ratio).

Scope–magnitude intervention. This intervention kept the supervised objective and stream fixed and compared three arms that differ only in adapted scope and step size (Table 13): BN-affine at $\eta = 0.03$ (reference; 106 tensors; $b = +0.0400$ $[+0.0347, +0.0455]$, $d = +0.0446$ $[+0.0395, +0.0498]$), classifier head at the same η (descriptive control; $b = -0.0091$ $[-0.0094, -0.0088]$, $d = -0.0090$ $[-0.0092, -0.0087]$), and classifier head at $\eta = 0.2078$, selected before outcome inspection by support-only log-space bisection on the predictive functional displacement $\mathrm{KL}(p_{\theta_0} \,\|\, p_{\theta_F})$ (fixed bracket $[0.12, 0.36]$; matching band $[0.80, 1.25]$; the same-eta head arm moves the predictions roughly 46× less than BN, so same-η is not a fair functional comparison). The evaluation-bank matching check gave $R_{\text{formal}} = 0.929 \in [0.80, 1.25]$, supporting the displacement-controlled comparison. Primary contrasts (Holm-4 adjusted $p = 4.0 \times 10^{-4}$): $\Delta_b = +0.185$ $[+0.153, +0.219]$, $\Delta_d = +0.196$ $[+0.163, +0.230]$. The BN arm reproduces the objective-intervention direction (71.4% base-name overlap with the prior TTA bank; not an independent replication). Descriptive ratios: the matched head arm carries 5.63× (b) and 5.39× (d) the BN arm's effect and recovers 41–44% of the fixed supervised-head reference; these are effect-size summaries, not thresholded claims. Per-corruption b^{HM} (descriptive only; per-corruption values are not tested): gaussian_noise -0.055, motion_blur $+1.048$, fog -0.031, pixelate -0.060 (the d pattern matches); the pooled level is the only confirmatory level.

Table 13: Scope–magnitude intervention inside the fixed TTA regime. Arms share the stream, model, objective, and sixteen graphs, differing only in adapted scope and step size; b = stale-pair slope, $d = \mathrm{CE}_{\mathrm{CO}} - \mathrm{CE}_{\textsc{fresh}}$ (positive d favors FRESH); 95% slot-cluster CIs, 125 paired slots; Holm-4 adjusted $p = 4.0 \times 10^{-4}$ for all four primary tests; matching ratio $0.929 \in [0.80, 1.25]$.

Arm	Scope	η	Matching	b	d
B	BN-affine (106 tensors)	0.03	reference	$+0.040\,[+0.035, +0.046]$	$+0.045\,[+0.040, +0.050]$
H0	head (2 tensors)	0.03	same-eta (descr.)	$-0.009\,[-0.009, -0.009]$	$-0.009\,[-0.009, -0.009]$
HM	head (2 tensors)	0.2078	function-matched	$+0.226\,[+0.194, +0.257]$	$+0.241\,[+0.210, +0.274]$
Contrasts				$\Delta_b{=}+0.185\,[+0.153, +0.219]$	$\Delta_d{=}+0.196\,[+0.163, +0.230]$

Table 14: CLIP refresh slopes ($\times 10^{-3}$ CE) across the step-size sweep. The k-multiplier scales the transplanted operator's α; $k^*{=}34$ matches the supervised median first-step support-side displacement. All slopes are negative at every scale, encoder, and dataset: the supervised positive direction never recovers. Clipping is never activated on CLIP features (rate 0 at all scales).

Encoder	Dataset	k=0.5	k=1	k=2	k=4	k=8	k=17	k=34
ViT-B/16	DTD	−0.1	−0.5	−1.6	−4.8	−12.6	−319	−5539
ViT-B/16	Flowers102	−0.2	−0.6	−2.3	−8.0	−25.3	−417	−6482
ViT-B/16	FGVC-Aircraft	−0.2	−0.6	−1.8	−4.8	−6.6	−348	−5476
ViT-B/16	EuroSAT	−0.1	−0.3	−0.9	−2.9	−7.3	−496	−6922
RN50	DTD	−0.2	−0.6	−1.9	−5.6	−13.5	−305	−5118
RN50	Flowers102	−0.2	−0.6	−2.1	−7.3	−22.1	−415	−6268
RN50	FGVC-Aircraft	−0.2	−0.5	−1.5	−3.5	−1.9	−321	−5467
RN50	EuroSAT	−0.1	−0.2	−0.8	−2.5	−6.5	−479	−7031

A.10 Foundation-Backbone Adaptation (CLIP LoRA)

Design: CLIP ViT-B/16 (official weights; official implementation), LoRA(Q,V) rank 8 with zero-initialized B, adapted sets exactly $\{6, 18\}$ tensors in total, comprising the selected LoRA tensors and the head weight and bias, LayerNorm and all other tensors unchanged; the zero-LoRA bridge reproduced the official feature cache at relative $L^2 \leq 7.7 \times 10^{-6}$, and independent Fresh/CO recomputations matched the block-parallel composition exactly. Calibration: query-blind, support-supervised, on the dedicated DTD calibration bank's first 50 episodes (class- and image-disjoint from all evaluation episodes); η=0.03 at the last-block rung and η=0.001 at the last-four-block rung; the all-block rung was omitted by a pre-specified compute rule (22.4 h > 8 h). Evaluation: both rungs × four datasets × 500 episodes × 16 graphs, equal-dataset pooled bootstrap, two Holm families; graph sensitivity with the Fresh direction at both rungs (Table 15). The frozen-feature reference is the pre-adaptation baseline; cross-rung magnitude comparisons are descriptive.

Second-backbone replication (ViT-L/14). OpenAI CLIP ViT-L/14 (official checkpoint) with the identical last-block Q, V rank-8 LoRA scope plus the prototype head (six adapted tensors): query-blind calibration on the dedicated DTD calibration bank's first 50 episodes (largest-in-band rule) selected η=0.003 (support-CE reduction 0.435, inside the $[0.05, 0.60]$ band), transplanted to all datasets and graphs; a new 500-episode × 4-dataset bank (DTD/Flowers102/FGVC-Aircraft/EuroSAT) ran all sixteen graphs under the same composition , with Fresh equal to sequential execution and common-origin equal to parallel-sum recomposition on every episode. Pooled equally over datasets: $\Delta_{\mathrm{FC}} = -0.1803$ $[-0.2211, -0.1404]$ and stale-pair slope $+0.1642$ $[+0.1253, +0.2039]$ (plus-one $p_{\mathrm{boot}} \leq 2 \times 10^{-4}$ at bootstrap resolution each, $B{=}10^4$); per-dataset means $+0.064/-0.079/-0.708/+0.003$ (DTD/Flowers/FGVC-Aircraft/EuroSAT): FRESH preferred in two of four, the pooled effect concentrated in FGVC-Aircraft.

Table 15: Foundation-backbone adaptation : LoRA(Q,V) inside the CLIP ViT-B/16 visual tower under the supervised support objective; four datasets, 500 episodes each, sixteen evaluation graphs, per-rung query-blind step size, equal-dataset pooled bootstrap (direction not pre-specified). Negative Δ_{FC} = Fresh (sequential) lower held-out CE than common-origin. Both rungs Holm-significant ($p = 4.0 \times 10^{-4}$, primary and slope families). The last-block rung is direction-uniform across 4/4 datasets; the last-four-block rung is clear on 3/4 and approximately null on DTD. The all-block LoRA rung was omitted by the fixed pre-outcome compute rule; the frozen-feature reference is the pre-adaptation baseline. Per-dataset rows descriptive.

Rung	Δ_{FC}	95% CI	$S(P)$ slope
Last-block LoRA (pooled)	−4.586	[−4.715, −4.462]	+4.760
DTD	−6.007	descriptive	+6.018
Flowers102	−3.230	descriptive	+3.206
FGVC-Aircraft	−4.990	descriptive	+4.766
EuroSAT	−4.117	descriptive	+5.050
Last-four-block LoRA (pooled)	−0.549	[−0.582, −0.516]	+0.495
DTD	−0.002	descriptive	−0.004
Flowers102	−0.206	descriptive	+0.166
FGVC-Aircraft	−0.825	descriptive	+0.753
EuroSAT	−1.162	descriptive	+1.067

Table 16: Longer-horizon replication : eight frozen five-shot support groups per class; the complete 2^7=128 contiguous evaluation graphs (eighteen distinct S_8 levels); 300 shared episodes; three strong checkpoints; supervised operator unchanged. Co-primary schedule-weighted and 18-level-balanced slopes (direction pre-specified positive); Holm across the six tests, $p = 0.0012$ throughout. Replication framing only: no causal horizon-magnitude claim (the support union changes with K).

Checkpoint	b_{sched}	95% CI	b_{level}	95% CI
strong-42	+1.509	[1.382, 1.641]	+1.767	[1.601, 1.938]
strong-43	+1.518	[1.395, 1.645]	+1.763	[1.606, 1.927]
strong-44	+1.321	[1.204, 1.441]	+1.555	[1.400, 1.708]

A.11 Longer-Horizon Replication (K=8)

Design: eight frozen five-shot support groups per class, prototype head from S_1, the complete 2^7=128 contiguous graphs with exactly eighteen distinct S_8 levels, 300 shared episodes, three strong checkpoints, the supervised operator unchanged. Co-primary schedule-weighted and 18-level-balanced slopes are Holm-significant across six tests (Table 16); no cross-K causal magnitude comparison is made.

A-III. Selection and Decision-Level Evidence

A.12 Supervised Selection: Support-Side Jackknife (PJGS)

Selector definition. PJGS selects the candidate graph that minimizes the mean paired jackknife validation-CE difference relative to the latest-protected anchor (TPR), the terminal-protected partition $P_{\mathrm{TPR}}(K, B) = [K-B+1, 1, \ldots, 1]$; the anchor's own difference is fixed at exactly zero and exact ties are awarded to the anchor. The selector has no tunable scientific hyperparameters and no query access at selection. The selectable family has 10 candidates ($B \in \{2, 3\}$) at K=5 and 119 candidates ($B \in \{2, \ldots, 6\}$) at K=8. We also evaluate a bootstrap-U95 variant (JCFGS) as a fixed secondary baseline; it performs worse than the ungated selector.

Validation protocol. Each fold leaves out one support image per class, giving five held-out validation images that play no role in that fold's adaptation, and every support image validates exactly once; deletion from the first support group initializes the prototype head

Table 17: Support-only selection gains on held-out query cross-entropy: mean paired CE(anchor) − CE(PJGS) per episode (positive favors the selector). miniImageNet rows range over the four checkpoints (every cell $p_{\text{boot}} \leq 2 \times 10^{-4}$ at bootstrap resolution, plus-one two-sided bootstrap); dataset rows run one transplant unit under the frozen transplant operator with a single global Holm family across the three datasets (CUB and FC100 Holm 0.0012; tieredImageNet raw $p = 0.144$, not separable from zero). Regret stays below the pre-specified bound in every family (Appendix A.12).

Setting	vs. TPR anchor	vs. balanced	Outcome
miniImageNet, K=5 head	+0.039–+0.055	+0.029–+0.039	confirmed
miniImageNet, K=8 (128 graphs)	+0.157–+0.219	+0.151–+0.178	confirmed
Full network (38 tensors)	+0.064–+0.130	+0.053–+0.162	confirmed
CUB-200 (transplant)	+0.0234	+0.0292	confirmed
FC100 (transplant)	+0.0526	+0.0749	confirmed
tieredImageNet (transplant)	+0.0184	+0.0161	not sep.

from the four surviving images per class, and deeper deletions use the full five-shot S_1 group. After selection, candidate graphs are re-executed on the full original support set and scored only on the disjoint held-out query episode.

Primary result. On a new 300-episode bank (disjoint from all previously used episode banks) the selected graphs beat the TPR anchor at all four checkpoints: +0.0394 [+0.0210, +0.0580] (main), +0.0474 [+0.0288, +0.0659], +0.0552 [+0.0351, +0.0763], +0.0483 [+0.0327, +0.0644] (strong 42/43/44), all $p_{\text{boot}} \leq 2 \times 10^{-4}$ (bootstrap resolution), strong-family Holm 0.0012; versus the balanced schedule +0.0289 to +0.0385 (Holm 0.0012). Median normalized regret against the per-episode oracle-best graph is 0.114 (within the pre-specified 0.15 bound); normalized regret stayed below 0.15 in all 2×4 checkpoint×B settings.

Breadth. We evaluate the selector along three additional axes: horizon, parameter scope, and dataset transfer (Table 17). Horizon: on the complete 128-graph family (150 episodes × 4 checkpoints) the gain over the anchor spans +0.157 to +0.219 across checkpoints, and +0.151 to +0.178 over balanced, every checkpoint interval excluding zero; normalized regret stays below 0.15 in all 40 settings. Parameter scope: with the full 38-tensor backbone adapting under the pre-specified fallback calibration scale (300 episodes × 4 checkpoints), the gain over the anchor spans +0.064 to +0.130 and +0.053 to +0.162 over balanced, every checkpoint interval excluding zero; regret stays below 0.15 in all 16 settings; because the parameter-scope experiment uses a different calibrated step size, we do not compare effect magnitudes across scopes. Dataset transfer: under the transplant operator on new 200-episode banks, CUB and FC100 confirm under a single global Holm family across the three datasets (0.0012 each), while tieredImageNet is not separable from zero (raw $p = 0.144$, Holm 0.2884). Selection used support-side jackknife losses only; all reported gains are realized on disjoint held-out query episodes. Per-checkpoint estimates and intervals for all three axes are recorded in the supplementary repository manifest.

Compute. Selection computes 25 × 10 fold×graph evaluations per episode at K=5 and 40 × 119 at K=8 (plus one query evaluation per candidate graph); we report this as a selection-compute trade-off and make no latency or deployment claim. The selector does not estimate any per-episode ground-truth utility, and the mechanism by which support-side jackknife losses predict held-out graph ranking remains unidentified.

Supervised CLIP-LoRA foundation-backbone bridge. A narrow bridge test asked whether the selector transfers verbatim to the supervised CLIP environment of the foundation-adaptation experiments: CLIP ViT-B/16 official weights, LoRA(Q,V) rank 8 on the last visual block plus the prototype head (six trainable tensors), the plain-SGD operator at the fixed η=0.03, the sixteen contiguous graphs under the block-parallel composition, and the selector run unchanged (fixed-budget families: four candidates at B=2, six at B=3; budget-level contrasts averaged within episode; dataset-stratified bootstrap,

B=10^4, equal-weight dataset pooling, plus-one two-sided p, Holm–2). On a confirmation bank disjoint from all previously used banks in this environment (200 episodes each of DTD, Flowers102, FGVC-Aircraft, and EuroSAT), the development-best fixed policy is the TPR anchor at both budgets, so the comparator identity is exact by construction. On held-out queries the selector beats both fixed-budget anchors (Table 18): over TPR +0.3135 [+0.2719, +0.3577] and over balanced +1.2279 [+1.1450, +1.3108] (Holm–2 4.0×10^{-4} each), direction-positive in 4/4 datasets for TPR. The step-zero (no-gradient-adaptation) reference (+0.1188 [+0.0681, +0.1707]) shows that the selected adaptation remains worse than no adaptation in this environment, so selection is not an improved end-to-end adaptation method.

Table 18: Supervised CLIP-LoRA foundation-backbone bridge: held-out query CE contrasts on the 200-episode × {DTD, Flowers102, FGVC-Aircraft, EuroSAT} confirmation bank (equal-weight dataset pooling; dataset-stratified bootstrap, B=10^4; plus-one two-sided p). Positive values favor the PJGS selection. A_FIXED is identical to A_TPR by the fixed comparator identity; D_step0 is a reporting-only no-adaptation reference (not in the confirmatory family).

Estimand	Pooled mean	95% CI	DTD / Flow. / FGVC / Euro	Holm-2	Role
A_TPR (H1)	0.3135	[0.2719, 0.3577]	0.272 / 0.410 / 0.399 / 0.172	4.0×10^{-4}	primary
A_BAL (H2)	1.2279	[1.1450, 1.3108]	1.045 / 1.261 / 0.666 / 1.940	4.0×10^{-4}	co-primary
A_FIXED	identical to A_TPR by fixed comparator identity (max dev 0.0)			—	comparator
D_step0	0.1188	[0.0681, 0.1707]	—	—	reference

A.13 Step-Zero Utility Comparison

This is a descriptive comparison; all eight comparisons are reported without multiplicity correction (Table 19). Primary-setting values come from the factorization records (step zero = the step-0 trajectory record: zero parameter displacement, zero logit RMS, zero output KL, and endpoint CE identical to the refresh analysis), and the canonical-five-shot values were recomputed exactly from stored per-episode records. In the primary setting, sequential adaptation is beneficial relative to its step-zero prototype while common-origin aggregation is at or below it; in the canonical five-shot setting the sequential CE gain is small and accuracy is essentially unchanged.

A.14 Label-Free Target-Input-Aware Selection in Entropy TTA

In the ImageNet-C TTA environment (CLIP ViT-B/16 zero-shot head, last-block LoRA(Q,V) rank 8, entropy-minimization adaptation at η=0.03, paired slot design), graph choice can also be made label-free: the selector S^* = argmin over graphs of the KL divergence from the unadapted anchor's prediction distribution to the graph-composed prediction distribution on the stream's target inputs (label-free but target-input-aware, i.e., transductive; ties to Fresh, then smallest index). On an independent 500-stream confirmation bank, the selector prospectively improves graph choice within the adapted family: median paired CE difference $D = -9.48$ versus Fresh, all four fixed comparators beaten (each median below zero with CI upper bound below zero, Holm–4 adjusted $p = 4.0\times10^{-4}$), and no selection degeneracy (maximum selection share 0.444; Table 20). The no-adaptation (θ_0) reference (mean +4.62 [+4.50, +4.73]) shows that the unadapted model outperforms the selected adaptation. This result shows that graph selection repairs a poor TTA family substantially, yet no adaptation remains better overall—successful graph selection does not imply beneficial adaptation. This result strictly covers the selector's chosen graphs; per-graph dominance of θ_0 over all sixteen adapted graphs was not tested and is not claimed.

A.15 Prospective Evaluation of PG-JGS

Design. PG-JGS keeps the selector rule, candidate family, and fold protocol of Section A.12 exactly, and replaces candidate execution with prediction: one Fresh trajectory per distinct fixed per-graph operator family, exact finite-defect evaluation shared across

Table 19: Step-zero utility comparison (descriptive; specified before the step-zero metrics were computed; not in any confirmatory family). Per-episode means over the 500 episodes per checkpoint; contrasts are endpoint minus the step-zero prototype head θ_0 (negative CE/Brier and positive accuracy favor the endpoint), with descriptive paired episode-bootstrap 95% CIs (B=10^4).

Setting	Checkpoint	Metric	θ_0	FRESH $-\theta_0$	CO $-\theta_0$
Primary	main	CE	4.632	-0.535 $[-0.583, -0.489]$	-0.009 $[-0.093, +0.074]$
		Acc	0.700	$+0.014$ $[+0.011, +0.017]$	-0.003 $[-0.007, +0.001]$
		Brier	0.564	-0.028 $[-0.033, -0.023]$	$+0.007$ $[-0.001, +0.014]$
	strong42	CE	3.705	-0.453 $[-0.499, -0.408]$	$+0.134$ $[+0.051, +0.219]$
		Acc	0.731	$+0.011$ $[+0.009, +0.014]$	-0.013 $[-0.018, -0.009]$
		Brier	0.500	-0.025 $[-0.029, -0.020]$	$+0.027$ $[+0.019, +0.035]$
	strong43	CE	3.931	-0.480 $[-0.528, -0.433]$	$+0.109$ $[+0.023, +0.196]$
		Acc	0.728	$+0.014$ $[+0.011, +0.017]$	-0.011 $[-0.016, -0.007]$
		Brier	0.507	-0.028 $[-0.033, -0.023]$	$+0.024$ $[+0.016, +0.032]$
	strong44	CE	3.664	-0.448 $[-0.491, -0.405]$	$+0.036$ $[-0.038, +0.113]$
		Acc	0.732	$+0.012$ $[+0.009, +0.015]$	-0.010 $[-0.014, -0.006]$
		Brier	0.499	-0.024 $[-0.029, -0.020]$	$+0.019$ $[+0.012, +0.027]$
Canonical 5-shot	main	CE	4.760	-0.057 $[-0.085, -0.029]$	-0.042 $[-0.074, -0.011]$
		Acc	0.695	-0.001 $[-0.003, +0.001]$	-0.002 $[-0.004, +0.000]$
		Brier	0.572	$+0.002$ $[-0.001, +0.005]$	$+0.003$ $[-0.000, +0.006]$
	strong42	CE	3.834	-0.048 $[-0.075, -0.022]$	-0.034 $[-0.063, -0.004]$
		Acc	0.720	-0.002 $[-0.004, +0.000]$	-0.002 $[-0.004, -0.000]$
		Brier	0.520	$+0.004$ $[+0.001, +0.007]$	$+0.005$ $[+0.001, +0.008]$
	strong43	CE	4.074	-0.045 $[-0.072, -0.018]$	-0.026 $[-0.056, +0.005]$
		Acc	0.720	-0.002 $[-0.004, +0.000]$	-0.003 $[-0.005, -0.001]$
		Brier	0.524	$+0.004$ $[+0.000, +0.007]$	$+0.005$ $[+0.002, +0.009]$
	strong44	CE	3.791	-0.045 $[-0.070, -0.020]$	-0.034 $[-0.061, -0.005]$
		Acc	0.720	-0.001 $[-0.003, +0.001]$	-0.002 $[-0.004, +0.000]$
		Brier	0.520	$+0.002$ $[-0.001, +0.005]$	$+0.003$ $[-0.000, +0.007]$

Table 20: Label-free selector result (500 held-out streams; evaluated once). D = CE(S0) − CE(comparator) per stream (negative favors the selector). Plus-one two-sided $p_{\text{boot}} \leq 1.0 \times 10^{-4}$ (bootstrap resolution, B=10^4) throughout (slot-cluster bootstrap); Holm–4 across the four superiority contrasts (adjusted $p = 4.0 \times 10^{-4}$ each). The no-TTA row is a mandatory reporting-only utility reference (not in the confirmatory family); its CI excludes zero on the adverse side.

Comparator	median D	mean D	95% CI (mean)
Fresh (primary)	-9.48	-10.26	$[-10.68, -9.83]$
Superiority family (Holm–4 adjusted)			
Fixed graph selected on development data (graph 8)	-0.19	-4.26	$[-4.51, -4.03]$
Common-origin aggregation (graph 0)	-0.63	-0.95	$[-1.05, -0.86]$
argmin endpoint-distance policy	-5.61	-6.89	$[-7.35, -6.44]$
argmin total-update-norm policy	-7.26	-8.65	$[-9.15, -8.17]$
External utility reference (reporting-only)			
No adaptation (θ_0)	$+2.58$	$+4.62$	$[+4.50, +4.73]$

candidates, transport by forward-mode JVPs along the Fresh path, and the fixed PJGS support-side choice rule applied to the paired predicted score; only the selected graph is executed. The information condition of the exact selector is preserved: held-out fold images enter the fold validation CE only. The algorithm, candidate set, and choice rule were fixed before the evaluation bank was created.

Table 21: Argmin-stability certificate for PG-JGS (descriptive post-hoc; stored per-candidate exact and predicted support-validation scores, 800-episode bank). Certificate: $2e_\infty < g$ (best–second candidate gap g in the exact score; centered worst-case score error e_∞). Agreement between the predicted and exact selectors is exact in all certified episodes, so their regret is zero by construction of the executed graph; median best–second gaps (0.0022–0.0067 CE) and median doubled score errors (0.0042–0.0133 CE) are of the same order, which is why most episodes fall outside this sufficient certificate.

Checkpoint	B	certified	med g	med $2e_\infty$	agr.|uncert	reg.|uncert
main	2	0.30	0.0067	0.0133	0.601	+0.0029
main	3	0.27	0.0024	0.0052	0.692	+0.0015
strong-42	2	0.40	0.0046	0.0064	0.605	+0.0007
strong-42	3	0.33	0.0025	0.0050	0.624	+0.0005
strong-43	2	0.41	0.0047	0.0062	0.624	+0.0027
strong-43	3	0.29	0.0024	0.0053	0.649	+0.0017
strong-44	2	0.44	0.0041	0.0054	0.647	+0.0016
strong-44	3	0.29	0.0022	0.0042	0.698	+0.0008

Development feasibility. On a 400-episode development bank the predicted jackknife scores track the exact scores at per-episode Spearman 0.800–1.000 (median across checkpoints and budgets), with identical selections on 72–82% of episodes.

Held-out evaluation. On an independent 800-episode bank (disjoint from all prior banks), four arms ran in full: PG-JGS, exact PJGS, the development-best fixed graph, and no adaptation. Primary budget-level estimands average the three strong checkpoints within each episode (raw cross-entropy): regret $r = \mathrm{CE}_{\mathrm{PG}} - \mathrm{CE}_{\mathrm{exact}}$ and utility $u = \mathrm{CE}_{\mathrm{fixed}} - \mathrm{CE}_{\mathrm{PG}}$. Noninferiority margins were fixed in advance at half the development-bank strong-family gain ($\delta_{B2} = 0.0015078$, $\delta_{B3} = 0.0010729$). Utility over the strong fixed policy: +0.001891 $[+0.000470, +0.003349]$ at B=2 (one-sided plus-one $p = 0.0052$, Holm–2 0.0104) and +0.000397 $[-0.000478, +0.001281]$ at B=3 (Holm 0.185). Noninferiority was not established at either budget (mean regret +0.000980/+0.000720; bootstrap upper 95% bounds 0.001789/0.001219 exceed the margins; point-estimate retention 0.659/0.355). Selected-graph agreement 0.72–0.80; same-α pairwise ranking accuracy 0.865–0.896; the main checkpoint's cross-α candidate is over-selected (41% versus 34%). Under the same continuous output calibration the contrasts are essentially null (regret +0.00013/+0.00008; utility +0.00017/+0.00006). PG-JGS therefore does not yet match exact PJGS closely enough to replace exhaustive selection under the stated noninferiority criterion, although the predicted structure still informs the selection: PG-JGS executes none of the 250 exact candidate fold-graph compositions per episode. In the stored per-episode wall-clock records, PG-JGS selection is slower than exhaustive selection (median 1.92 s versus 1.52 s per episode); the stored timing covers the selection pass and is budget-independent, so no speed or cost claim is made.

Why rank correlation does not imply identical selection. Selection depends on the argmin over candidates, not on the global ranking. For each episode, budget, and checkpoint, let g be the gap between the best and second-best candidate in the exact support-validation score, $r_j = \hat{\mu}_j - \mu_j$ the per-candidate score errors, and $e_\infty = \frac{1}{2}(\max_j r_j - \min_j r_j)$ the centered worst-case error (the choice is invariant to a common score offset). If $2e_\infty < g$, then for every non-optimal candidate $\hat{\mu}_j - \hat{\mu}_{j^*} \geq g - 2e_\infty > 0$, so the predicted and exact selectors coincide; the certificate is exact and involves no data-selected threshold. On the stored evaluation scores the certificate certifies 27–44% of episodes per checkpoint and budget; agreement is exact within certified episodes and 0.60–0.70 outside them, and the measured regret is confined to the uncertified episodes, whose median best–second margin (0.002–0.007 CE) is of the same size as the score error (Table 21). High global rank correlation and imperfect argmin selection are thus consistent: many episodes have candidate margins too small for a predictor of this accuracy to resolve.

Table 22: PG-JGS versus exact PJGS and the strong fixed policy (strong-checkpoint-family mean raw CE; independent 800-episode bank; margins fixed in advance at half the development-bank gain). *Holm-adjusted $p = 0.0104$. Positive $\mathrm{CE}_{\mathrm{PG}} - \mathrm{CE}_{\mathrm{exact}}$ denotes regret relative to exhaustive PJGS; positive $\mathrm{CE}_{\mathrm{fixed}} - \mathrm{CE}_{\mathrm{PG}}$ means PG-JGS improves over the fixed policy.

Strong-family raw CE	B=2	B=3
$\mathrm{CE}_{\mathrm{PG}} - \mathrm{CE}_{\mathrm{exact}}$ (regret)	+0.00098	+0.00072
upper 95% bound / margin	0.00179 / 0.00151	0.00122 / 0.00107
$\mathrm{CE}_{\mathrm{fixed}} - \mathrm{CE}_{\mathrm{PG}}$	+0.00189*	+0.00040
$\mathrm{CE}_{\mathrm{fixed}} - \mathrm{CE}_{\mathrm{exact}}$	0.00287	0.00112

A-IV. Initialization, Scale, Calibration, and Composition Controls

A.16 Initialization and Step-Size Controls

A 2×2 control crossed initialization (S_1-only prototype versus full-support All25) with learning-rate policy (shared α_0 versus equal-budget graph-specific selection on the fixed bounded seven-point grid; 500 shared evaluation episodes per checkpoint, as an independent bank draw). Full per-checkpoint FRESH–CO means: cell A $-0.584/-0.636/-0.631/-0.537$, cell B $-0.129/-0.122/-0.128/-0.118$, cell C $-0.444/-0.510/-0.520/-0.449$, cell D $-0.008/-0.006/-0.010/-0.002$ CE (main/strong-42/43/44). Same-(S, B) positional dispersion D_{pos}: A 0.167–0.198, C 0.220–0.264, D 0.009–0.013 CE. In cell B, 19 of 64 nominal $2\times$ selections were clamped to the upper learning-rate bound (effective $\approx \alpha_0$); cell D selected $0.25\times$ on 51 of 64 graphs with two lower-boundary selections and no clamping.

No-adaptation baseline. The All25 no-adaptation means (corrected values; Appendix A.1) are CE 2.441/1.947/2.080/1.916 (main/strong-42/43/44) at accuracies 78.1%/80.4%/80.2%/80.4%. At every checkpoint the hindsight-best tuned fixed graph has lower mean CE than the All25 no-adaptation baseline, whereas the hindsight-best shared-α fixed graph does not; these are evaluation-only hindsight statistics, not deployment claims.

Temperature sensitivity (post-hoc descriptive). On cells A and C (fixed operator), 100 fixed evaluation episodes, the four graphs CO, $[4,1]$, $[1,4]$, and FRESH, and four checkpoints, dividing logits by $\tau \in \{0.5, 1, 2, 4\}$ leaves accuracy exactly invariant and preserves the signs of the tested FRESH–CO and $[1,4]$–$[4,1]$ contrasts at every checkpoint and temperature, while their CE magnitudes scale approximately as $1/\tau$ (cell A FRESH–CO: $-1.270/-0.634/-0.315/-0.153$ at $\tau = 0.5/1/2/4$, main checkpoint); Brier decreases with τ, consistent with overconfident logits at $\tau = 1$. No ECE is reported (no verified implementation); the tested contrasts are four graphs on two of the four control settings, not the full sixteen-graph family.

D_{pos} definition. Strata C_j group the sixteen graphs by identical stale-pair fraction $S(P)$ and block count $B(P)$ with $|C_j| \geq 2$ (five strata with sizes 4/3/3/2/2). For episode i and stratum j,

$$D_{i,j} = \frac{2}{|C_j|(|C_j|-1)} \sum_{g<h,\, g,h \in C_j} \left|\mathrm{CE}_i(g) - \mathrm{CE}_i(h)\right|, \tag{4}$$

$D_{\mathrm{pos},i} = \frac{1}{J}\sum_j D_{i,j}$ (strata equal-weighted), and D_{pos} averages $D_{\mathrm{pos},i}$ over episodes; inference is episode bootstrap. In the shared-α settings A and C, D_{pos} measures fixed-operator positional graph heterogeneity; in the tuned settings B and D the learning rate varies across graphs, so it instead measures residual heterogeneity within the equal-budget tuned family rather than a pure positional intervention. Table 23 reports the primary CE and positional estimands; Table 24 reports the Fresh–CO accuracy/Brier contrasts and paired no-adaptation comparisons, all with percentile 95% episode-bootstrap CIs where applicable. In cell D, the $[4,1]$ and $[1,4]$ selections coincided in effective learning rate at three of four checkpoints (strong-42/43/44, all at $\alpha \approx 0.125$); at those checkpoints the $[1,4]$–$[4,1]$

Table 23: Initialization × step-size control: key estimands per checkpoint (Fresh−CO; independent 500-episode bank draw; percentile 95% episode-bootstrap CIs). D_{pos}: same-(S, B) dispersion (A/C fixed-operator positional heterogeneity; B/D tuned-policy heterogeneity within the equal-budget family).

Cell	Ckpt	Fresh−CO CE	D_{pos}	$[1,4]-[4,1]$ CE
A	main	-0.584 [-0.654, -0.515]	0.193 [0.179, 0.207]	0.418 [0.367, 0.471]
A	s42	-0.636 [-0.707, -0.567]	0.196 [0.182, 0.210]	0.444 [0.392, 0.499]
A	s43	-0.631 [-0.700, -0.564]	0.198 [0.185, 0.212]	0.439 [0.385, 0.494]
A	s44	-0.537 [-0.598, -0.478]	0.167 [0.155, 0.179]	0.362 [0.319, 0.409]
B	main	-0.129 [-0.153, -0.105]	0.123 [0.116, 0.130]	0.048 [0.022, 0.075]
B	s42	-0.122 [-0.147, -0.099]	0.054 [0.051, 0.057]	0.029 [0.019, 0.039]
B	s43	-0.128 [-0.152, -0.105]	0.107 [0.102, 0.113]	0.028 [0.004, 0.051]
B	s44	-0.118 [-0.140, -0.096]	0.096 [0.091, 0.101]	0.027 [0.007, 0.048]
C	main	-0.444 [-0.505, -0.388]	0.259 [0.240, 0.280]	0.591 [0.511, 0.673]
C	s42	-0.510 [-0.572, -0.452]	0.258 [0.238, 0.279]	0.616 [0.540, 0.695]
C	s43	-0.520 [-0.585, -0.461]	0.264 [0.244, 0.284]	0.645 [0.568, 0.726]
C	s44	-0.449 [-0.507, -0.396]	0.220 [0.203, 0.237]	0.500 [0.431, 0.568]
D	main	-0.008 [-0.011, -0.005]	0.013 [0.012, 0.013]	0.006 [0.002, 0.011]
D	s42	-0.006 [-0.014, 0.000]	0.010 [0.010, 0.011]	0.005 [0.002, 0.007]
D	s43	-0.010 [-0.018, -0.003]	0.011 [0.010, 0.012]	0.004 [0.001, 0.007]
D	s44	-0.002 [-0.009, 0.004]	0.009 [0.009, 0.010]	0.003 [0.000, 0.006]

contrast remains positive and small (+0.005/+0.004/+0.003 CE, descriptive), a near-pure positional comparison within the tuned family.

Secondary metrics track the large Fresh–CO CE contrast in the shared-α cells A and C, are much smaller after graph-specific selection in B, and are essentially neutral or mixed in the jointly controlled cell D (Table 24). Paired comparisons against the same-initialization corrected no-adaptation baseline appear in the same table (hindsight, evaluation-only).

Step-size calibration details. Per-checkpoint base rates $\alpha_0(c)$: main 0.498128, strong-42 0.499766, strong-43 0.500000, strong-44 0.473337 (β=0.01, clip 1.0). Candidate multipliers run over the fixed grid $M = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$ with $\alpha_{\text{candidate}} = m\,\alpha_0(c)$; the pre-specified learning-rate bounds $[10^{-5}, 0.5]$ clamp nominal candidates $\geq 2\times$ to a shared effective value (at strong-43 the base rate itself equals the upper bound), while each candidate still consumes its own development evaluation, so the comparison remains equal-budget. Development used 200 episodes and the evaluation run 500 new episodes with disjoint episode identities; all 128 α^* values were fixed before the evaluation bank was generated, and grid collapse arithmetic, tie handling, and seed domains are recorded in the supplementary repository manifest. The full sixteen-graph family per cell is shown in Figure 3.

Supplementary repository. The supplementary repository contains the pre-specified experiment specifications, episode-bank manifests, execution code, and per-run statistics used for the analyses reported here; the repository README maps each appendix experiment to its implementation and stored records. No external registry was used; stored content hashes support record consistency rather than temporal priority.

Table 24: Initialization × step-size control: secondary metrics per checkpoint; percentile 95% episode-bootstrap CIs. **Panel A**: FRESH−CO accuracy (pp) and Brier contrasts. **Panel B**: paired FRESH–no-adaptation / CO–no-adaptation CE against the same-initialization corrected no-adaptation baseline (reconstructed records).

Panel A: secondary endpoint contrasts			
Cell	Ckpt	ΔAcc (pp)	ΔBrier
A	main	+1.8 [+1.5, +2.2]	-0.040 [-0.047, -0.034]
A	s42	+2.1 [+1.7, +2.4]	-0.043 [-0.049, -0.036]
A	s43	+1.9 [+1.6, +2.3]	-0.039 [-0.046, -0.033]
A	s44	+1.8 [+1.5, +2.1]	-0.040 [-0.046, -0.034]
B	main	+0.1 [-0.1, +0.3]	-0.006 [-0.009, -0.002]
B	s42	+0.4 [+0.2, +0.7]	-0.007 [-0.011, -0.003]
B	s43	+0.4 [+0.2, +0.6]	-0.007 [-0.010, -0.003]
B	s44	+0.4 [+0.1, +0.6]	-0.008 [-0.011, -0.005]
C	main	+2.3 [+1.9, +2.7]	-0.045 [-0.052, -0.038]
C	s42	+2.8 [+2.4, +3.2]	-0.059 [-0.066, -0.051]
C	s43	+2.8 [+2.4, +3.2]	-0.055 [-0.063, -0.048]
C	s44	+2.6 [+2.2, +2.9]	-0.052 [-0.059, -0.045]
D	main	+0.0 [-0.1, +0.1]	+0.000 [-0.000, +0.001]
D	s42	-0.1 [-0.2, +0.1]	+0.001 [-0.001, +0.002]
D	s43	-0.0 [-0.2, +0.1]	+0.000 [-0.001, +0.002]
D	s44	-0.1 [-0.2, +0.0]	+0.001 [-0.000, +0.002]
Panel B: utility relative to no adaptation			
Cell	Ckpt	FRESH−NoAdapt CE	CO−NoAdapt CE
A	main	-0.484 [-0.527, -0.440]	+0.100 [+0.017, +0.185]
A	s42	-0.445 [-0.486, -0.405]	+0.191 [+0.114, +0.272]
A	s43	-0.464 [-0.507, -0.422]	+0.167 [+0.087, +0.248]
A	s44	-0.439 [-0.478, -0.401]	+0.098 [+0.027, +0.170]
B	main	-0.484 [-0.528, -0.440]	-0.355 [-0.393, -0.318]
B	s42	-0.445 [-0.486, -0.405]	-0.322 [-0.359, -0.286]
B	s43	-0.464 [-0.507, -0.422]	-0.336 [-0.375, -0.299]
B	s44	-0.440 [-0.480, -0.401]	-0.322 [-0.357, -0.289]
C	main	+0.034 [+0.009, +0.059]	+0.478 [+0.417, +0.544]
C	s42	+0.013 [-0.007, +0.034]	+0.523 [+0.462, +0.589]
C	s43	+0.005 [-0.015, +0.026]	+0.525 [+0.461, +0.594]
C	s44	+0.014 [-0.004, +0.033]	+0.463 [+0.407, +0.525]
D	main	-0.030 [-0.038, -0.022]	-0.022 [-0.027, -0.017]
D	s42	-0.032 [-0.044, -0.019]	-0.025 [-0.035, -0.015]
D	s43	-0.037 [-0.050, -0.024]	-0.027 [-0.037, -0.016]
D	s44	-0.026 [-0.038, -0.013]	-0.023 [-0.033, -0.013]

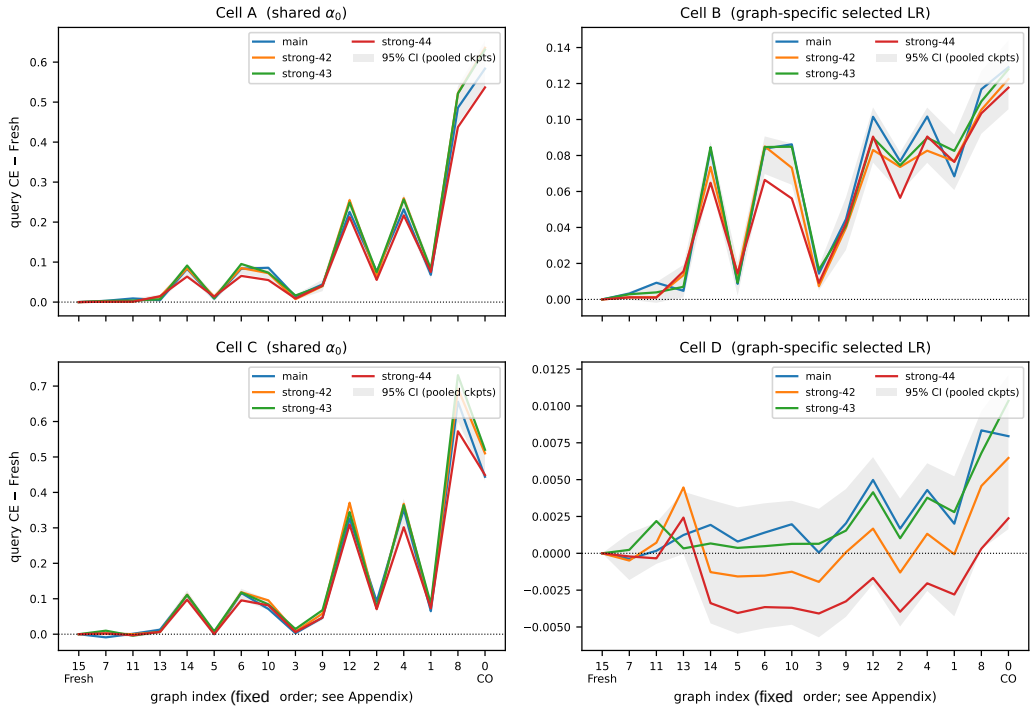


Figure 3: Initialization × step-size control, full sixteen-graph result: evaluation query CE relative to FRESH per graph in the fixed display order (leftmost point: graph 15, FRESH; rightmost point: graph 0, CO; 500 shared episodes per checkpoint; four checkpoint curves; shaded band = pooled-checkpoint 95% episode-bootstrap CI). Cells B and D use graph-specific selected learning rates over the fixed bounded grid, so their within-panel differences mix position and step-size policy; cells A and C are fixed-operator comparisons. Graph indices 0–15 are the block-size signatures of the sixteen contiguous partitions in the fixed display order: 0=(5), 1=(4,1), 2=(3,2), 3=(3,1,1), 4=(2,3), 5=(2,2,1), 6=(2,1,2), 7=(2,1,1,1), 8=(1,4), 9=(1,3,1), 10=(1,2,2), 11=(1,2,1,1), 12=(1,1,3), 13=(1,1,2,1), 14=(1,1,1,2), 15=(1,1,1,1,1).

A.17 Canonical Support Scale, Initialization Reuse, and a Non-Visual Boundary

Design and calibration. Seven cells crossed a per-update-group support budget $m \in \{1, 2, 5\}$ (totals 5/10/25 images per class; no total-one-shot condition, which cannot form five distinct real update groups without duplication) with two initialization policies (S_1 = first update group; FULL = the whole adaptation union) and one disjoint-initialization arm (DISJ: a separate 5-shot-per-class initialization set with zero overlap with the 25-image-per-class adaptation support and the query). Episodes were shared across cells within each split (same classes and queries; support nested in m); banks were generated with independent reproducible seeds; class/image identities and bank hashes are recorded in the supplementary repository manifest. Each cell–checkpoint used one graph-independent, query-blind step size chosen before outcomes by nearest-grid median first-step parameter displacement targeting of the fixed reference target (the $(m{=}5, S_1, \times 1)$ cell) on the fixed multiplier grid $\{\frac{1}{8}, \ldots, 8\}$; the operator's learning-rate bounds were never bypassed. All sixteen graphs ran on every episode; Fresh and CO endpoints were re-verified against an independent structural recomposition (worst max-abs 0.0 over all 8,400 episodes × checkpoints). Table 25 reports the main-checkpoint contrasts with slopes, dispersion, and step-zero references.

Results. Table 25 reports the main checkpoint; the strong checkpoints replicate direction at every cell (pooled strong: $m{=}1$ S_1 $+0.155$ $[+0.111, +0.199]$; DISJ -0.817 $[-0.920, -0.721]$; both Holm-adjusted 4.0×10^{-4} within the pre-specified two-test family). The S_1 magnitude summary across m is non-monotone (0.150/0.067/0.519 for totals 5/10/25) and the full-init summary monotone (0.021/0.203/0.486), both exactly as pre-specified in the scaling estimand; D_{pos} grows with support in both families. Ratios of graph-effect magnitude to adaptation-utility magnitude are reported only when $|\text{FRESH} - \text{step zero}| \geq 0.02$ CE. Figure 4 shows the full family.

Calibration disclosure. Table 26 reports, per cell over the four checkpoints, the selected multiplier on the fixed grid, the achieved-to-target median first-step displacement ratio, and the rows limited by the learning-rate upper bound. The fixed rule selected the nearest candidate on the discrete grid and specified no match tolerance: the S_1 $m{=}1$ family remains 26.6–28.8% from target and the S_1 $m{=}2$ family 8.7–9.8%, while every remaining cell is within 4.6%. At strong-43 the base rate itself equals the operator upper bound ($\alpha_0{=}0.500000$) in every cell—a property of the operator constants, not a calibration deviation.

Support-scale summary. Figure 5 plots FRESH − CO held-out CE across per-class totals 5/10/25 for first-group (S_1) and full-union initialization, plus the disjoint-initialization arm (all points are Table 25 estimates). Graph sensitivity is present at every budget in every cell; in the S_1 family the endpoint preference reverses across support-scale cells conditional on the fixed calibration policy of the least-targeted cell ($m{=}1$), while under full-union initialization FRESH is preferred throughout, and the disjoint arm retains the largest observed contrast, so initialization/adaptation data reuse is not required.

A.18 Continuous Output Calibration

Selection after joint initialization-and-step-size calibration. A follow-up asks whether per-episode selection retains value once the two strongest levers are already controlled: the head starts from the full-support (All25) prototype and each candidate graph uses its own learning rate α_g^* selected on the independent development bank of the initialization–tuning control; the selector rule is unchanged, fold initialization uses all 24 non-held-out support images per class, and selection ran on a new independent 800-episode bank at the four checkpoints. All CE comparisons, including the non-separable ones, are reported in Table 27. Descriptive secondary metrics from the same raw arrays show a metric split: accuracy point estimates (percentage points) agree in sign with the CE contrasts on the strong checkpoints ($+0.4$ to $+0.7$ pp against both comparators) and are negative on the main checkpoint (-1.7 and -1.6 pp), while Brier point estimates reverse on the strong checkpoints (-0.007 to -0.015) and agree on the main checkpoint ($+0.033$ to $+0.034$); every secondary contrast's descriptive 95% CI includes zero, so the CE ordering does not carry

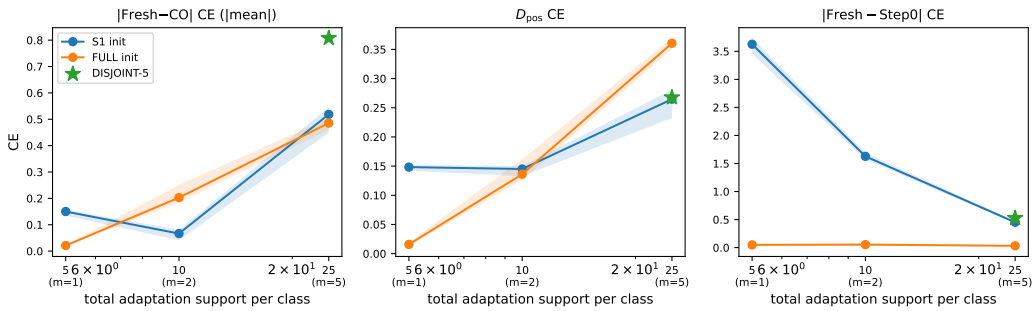


Figure 4: Effect magnitudes vs total adaptation support per class (log scale): $|\textsc{fresh} - \textsc{co}|$, D_{pos}, and $|\textsc{fresh} - \text{step zero}|$ (pooled-checkpoint means; band = per-checkpoint envelope), with the disjoint-initialization arm marked at m=5; note the panel-specific vertical scales.

Table 25: Support-scale and initialization moderation (main checkpoint; all four checkpoints in Appendix A.17). Per-update-group budget m (totals 5/10/25 images per class); S_1 vs full-union initialization; one query-blind, graph-independent displacement-targeted step size per cell (fixed before outcomes); all sixteen graphs; 300 shared episodes per cell and checkpoint. Δ: FRESH−CO CE; percentile 95% episode-bootstrap CIs. DISJ: disjoint 5-shot initialization set, no reuse in adaptation.

Cell	m	Δ CE	slope	D_{pos}	θ_0 CE
S_1	1	$+0.135$ $[+0.086, +0.184]$	-0.144	0.142	20.87
S_1	2	-0.069 $[-0.116, -0.023]$	$+0.066$	0.149	9.87
S_1	5	-0.538 $[-0.622, -0.459]$	$+0.535$	0.271	4.64
Full	1	-0.023 $[-0.042, -0.008]$	$+0.023$	0.017	4.64
Full	2	-0.251 $[-0.312, -0.193]$	$+0.256$	0.162	3.30
Full	5	-0.460 $[-0.533, -0.391]$	$+0.532$	0.350	2.48
DISJ	5	-0.779 $[-0.881, -0.682]$	$+0.731$	0.277	4.72

over to the secondary metrics. The eight fixed comparators in Table 27 are, at B=2/B=3: $[3,2]/[2,2,1]$ (main), $[2,3]/[1,3,1]$ (strong-42), $[2,3]/[2,2,1]$ (strong-43), and $[1,4]/[1,3,1]$ (strong-44), all at effective learning rate 0.118–0.125 after the bound clamp.

Fit. The inverse temperature $s \geq 0$ minimizes the convex objective $L(s) = \frac{1}{N}\sum_i[-s\, z_{i,y_i} + \log\sum_c e^{s\, z_{i,c}}]$ on the query cross-entropy of a 400-episode continuous-calibration development bank, fit by a deterministic Newton–bisection solver to $|L'(s)| \leq 10^{-12}$ (boundary rule $s^* = 0$ if $L'(0) \geq 0$); there is one fit per (checkpoint, budget, policy). All 24 fits found interior optima: $s^* \in [0.117, 0.139]$ ($\tau^* \in [7.2, 8.5]$), residuals $\leq 3.6 \times 10^{-14}$, and no refit occurred on evaluation data.

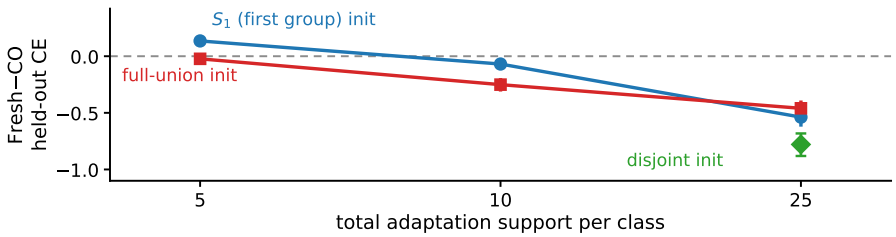


Figure 5: Support scale and initialization jointly moderate the preferred endpoint: FRESH−CO held-out CE (mean ± 95% CIs) across per-class totals 5/10/25 for first-group (S_1) and full-union initialization, plus a disjoint-initialization arm; all points are Table 25 estimates (300 shared episodes per cell).

Table 26: Step-size calibration disclosure (one query-blind, graph-independent, support-only step size per cell–checkpoint; multipliers in checkpoint order main, strong-42/43/44 on the fixed grid $\{\frac{1}{8}, \ldots, 8\}$). Ratio = achieved-to-target median first-step displacement; the fixed rule selected the nearest grid candidate and specified no match tolerance. In bound-limited rows, the selected 2× multiplier gives a nominal $\alpha > 0.5$, so the effective rate is clamped to 0.5; at strong-43 the base rate itself equals that bound (α_0=0.500000) in every cell, a checkpoint property of the operator constants. These rows summarize the calibration records, not additional experiments.

Cell	Multipliers	Ratio range	Max mismatch	Bound-limited
S_1, m=1	1/1/1/1	1.266–1.288	28.8%	0
S_1, m=2	1/1/1/1	1.087–1.098	9.8%	0
S_1, m=5	1/1/1/1	1.000	0% (target)	0
Full, m=1	1/1/1/1	0.993–1.002	0.75%	0
Full, m=2	2/2/1/2	0.964–1.023	3.62%	3
Full, m=5	2/2/1/2	0.954–1.016	4.57%	3
DISJ, m=5	1/1/1/1	1.009–1.029	2.87%	0

Table 27: Calibrated-utility follow-up (Section 8): held-out CE gains of the selected graph over the best fixed tuned graph (H1) and over the All25 no-adaptation reference (H2), per checkpoint and refresh budget. Orientation: CE(comparator) − CE(PJGS); positive favors the selector. n=800 shared episodes from a new untouched bank; paired episode bootstrap (B=10,000). Holm correction is applied separately to the six strong-checkpoint H1 tests and the six strong-checkpoint H2 tests; the main-checkpoint rows are descriptive only (^{d}, not in any family). Ordinary 95% CIs are not multiplicity-adjusted, so a CI may exclude zero while the Holm-adjusted comparison is not significant (strong-43, B=2, H1). All eight comparisons are reported; the two strong-43 H1 tests and both main-checkpoint H1 tests are not separable from zero.

		vs. best fixed tuned			vs. no adaptation		
Checkpoint	Budget	gain	95% CI	p	gain	95% CI	p
main	B=2	+0.0004	$[-0.0017, +0.0024]$	0.741^d	+0.023	$[+0.017, +0.028]$	$\leq 2 \times 10^{-4\,d}$
main	B=3	+0.0003	$[-0.0011, +0.0016]$	0.674^d	+0.025	$[+0.018, +0.031]$	$\leq 2 \times 10^{-4\,d}$
strong-42	B=2	+0.0032	$[+0.0014, +0.0050]$	0.0020	+0.022	$[+0.016, +0.029]$	0.0012
strong-42	B=3	+0.0026	$[+0.0011, +0.0042]$	0.0024	+0.023	$[+0.017, +0.030]$	0.0012
strong-43	B=2	+0.0021	$[+0.0002, +0.0040]$	0.0628	+0.023	$[+0.016, +0.030]$	0.0012
strong-43	B=3	+0.0008	$[-0.0007, +0.0021]$	0.2932	+0.025	$[+0.018, +0.031]$	0.0012
strong-44	B=2	+0.0061	$[+0.0042, +0.0082]$	0.0012	+0.019	$[+0.013, +0.025]$	0.0012
strong-44	B=3	+0.0021	$[+0.0007, +0.0034]$	0.0084	+0.020	$[+0.014, +0.026]$	0.0012

Raw versus calibrated results. Four experiments measure selection utility, each on its own episode bank and step-size regime; Table 29 maps them, and their magnitudes should not be compared directly. On the raw-logit selection bank, PJGS beats the development-best fixed graph in four of six strong-checkpoint×budget tests (+0.002 to +0.006 CE) and no adaptation in all six (+0.019 to +0.025 CE; Table 27). Under the grid-τ calibration (τ^*=4) the B=2 fixed-policy advantage persists at all three strong checkpoints (+0.0005/+0.0009/+0.0010 CE; Holm 0.0012–0.017). Under the continuous calibration of this section it survives at two of three strong checkpoints (+0.00023 and +0.00046 CE; Holm 0.041 and 0.0012), the B=3 advantage is null, and the advantage over no adaptation, though point-positive in every cell (+0.0006 to +0.0009 CE), is no longer separable (minimum Holm 0.073). The PG-JGS calibrated contrasts transplant this fixed s^* table to all four arms.

Structural contrasts under output calibration. A separate post-hoc robustness analysis asks how much of the raw-CE magnitude of the paper's structural contrasts is output-scale amplification. Temperatures are fitted on the disjoint development bank, pooled over the six structural arms (S1 head, native per-checkpoint operator, natural order; one shared s per checkpoint, $\tau \approx 10.1$–11.8), then applied to the FRESH–CO, $[4,1]$–$[1,4]$, and $[3,2]$–$[2,3]$ contrasts on an independent 500-episode evaluation bank (shared with the matched-delay

Table 28: Post-hoc output calibration of the three structural contrasts (S1 head, native per-checkpoint operator, natural order; one shared s per checkpoint, fitted on the disjoint 400-episode continuous-calibration development bank; $s^* \in [0.0845, 0.0988]$; the contrasts are replayed on the pre-specified first-500-episode subset of an independent 800-episode evaluation bank—the same episodes as the matched-delay analysis of Table 3 and Appendix B.7 and as the reliability subset of Appendix A.22, distinct from the primary bank behind Tables 2, 7, and 10). Ranges over the four checkpoints of the raw and calibrated mean paired CE differences and of the removed magnitude; the calibrated residual keeps the raw sign in all twelve cells with paired-bootstrap CIs excluding zero.

Contrast	raw CE	calibrated CE	magnitude removed	residual |CE|
FRESH−CO	−0.570–−0.488	−0.044–−0.036	92.1%–93.7%	0.036–0.044
[4, 1]–[1, 4]	−0.498–−0.412	−0.037–−0.027	92.4%–93.8%	0.027–0.037
[3, 2]–[2, 3]	−0.200–−0.175	−0.015–−0.012	92.7%–94.0%	0.012–0.015

Table 29: Map of selector-utility evaluations. The four experiments use different episode banks (each an 800-episode pre-specified evaluation bank, distinct from the banks behind Tables 2 and 24), step-size regimes, aggregations, and calibration maps; their magnitudes should not be compared directly. Bank identities and content hashes are listed in the supplementary repository manifest. All values are cross-entropy differences in which positive numbers favor selection.

Where	Bank	Arms / step-size policy	Agg. / calib.	Reported gain
Table 27	raw-utility evaluation bank (n=800)	PJGS / DEV-best fixed / no adaptation; All25 init, per-graph DEV-fixed α_g^*	per checkpoint; raw	vs fixed +0.002 to +0.006 (4/6 Holm); vs no-adapt +0.019 to +0.025 (6/6)
App. A (grid)	grid-temperature calibration evaluation bank (n=800)	same policies; output temperature τ^*=4 (grid edge)	per checkpoint; grid τ	vs fixed +0.0005 / +0.0009 / +0.0010 (Holm 0.0012–0.017)
App. A (cont.)	continuous-temperature calibration evaluation bank (n=800)	same policies; continuous $s^* \in [0.117, 0.139]$	per checkpoint; cont. s^*	vs fixed: B=2 +0.00023, +0.00046 (Holm 0.041, 0.0012); B=3 null; vs no-adapt +0.0006 to +0.0009, 0/6 (min 0.073)
Table 22	predictor-guided selection confirmation bank (n=800)	PG-JGS / exact PJGS / fixed / no adaptation; All25, graph-specific fixed α	strong-family mean; raw, cont. s^*	fixed−exact +0.00287 / +0.00112 raw; selection better than no adaptation by 0.0290 / 0.0301 raw and 0.0007 / 0.0009 calibrated CE for B=2/3

analysis of Appendix B.7, distinct from the primary bank behind Tables 2, 7, and 10); the raw structural contrasts differ modestly in magnitude across independent bank draws while preserving direction. Calibration removes 92.1–94.0% of every raw contrast magnitude while preserving all twelve signs (paired-bootstrap CIs excluding zero), leaving residuals of 0.012–0.044 CE (Table 28): the ordering is not created by output scale, but its raw-CE magnitude is dominated by it. These temperatures are estimated for the structural-contrast question and are neither the selection-utility fits above nor interchangeable with them.

A.19 First-Step Scale-Calibrated Control and a Block-Mean Alternative

The control rescales each graph's single effective step size by a multiplier m chosen on the fixed grid $\{0.125, 0.25, 0.5, 1, 2, 4, 8\}$ to target the first-commit parameter displacement of the FRESH graph (median over the 400-episode development bank; proxy linearity 4×10^{-7} to 2×10^{-6} relative; the first eight episodes executed twice with identical results at every checkpoint; paired mean CE with bootstrap CIs on the pre-specified 500-episode bank, no p-values). The selected multiplier tracks the first block size on the coarse grid ($m = 0.25$ for CO, $[4, 1]$, $[3, 2]$, and $[3, 1, 1]$; 0.5 for the four graphs whose first block has size two; exactly 1.0 for the eight graphs whose first block is a single update, including $[1, 4]$ and $[1, 1, 3]$), but the fixed grid cannot exactly represent targets such as 1/3 (size-3 first blocks)

Table 30: First-step scale-calibrated control (paired mean CE, pre-specified 500-episode bank; no p-values). Top: selected step-size multiplier m and matching ratio $|d_G - d_F|/d_F$ (ranges over the four checkpoints); the CO arm sits at the grid-resolution boundary, so the study is *scale-calibrated*, not displacement-matched. Bottom: evaluation contrasts at the grid-selected scales, with native raw ranges in Table 28; entries are contrasts of the form CE(first)−CE(second), so negative values favor the first-listed graph.

Panel A: selected multiplier m and matching ratio $|d_G - d_F|/d_F$ (range over checkpoints)

CO	$m = 0.25$; ratio 0.247–0.251 (at the pre-specified tolerance boundary)
$[3, 2]$	$m = 0.25$; ratio 0.249–0.251
$[1, 4]$	$m = 1.0$; ratio 0.000–0.000
FRESH	$m = 1.0$; ratio 0.000–0.000 (fixed at 1.0)

Panel B: evaluation contrasts at the grid-selected scales (paired mean CE [bootstrap CI])

Contrast	main	strong-42	strong-43	strong-44
FRESH−CO	−0.249 [−0.280, −0.219]	−0.206 [−0.236, −0.176]	−0.220 [−0.253, −0.189]	−0.204 [−0.232, −0.176]
$[4, 1]$–$[1, 4]$	−0.235 [−0.308, −0.163]	−0.335 [−0.406, −0.263]	−0.284 [−0.354, −0.212]	−0.248 [−0.313, −0.185]
$[3, 2]$–$[2, 3]$	+0.124 [+0.106, +0.142]	+0.109 [+0.092, +0.127]	+0.120 [+0.101, +0.138]	+0.111 [+0.095, +0.128]

or 1/5 (CO), and the CO matching ratio lands at the grid-resolution boundary (0.247–0.251, straddling the pre-specified 0.25 tolerance), so we describe the study as *scale-calibrated* rather than displacement-matched, and we use that name throughout. At the grid-selected scales the FRESH−CO contrast attenuates but remains sizable (−0.204 to −0.249 CE), and the $[4, 1]$–$[1, 4]$ comparison keeps its direction with the two arms at their own calibrated scales (−0.235 to −0.335 CE, arm multipliers 0.25 versus 1.0; this is a graph-specific scale comparison, not a fixed-operator placement test), while the $[3, 2]$–$[2, 3]$ contrast reverses sign (+0.109 to +0.124 against native −0.175 to −0.200; all CIs exclude zero; Table 30). Because the grid cannot represent the size-3 target exactly, it is not clean evidence that scale alone reverses the preference. As a descriptive alternative-composition arm, block-mean within-block accumulation at native scale leaves FRESH−CO sizable (−0.293 CE, main checkpoint) while the two matched-delay pair contrasts fall below the 0.02-CE materiality floor (−0.008 and −0.002); this is an alternative algorithm, not a read-placement control.

A-V. Language and Cross-Modality Transfer Boundaries

A.20 Language Adaptation in a De-Saturated Regime (Banking77)

Regime. Episodic intent classification over Banking77 with a frozen sentence encoder (`BAAI/bge-base-en-v1.5`; CLS pooling, L_2-normalized embeddings; tokenizer length 64) and a cosine prototype head ($w = 2\hat{P}$, $b = -\|\hat{P}\|^2$). Episodes are 5-way with 15 held-out query texts per class; the adaptation support is 25 texts per class organized as five across-class update groups (five texts per class per group), disjoint from the queries within an episode. Prototype initialization uses one shot per class (the first text of the first update group); a pre-specified pool-size feasibility rule admitted 73 of 77 intents, and episodes sample classes from the admissible pool only. The one-shot initialization is the pre-specified de-saturation rule: full-support prototype initializations on both candidate datasets left step-zero accuracy at 95.4–98.4%, and the pre-specified saturation band requires step zero to lie in $[35\%, 95\%]$—without it, almost no held-out error remains for any graph to change.

Calibration and evaluation run. The step size is calibrated graph-neutrally: a single one-step update on the first group, evaluated before any graph trajectory diverges, chosen by median relative calibration-validation CE reduction over a pre-specified six-point grid $\{0.01, \ldots, 3.0\}$; the selection is support-only and query-blind. The selected scale, α=3.0, sits at the upper edge of the pre-specified grid—broader operator calibration remains untested. The evaluation run executes all sixteen graphs on 300 new episodes under this fixed graph-independent operator; utterances are split between calibration and evaluation pools at the string level (zero overlap). An independent reimplementation reproduced every reported comparison exactly (maximum absolute CE difference 0).

Table 31: Banking77 language regime (frozen encoder, cosine prototype head, one-shot initialization, graph-neutrally calibrated head-only operator at α=3.0): step-zero, FRESH, and CO endpoints over 300 episodes × sixteen graphs. The last column reports the FRESH−CO contrast with per-episode bootstrap 95% CIs (10,000 resamples); the CE contrast and the stale-pair slope are Holm-adjusted over the six-slot family ($p = 1.2 \times 10^{-3}$ each). U_{rel} is FRESH's relative CE gain over step zero; M is the CE contrast relative to that gain. The CE ordering and the accuracy ordering disagree: the preferred endpoint is metric-dependent.

Metric	Step zero	FRESH	CO	FRESH−CO
Held-out CE	1.2652	0.9185	0.9057	$+0.01279$ $[+0.01094, +0.01458]$
Accuracy (%)	93.00	97.97	94.47	$+3.50$ $[+2.81, +4.24]$
Brier	0.6434	0.4519	0.4433	$+0.0086$
Stale-pair slope (CE)	–	–	–	-0.01608 $[-0.01750, -0.01463]$
D_{pos} (CE)	–	–	–	0.0009 [0.0008, 0.0010]
U_{rel} (%)	–	–	–	27.5 [27.1, 27.9]
M	–	–	–	0.037

Boundaries. A direct RoBERTa/CLINC150 operator transplant did not enter its pre-specified calibration range (Appendix A.21).

A.21 Non-Visual Transformer Transfer (Operator-Scaling Boundary)

Setting. `roberta-base` (124.6M parameters) with the CLINC150 train pool (150 intent classes, ≥ 100 examples each); episodic geometry 5-way, K=5 groups × 5 real texts per class per group (25 per class), 15 held-out query texts per class, no within-episode reuse, tokenizer max length 64. Episode support and query examples are disjoint but drawn from the CLINC150 training pool; they are not the official CLINC test split.

Operator and adaptation. The supervised operator was transplanted: mean-CE gradients over each 25-text update group, joint gradient clipping over all adapted tensors (rank-8 LoRA A/B on the final block's attention Q, V, B zero-initialized, plus the squared-Euclidean prototype head over the CLS representation), weight decay $\beta = 0.01$, explicit SGD updates, five updates, sixteen contiguous graphs, block-parallel-at-base composition, no optimizer state. A single graph-independent step size was selected query-blind on a dedicated 100-episode development bank by the pre-specified support-validation band.

Calibration outcome. The band $[0.05, 0.60]$ contained no grid point: median first-step CE reduction peaked at $+0.14\%$ (α=0.1) and turned negative at $\alpha \geq 0.5$. The pre-specified fallback rule therefore selected the smallest grid step, α=0.005. The fixed operator thus admits no admissible calibration scale on this parameterization: joint clipping normalizes updates to a magnitude negligible relative to the prototype head's weight scale.

Result and interpretation. At the fallback scale (300 new episodes; Fresh/CO replay worst max-abs 0.0): FRESH−CO held-out CE $= -4.19 \times 10^{-5}$ $[-5.88 \times 10^{-5}, -2.58 \times 10^{-5}]$; stale-pair slope $+4.05 \times 10^{-5}$ $[+2.43 \times 10^{-5}, +5.74 \times 10^{-5}]$; D_{pos} 1.0×10^{-5}; graph CE range 9.6×10^{-5}; FRESH−step zero -0.0003. Both primary two-sided tests are Holm-significant (4.0×10^{-4}), and the direction matches the primary supervised setting (Fresh preferred, slope consistent), but the practical magnitude is four orders below the primary supervised effect and the adaptation itself moved held-out CE by 0.03%. We interpret this result as an operator-scaling boundary rather than evidence for practically meaningful language-side transfer: under a verbatim transplant of the supervised operator, no meaningfully calibrated adaptation regime exists on this transfer, and cross-modality generality of practically scaled graph effects remains open.

A-VI. Predictor Reliability

A.22 Support-Side Reliability Diagnostics

We pre-specified exactly three candidate diagnostics. For episode e and support order π of the [3, 2] task, with the Fresh path x_t, defects $\tilde{\eta}_t^G$, and transport Φ_F as in Section 6, the diagnostics are $\kappa_0 = \|y_K''(0)\|/(\|y_K'(0)\| + \varepsilon)$ (the ratio of the norm of the second intervention response to that of the first intervention response at λ=0, the Fresh-read end of the intervention path, with $\varepsilon = 10^{-12}$; realized by exact nested forward-mode differentiation in the supervised arm and by a frozen forward second-difference stencil validated against exact differentiation to $\leq 1.6 \times 10^{-3}$ relative in the CLIP-LoRA arm), $A_G = \|\sum_t \Phi_F(K{:}t)\tilde{\eta}_t^G\|/(\sum_t \|\tilde{\eta}_t^G\| + \varepsilon)$, and $D_G = \sum_t \|\tilde{\eta}_t^G\|/(\|x_K - x_0\| + \varepsilon)$. Every norm is the global Euclidean norm of the concatenated adapted-parameter vector in the engine's fixed canonical tensor order, used identically in both arms. The error estimand is the absolute query-CE prediction error $e_{e,o} = |\mathrm{CE}_q(\hat{\theta}_{e,o}) - \mathrm{CE}_q(\theta_{e,o})|$ on held-out queries; per-episode values are means over the 120 orders.

Evaluation criteria. The supervised arm evaluates the predictor under the matched-delay conventions of Appendix B.7 (All25 prototype-head initialization with per-(checkpoint,graph) selected step sizes, native β and clipping) on the same pre-specified first-500-episode subset of the 800-episode evaluation bank. It therefore differs from both the native-operator primary setting of Section 6 (whose S_1-prototype results use a separate confirmation bank) and the independent 300-episode confirmation bank in Appendix B. The CLIP-LoRA arm re-executes the predictor diagnostics on the last-block LoRA configuration of Appendix A.10 (six adapted tensors, plain-SGD step 0.03, no clipping or weight decay, fp32) over all four datasets $\times$ 200 episodes = 800. Criteria: within-regime confirmation Spearman ≥ 0.30 at $\geq 3/4$ checkpoints, and cross-regime separation (one-sided Mann–Whitney with alternative "CLIP-LoRA larger," Holm-adjusted over the three diagnostics, $p < 0.01$); for the cross-regime test, the supervised sample comprises the 500 episode-level diagnostic values of the main checkpoint only, the CLIP-LoRA sample all 800 episode-level values, and the remaining supervised checkpoints enter only the within-regime assessment. We treat κ_0 as an analysis diagnostic rather than an inexpensive online screen: in the CLIP-LoRA arm its finite-difference implementation requires additional mixed-read trajectory evaluations and is numerically sensitive across devices in fp32.

Results and boundaries. κ_0 and D_G satisfy both criteria (Table 32): ρ spans 0.824–0.843 and 0.790–0.826, worst-decile AUC 0.91–0.94, Holm-adjusted regime $p \leq 2.1 \times 10^{-202}$ and $\leq 5.5 \times 10^{-202}$, with CLIP medians 7.85 and 2.25 versus 0.200 and 0.022 supervised (about 39$\times$ and 103$\times$). A_G separates the regimes (median 0.976 versus 0.582; Holm 9.5×10^{-191}) but is negatively associated with within-regime prediction error (ρ -0.596 to -0.608, AUC 0.21–0.27), so it does not provide the intended reliability signal. Descriptive, outside the confirmatory family: CLIP per-dataset Spearman correlations are 0.580–0.681 (κ_0), 0.690–0.802 (A_G), and -0.122–0.747 (D_G, EuroSAT negative), so the within-CLIP signal is not uniform. The regime separation is distributional, and the CLIP diagnostic magnitudes are much larger in absolute terms; the separating claim therefore does not assert a per-episode normalized effect. In the pre-specified analysis, absolute prediction error can covary with the magnitude of the graph-induced effect, and those two sources are not separated. The diagnostics distinguish the two tested regimes and track predictor error on the pre-specified reliability subset; this is not a validated failure detector, since the regimes also differ in representation, parameterization, scope, operator, data, and predictor behavior.

Post-hoc specificity analysis. A post-hoc analysis asks whether the diagnostics track prediction quality beyond absolute graph-effect magnitude. On the same subset, define per-episode order-ranking error $R = 1 - \rho_\pi(\hat{L}, L)$ over the 120 orders (average ranks for exact ties), reuse the absolute error E above, and control the episode's mean absolute graph-versus-Fresh CE gap M — an analysis-only effect-size covariate computed from held-out query outcomes, hence not deployable. Partial Spearman is computed within each checkpoint by residualizing the ranked diagnostic and ranked E on ranked M and correlating the two residual vectors; each episode-bootstrap replicate recomputes the same procedure, the

Table 32: Support-side reliability diagnostics (pre-specified protocol; per-episode values are means over the 120 orders of the $[3, 2]$ task). ρ: Spearman correlation between diagnostic and per-episode absolute query-CE prediction error (ρ_{conf} = median over the four checkpoints); AUC: worst-decile absolute-CE-prediction-error detection; Sup./CLIP-LoRA: regime medians; comp.: regime separation (one-sided Mann–Whitney, Holm over the three diagnostics, all $p < 10^{-190}$). A_G does not meet its within-regime criterion (stable negative association); CLIP-side per-dataset correlations are descriptive.

	Supervised (4 ckpts)		AUC	Median		Criterion
Diagnostic	ρ span	ρ_{conf}	(decile)	Sup.	CLIP-LoRA	comp.
κ_0	0.824–0.843	0.836	0.91–0.93	0.200	7.85	yes
A_G	−0.608–−0.595	−0.599	0.21–0.27	0.582	0.976	no (negative)
D_G	0.790–0.826	0.813	0.92–0.94	0.022	2.25	yes

four checkpoints use shared episode resamples, and the across-checkpoint median is recomputed within each replicate. Both diagnostics correlate positively with ranking error at every checkpoint (ρ 0.51–0.58 for κ_0 and 0.46–0.52 for D_G). Their association with absolute error also remains positive after controlling M (partial Spearman given M: 0.87–0.90 and 0.72–0.77, respectively). Across the two analyses, all sixteen checkpoint-level episode-bootstrap 95% CIs exclude zero (eight ranking associations and eight magnitude-controlled partial associations). Post-hoc analysis therefore suggests that the diagnostics contain information about prediction quality beyond absolute graph-effect magnitude; the analysis is descriptive (no p-values or multiplicity family) and does not make the diagnostics a per-episode reliability guarantee.

Native fixed-operator confirmation. The calibrated arm above varies step size across graphs and therefore does not answer the primary fixed-operator question: identifying inaccurate episodes *within* one checkpoint at its fixed native operator. A two-stage development/confirmation test therefore re-examines the diagnostics in the native supervised configuration of Section 6 (frozen per-checkpoint operator, S_1-prototype initialization, K=5, sixteen-graph family, canonical order), with the paired contrast error E_{pair} against Fresh over the fifteen stale graphs, episode-reduced as above, as the error estimand. On the development bank (300 episodes, four checkpoints; median E_{pair} 0.057–0.078 CE) the selection rule ($\rho \geq 0.30$ at $\geq 3/4$ checkpoints) selected $\bar{\kappa}_0$ (per-checkpoint ρ 0.765–0.807) and $\bar{D}_G$ (0.702–0.731) but not $\bar{A}_G$ (negative, as in the calibrated arm), and fixed the predicted-displacement baseline $\bar{b}_{\text{disp}} = \frac{1}{15}\sum_g \|\hat{\theta}_g^{\text{full}} - x_K\|_2$ as the incremental comparator (the mean over the fifteen stale graphs of the Euclidean displacement of the Full prediction $\hat{\theta}_g^{\text{full}}$ from the Fresh endpoint x_K; same concatenated-parameter norm as the diagnostics). On an independent confirmation bank (300 fresh episodes, zero overlap), $\bar{\kappa}_0$ replicated at ρ 0.724–0.806 at 4/4 checkpoints, retained partial ρ of 0.579–0.715 given predicted displacement (all episode-bootstrap intervals positive), and tracked the paired direction-error rate (ρ 0.415–0.469); $\bar{D}_G$ also satisfied every criterion (ρ 0.660–0.735), and the episode×graph cell-level analysis agrees ($\bar{\kappa}_0$ 0.755–0.786). Within the native fixed-operator supervised regime, support-side diagnostics thus discriminate episodes with larger paired prediction error beyond the predicted-displacement baseline. This is a distinct experiment from the calibrated arm above (neither substitutes for the other), and it does not make the diagnostics a per-episode reliability guarantee.

A.23 Beyond Finite-Horizon Adaptation: Scope Schematic

Figure 6 summarizes the scope of the evaluation-graph view. Any learning procedure in which computations read historical parameter states induces such a dependency structure. Our experiments intervene on this structure in finite-horizon adaptation, where the graph family can be enumerated and compared directly. Local and distributed SGD, asynchronous optimization, continual learning, meta-learning, and test-time adaptation share the same

structural feature, but we do not test whether the effect sizes, predictors, or selection results transfer to those settings.

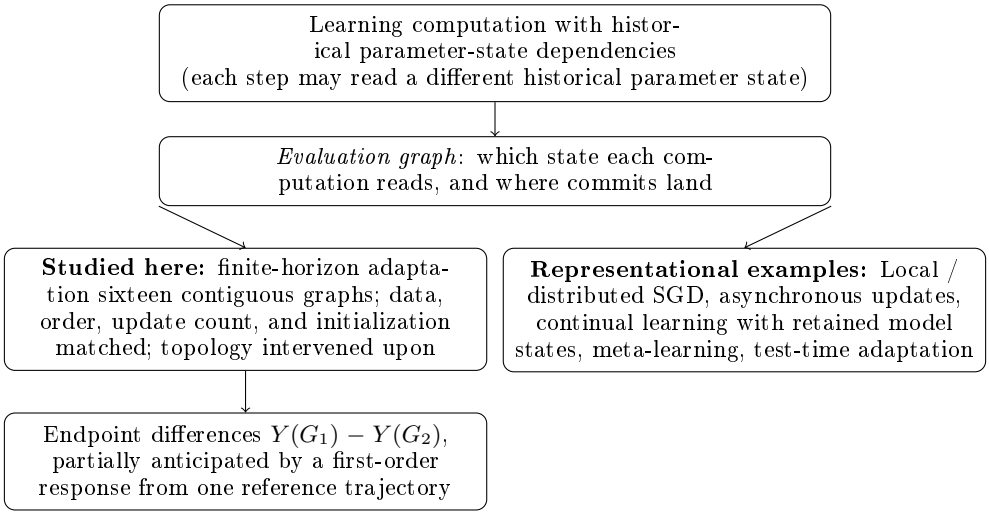


Figure 6: Scope of the evaluation-graph abstraction. The abstraction describes learning computations with historical parameter-state dependencies in general; systematic intervention, endpoint effects, and prospective prediction are established in this paper only within the controlled finite-horizon regime (left). Instances on the right are representational, not empirical validations.

B Fresh-Reference Predictor: Derivation and Evaluation

This appendix develops the Fresh-reference predictor used in Section 6, evaluates its behavior in the supervised setting, and examines its transfer to the TTA regime discussed in Section 7. Predictor development uses support-side quantities; held-out query outcomes are introduced only for the final evaluation.

B-I. From Defects to the Fresh-Reference Predictor

B.1 Initialization-Anchored Approximation and Its Limitation

A jet-propagated expansion of the exact operator gives second-order pairwise terms $O(\alpha^2)$ and third-order terms involving chained Jacobians and curvature. Synthetic tests recover the expected small-step scaling. On 100 development episodes from the main checkpoint, the second-order initialization-anchored predictor has a fitted log–log slope of 2.78 for median absolute endpoint error over the four smallest fixed scales (0.05–0.35× native). This is consistent with an $O(\alpha^3)$ absolute residual over the tested range. Because endpoint displacement is itself first order in α when nondegenerate, the corresponding relative error is asymptotically $O(\alpha^2)$.

The fitted slope is a finite-range diagnostic rather than an asymptotic verification. At the operational step size $\alpha \approx 0.5$, the third-order correction worsens the endpoint error in 85% of cases (median relative error 3.61 versus 1.05), so we do not use the perturbative expansion for prediction at the operational scale.

B.2 Exact Stale-Read Defects and Pathwise Transport

On 100 development episodes with the native operator, every stale read t receives an exact local defect $\ell_t = v_t(r_t) - v_t(y_{t-1})$ (the nonlinear update increment evaluated at the stale block-start state versus the current accumulated state; no Taylor approximation), an exact

single-read repair effect q_t obtained by re-executing the trajectory with that one read restored, and an exact Shapley attribution over stale reads. Findings: (i) terminal stale reads (R_t=0) satisfy $q_t = \ell_t$ to 6×10^{-14}—defects pass to the endpoint unmodified, an identity of the block-commit semantics; (ii) nonterminal defects retain a median propagation ratio of 0.890 with nearly unchanged direction (the terminal–nonterminal difference is 0.110, episode-cluster bootstrap 95% CI $[0.086, 0.125]$; median $\cos(q_t, \ell_t) = 0.99986$); (iii) linear transport of the exact defect through Jacobians evaluated at the *visited* states predicts the exact repair effect with median cosine 0.985 and median relative error 0.18 ($\|\hat{q}_t - q_t\|_2 / \|q_t\|_2$, with $\hat{q}_t$ the transported defect; 74.9% of instances within cosine 0.9 and relative error 0.5); (iv) the Shapley sum reproduces $\theta_G - \theta_{\text{FRESH}}$ to 1.4×10^{-16} on every episode and graph; (v) clipping-branch changes are rare (1%) but are associated with roughly twice larger defects and three times larger transport residuals, suggesting that they are a secondary source of prediction error rather than the dominant effect. All analyses in this subsection use support-side trajectories only.

B.3 Fresh-Reference Predictor and Prospective Evaluation

For each episode and support order π: run only the Fresh trajectory $x_t = F_{\pi(t)}(x_{t-1})$ (so x_t is the post-update state and update t reads x_{t-1}); form exact finite defects $\tilde{\eta}_t^G = v_{\pi(t)}(x_{a_G(t)-1}) - v_{\pi(t)}(x_{t-1})$, where $a_G(t)$ is the first update position of the block containing t (the defect is zero for block-start updates), each evaluated at pre-update Fresh-path states; transport each by $\Phi_F(K{:}t) = A_K \cdots A_{t+1}$ with $A_s = DF_{\pi(s)}(x_{s-1})$ (JVPs of the operator map); and predict $\hat{\theta}_G = x_K + \sum_t \Phi_F(K{:}t)\, \tilde{\eta}_t^G$. Here $\tilde{\eta}_t^G$ is an exact field difference evaluated between Fresh-reference states, not the exact target-graph perturbation after earlier blocks have departed from Fresh. Approximation enters by replacing target-graph block-start states with Fresh-reference states and linearly superposing and transporting the resulting defects along Fresh-path Jacobians. This predictor was fixed before evaluation on the held-out confirmation bank.

Algorithm 1 (Section 6) summarizes the predictor compactly. In the held-out evaluations, endpoint predictions are computed before query CE is evaluated. For PG-JGS, predicted endpoints are scored only on the held-out support-validation fold, so query data do not enter graph selection.

Continuous-intervention interpretation. The predictor also admits an exact variational reading. For a fixed support order and target graph, define the read-mixing path $y_0^\lambda = x_0$ and $y_t^\lambda = y_{t-1}^\lambda + (1-\lambda)\, v_{\pi(t)}(y_{t-1}^\lambda) + \lambda\, v_{\pi(t)}(y_{a_G(t)-1}^\lambda)$. At $\lambda = 0$ this is the Fresh trajectory; at $\lambda = 1$ it reproduces the target graph's block-frozen read semantics exactly, since every update in a block then reads the committed block-start state of the interpolated path. With $h_t = \partial y_t^\lambda / \partial \lambda|_{\lambda=0}$, the λ-prefactor of the block-start read removes its $h_{a_G(t)-1}$ chain term at first order, and differentiating the recursion gives $h_t = A_t h_{t-1} + \tilde{\eta}_t^G$ with $A_t = DF_{\pi(t)}(x_{t-1})$; hence $h_K = \sum_t \Phi_F(K{:}t)\, \tilde{\eta}_t^G$ and $\hat{\theta}_G = x_K + h_K$ — Algorithm 1 is exactly the first-order response at Fresh of this continuous intervention. If the interpolating path remains in a region where the active clipping branches are fixed and the update fields are twice differentiable, then $\theta_G - \hat{\theta}_G = \int_0^1 (1-\lambda)\, \partial^2 y_K^\lambda / \partial \lambda^2 \, d\lambda$ and $\|\theta_G - \hat{\theta}_G\| \le \frac{1}{2} \sup_{\lambda \in [0,1]} \|\partial^2 y_K^\lambda / \partial \lambda^2\|$. This is an expansion in the read-placement intervention strength λ, not in the step size; it characterizes where the first-order account should be reliable, but it is not usable as a real-time guarantee, since the path curvature is not observable at selection time. The predictor fits no graph-level coefficients, does not use query labels during construction, and does not execute the candidate graph. Its distinct Fresh-path field evaluations grow quadratically in K, while the candidate family grows as 2^{K-1}. We compare against two simpler approximations: local defects without transport, and the initialization-anchored predictor under revised position-space masks, $\hat{\theta}_P^{\text{anch}} = \theta_{\text{CO}} + \alpha^2 \sum_{\text{block}(s) < \text{block}(t)} J_{\pi(t)}(\theta_0)\, u_{\pi(s)}(\theta_0)$ (the CO-anchored first-order interaction model of Section 5; Table 34). The earlier equal-count analysis associated schedule masks with group identity rather than update position, so it did not preserve staleness under support permutation. Under the revised position-space definition, the initialization-anchored predictor reaches median rank correlations of 0.49–

0.51, with sign agreement 0.65–0.70 (Table 34). The 0.40–0.42 values reported in Section 5 correspond to the original identity-conditioned diagnostic.

Because the Fresh endpoint is exact for its own trajectory and the common-origin endpoint is invariant to support order, we evaluate the harder partial graph $[3, 2]$ over all 120 support permutations. All permutations share the same S_1 prototype initialization and differ only in update order. On support-side development data, the Fresh-reference predictor reduces median endpoint error by 76% relative to the initialization anchor.

On a separate 300-episode evaluation bank, median 120-order Spearman correlations are 0.722–0.772 for the full predictor, 0.569–0.616 for local defects, and 0.448–0.514 for the initialization anchor (Table 33). Across the full sixteen-graph family, median correlations range from 0.824 to 0.871. For the revised equal-staleness comparison they range from 0.695 to 0.736, with natural-order sign agreement of 0.780–0.813.

These results satisfy the pre-specified evaluation criteria; the criteria are reported separately and are not combined into a familywise-error-controlled claim. Because the block-parallel target depends only on group membership within each block, the 120 permutations induce 10 distinct target values for the $[3, 2]$ endpoint; the $[3, 2]$–$[2, 3]$ and $[4, 1]$–$[1, 4]$ contrasts induce 30 and 20 distinct target values, respectively. Exact target ties are handled by average ranks in the reported Spearman correlations.

A preliminary bank used during pipeline debugging was discarded after query labels were accessed. All reported results use a separately generated evaluation bank whose predictions were fixed before query evaluation.

Table 33: Fresh-reference finite-step prediction on an independent 300-episode confirmation bank. Entries are medians of per-episode Spearman correlations between predicted and actual query CE over all 120 support orders of the $[3, 2]$ partial graph (columns 2–4); the final row reports the revised equal-staleness contrast (the revised position-space matched-staleness estimand: CE difference of the two S=0.4 position shapes across the same 120 orders). The predictor and all thresholds were fixed before any query from this bank was accessed; all predictions were recorded beforehand, and stored hashes verified that the records were unchanged.

Checkpoint	θ_0 anchor	Local defects only	Fresh-path transport
main	0.503	0.609	**0.770**
strong-42	0.514	0.569	**0.729**
strong-43	0.448	0.573	**0.722**
strong-44	0.503	0.616	**0.772**
Equal-staleness contrast ρ: **0.736** / **0.701** / **0.695** / **0.723**			

Table 34: Scope of the pre-specified CO-anchored first-order interaction predictor (median per-episode Spearman, main / strong-42 / strong-43 / strong-44; s.a. = sign agreement). The sixteen-schedule ranking meets its confirmatory criterion; the 120-order ranking falls below the pre-specified criterion, and a downstream-transport correction improves it slightly while remaining below the criterion. The earlier equal-count diagnostic used identity-conditioned masks; the revised position-space variant is descriptive (bottom). Table 33 reports the Fresh-reference predictor used for the supervised order-ranking analysis.

Test	Result	Outcome
Rank 16 schedules	0.51/0.52/0.50/0.52	meets criterion
post-hoc: S-residualized	$\rho^{\text{res}} \approx 0.49$–$0.53$	descriptive
post-hoc: within-S sign agr.	0.69–0.79	descriptive
Rank 120 orders	0.40/0.40/0.40/0.42	below criterion
+ downstream transport	0.46/0.43/0.43/0.43	below criterion
Earlier equal-count (identity masks)	$-0.02/-0.01/-0.00/-0.01$ (s.a. 0.50)	historical
Revised position-space, Φ=I	0.49/0.49/0.49/0.51	descriptive
Revised position-space + transport	0.37/0.36/0.36/0.38 (s.a. 0.65)	below criterion

Table 35: Pooled TTA reversal on three independent untouched banks built after the original bank of Table 12 (whose graded slope was −0.199). Each bank froze its own pooled H0 estimand before execution (first bank: pooled actual S-slope; second and third banks: mean per-stream CE–S slope); slot-cluster bootstrap over 125 paired slots carrying all four corruptions. Negative means common-origin adaptation has lower held-out CE.

Bank	Image pool	CE–$S(P)$ slope	95% CI
Confirmation I	val split (disjoint)	−0.346	[−0.381, −0.311]
Confirmation II	test split	−0.415	[−0.436, −0.394]
Confirmation III	train split	−0.424	[−0.446, −0.402]

B-II. Predictor Evaluation and Scope Boundaries

B.4 Predictive Account in the TTA Regime

Wrong-class confidence strongly tracks TTA graph effects. On an independent evaluation set (test-split image pool, same 125-slot paired design and operator), the median within-stream Spearman correlation between the change in wrong-class confidence $R(z, y) = \log \sum_{j \neq y} e^{z_j}$ and the change in held-out CE across the fifteen stale graphs is 0.784 [0.711, 0.821], reaching 0.954 on motion blur and 0.996 on pixelate, and the pooled signed wrong-class slope is −0.077 [−0.080, −0.074]. A second independent evaluation set (train-split pool) with stream-level statistics—used because all cells of a stream share that stream's FRESH reference CE, so cell-wise resampling is uninformative—yields median $\rho = 0.946$ [0.943, 0.950], sign agreement 0.968 [0.963, 0.972], and mean absolute error 0.042 [0.040, 0.045] over the fifteen stale graphs, with per-corruption median ρ 0.946–0.950. Adding Fresh-path transport does not improve this account (increment bounded by +0.014 against a +0.02 materiality margin; MAE 0.042 versus 0.079 with transport). The pooled reversal also appears on all three subsequent evaluation sets (Table 35).

B.5 Scope Boundary in CLIP-LoRA

Two natural extensions tested on the development split did not close the CLIP-LoRA transfer gap. A block-recursive first-order variant reaches pooled median rank correlation 0.075, while adding a quadratic curvature correction gives a residual ratio of 1.965. Neither extension materially improves the transfer result, so we treat CLIP-LoRA as a current scope boundary of the predictor. These development-only results are descriptive rather than confirmatory.

B.6 Incremental Predictive Value

How much of the predictor's 120-order ranking is inherited from the Fresh trajectory's own order dependence, and does the transported-defect correction add incremental information about the target graph's deviation from Fresh? A post-hoc diagnostic on the same independent bank and its stored endpoint predictions (deterministic replay; no new model or adaptation run) answers both. Two per-episode quantities join the comparator set: the Fresh-query-only oracle, the Spearman correlation between the actual Fresh per-order query CE and the actual $[3, 2]$ per-order query CE (an evaluation diagnostic that reads held-out queries; not a deployable predictor), and the residual diagnostic, the Spearman correlation between the predicted and actual per-order deviations of the $[3, 2]$ query CE from the Fresh value. Medians over 300 episodes: the Fresh-query-only oracle retains moderate order information (0.501–0.532), the initialization anchor 0.448–0.514, local defects 0.569–0.616, the full predictor 0.722–0.772, and the residual-to-Fresh diagnostic 0.659–0.698—a complementary residual diagnostic, since predicted and actual residuals share the Fresh reference. A paired per-episode increment, $\Delta\rho = \rho_{\text{full}} - \rho_{\text{oracle}}$, is positive at every checkpoint: per-checkpoint medians 0.196–0.230 with descriptive 95% episode-bootstrap CIs whose lower bounds range 0.19–0.23. Fresh query ranking alone therefore does not explain the full predictor's order ranking, and the transported-defect correction carries additional rank information beyond

the Fresh-query-only oracle. This analysis is post hoc and is not part of the confirmatory evaluation.

Native matched-delay paired increment (post-hoc). A second post-hoc diagnostic evaluates the three matched-delay placement pairs of Section 5 under the native operator on the same 300-episode confirmation bank: for each pair, the paired-contrast error $E_{\text{pair}} = |(\widehat{\text{CE}}_P - \widehat{\text{CE}}_Q) - (\text{CE}_P - \text{CE}_Q)|$ and the pairwise sign error of the three predictor variants (Table 36). Local defects alone already predict the binary placement contrasts well (median E_{pair} 0.063–0.101, sign error 14–20%, strong-family pooled), and the transport increment is mixed: it reduces paired-contrast error for $[3,2]$ vs. $[2,3]$, has little detectable effect for $[3,1,1]$ vs $[1,1,3]$, and increases error for $[4,1]$ vs $[1,4]$; the corresponding paired episode-cluster bootstrap intervals appear in Table 36. The predictor's increment therefore concentrates in the harder order-ranking regime (median rank correlation 0.722–0.772 versus 0.569–0.616 for local defects), while on the coarse binary placement contrasts local defects capture most of the pair-level difference.

Table 36: Native-operator paired prediction on the three matched-delay placement pairs (post-hoc; strong-family pooled, 900 episode×checkpoint records from 300 shared episodes; per-pair values are median E_{pair} with sign-error rate in parentheses; the Full−Local column is the paired mean difference of E_{pair}, not the difference of the median columns, with 95% episode-cluster bootstrap CI (episodes resampled with all three checkpoint records).

Pair	Anchor	Local	Full	Full−Local ΔE [95% CI]
$[4,1]$ vs $[1,4]$	0.196 (0.26)	0.101 (0.20)	0.160 (0.24)	$+0.076$ $[+0.051, +0.102]$
$[3,2]$ vs $[2,3]$	0.188 (0.29)	0.089 (0.20)	0.086 (0.21)	-0.059 $[-0.081, -0.039]$
$[3,1,1]$ vs $[1,1,3]$	0.242 (0.32)	0.063 (0.14)	0.060 (0.13)	-0.008 $[-0.027, +0.010]$

K=8 extension of the predictor (confirmatory). A confirmatory run applies Algorithm 1 to the complete 2^7=128-graph K=8 contiguous family at the native operator on the three strong checkpoints, using a new 300-episode bank with zero overlap with the K=8 sensitivity bank; composition identities hold exactly (Fresh equals sequential execution and common-origin equals parallel sum, and the predictor reproduces the Fresh endpoint on the all-fresh graph). Pooled over 900 episode×checkpoint evaluations, median rank correlation over the 128 graphs is 0.633 $[0.601, 0.664]$ (episode-cluster bootstrap) for Full versus 0.520 $[0.493, 0.561]$ for Local and -0.052 for the initialization anchor; the paired Full−Local increment is $+0.080$ $[0.067, 0.099]$, defined as the median over the 900 episode×checkpoint cells of the per-cell $\rho_{\text{Full}} - \rho_{\text{Local}}$, with the interval from an episode-cluster bootstrap (10,000 resamples; all three checkpoints of a sampled episode move together)—a different statistic from the difference of the marginal medians, $0.633 - 0.520 \approx 0.113$, since medians do not commute with differencing. Absolute fidelity remains below the K=5 level (0.722–0.772); the extension is partial. Screening is scored oracle-style (graphs ranked by query CE of their predicted heads): the predicted top-16 contains the actual-best graph in 49.6% of episodes with median best-in-subset regret 0.0003 CE (top-8: 38.2%, 0.012). Within matched block-multiset classes, Full preserves placement discrimination (within-class centered Spearman 0.551, sign agreement 74%; Local 0.411, 65%), so the 128-graph ranking is not carried by the coarse staleness trend alone. Exhaustive execution applies 1024 operator instances, whereas the current Full predictor performs 1792 field evaluations, although these involve only 34 unique historical-state/update-group pairs that are reused across candidate graphs. The present implementation therefore does not reduce raw computation at K=8; its advantage is reuse of one reference-path construction across the full candidate family.

B.7 Joint-Calibrated Matched-Delay Prediction

Design. The matched-delay experiment of Section 6 runs on the first 500 episodes of the 800-episode evaluation bank, under joint full-support initialization with per-graph fixed step sizes (clamped to the operator bounds), 120 support orders, and the three pre-specified matched-delay pairs. For every episode, checkpoint, and order, the predictor endpoint

Table 37: Incremental matched-delay prediction: all three predictors under the same prospective design (descriptive post-hoc; the prespecified 500-episode subset of the 800-episode evaluation bank, 120 support orders, and pair contrasts of Table 38). Entries are medians of per-episode Spearman correlations between the predicted and actual pair contrast. Paired per-episode increments are summarized in the text; all 24 episode-bootstrap intervals exclude zero.

Pair	Checkpoint	ρ Full	ρ Local	ρ Anchor
$[3,2]$ vs $[2,3]$	main	0.986	0.981	0.914
$[3,2]$ vs $[2,3]$	strong-42	0.987	0.982	0.922
$[3,2]$ vs $[2,3]$	strong-43	0.987	0.982	0.915
$[3,2]$ vs $[2,3]$	strong-44	0.988	0.983	0.929
$[4,1]$ vs $[1,4]$	main	0.938	0.950	0.929
$[4,1]$ vs $[1,4]$	strong-42	0.979	0.983	0.947
$[4,1]$ vs $[1,4]$	strong-43	0.978	0.982	0.948
$[4,1]$ vs $[1,4]$	strong-44	0.980	0.984	0.952
$[3,1,1]$ vs $[1,1,3]$	main	0.996	0.992	0.924
$[3,1,1]$ vs $[1,1,3]$	strong-42	0.996	0.991	0.928
$[3,1,1]$ vs $[1,1,3]$	strong-43	0.996	0.992	0.923
$[3,1,1]$ vs $[1,1,3]$	strong-44	0.996	0.992	0.935

$\hat{\theta}_G$ (Algorithm 1, full-transport variant) and the local-defect and initialization-anchored comparator endpoints were computed before any held-out query was read; the query evaluation then ran as a single pass. On the three strong checkpoints all six candidate shapes share one calibrated step size, so the pair contrasts isolate placement; on the main checkpoint the $[4,1]/[1,4]$ pair includes a distinct tuned step size, so that pair contrasts jointly tuned policies.

Comparators under the same design. The local-defect and initialization-anchored comparator endpoints were computed under exactly the same prospective design as the full predictor. Table 37 reports all three predictors on each pair and checkpoint. Transport improves on the initialization anchor across all 12 pair-by-checkpoint comparisons, and the Full−Local increment is small and regime-dependent: positive on the eight $[3,2]/[2,3]$ and $[3,1,1]/[1,1,3]$ cells and negative on the four $[4,1]/[1,4]$ cells, consistent with rank saturation at these small calibrated steps; in the native-operator 120-order test the same transport contributes a much larger increment (Section 6). The paired per-episode increments (descriptive post-hoc means of $\rho_{\text{Full}} - \rho_{\text{comp}}$ with episode-bootstrap 95% intervals, B=10^4; no confirmatory multiplicity family) are $+0.043$ to $+0.115$ for Full−Anchor with all twelve intervals excluding zero, $+0.005$ to $+0.009$ for Full−Local on its eight positive cells, and -0.024 to -0.006 on the four $[4,1]/[1,4]$ cells, all with intervals excluding zero. These paired means are distinct from the median-increment summaries used by the six-shape criteria; per-cell values are recorded in the supplementary repository manifest.

Error decomposition. For episode i let $\hat{d}_{i,\pi}$ and $d_{i,\pi}$ be the predicted and actual pair contrasts over the 120 orders. The centered mean absolute error is $\text{cMAE}_i = \frac{1}{120}\sum_\pi \left|(\hat{d}_{i,\pi} - \bar{\hat{d}}_i) - (d_{i,\pi} - \bar{d}_i)\right|$, the uncentered error is $\text{MAE}_i = \frac{1}{120}\sum_\pi |\hat{d}_{i,\pi} - d_{i,\pi}|$, and the signed bias is $\text{Bias}_i = \frac{1}{120}\sum_\pi (\hat{d}_{i,\pi} - d_{i,\pi})$. Across the twelve pair×checkpoint cells the uncentered MAE ranges from 0.0011 to 0.0029 CE and the signed mean bias is small, at most 0.0009 CE in magnitude: the predicted pair contrasts are rank-correct, with only a small mean offset.

Per-checkpoint results. Table 38 reports, for each pair and checkpoint, the per-episode Spearman rank correlation between the predicted pair contrast $\widehat{CE}_P(\pi) - \widehat{CE}_Q(\pi)$ and the actual pair contrast $CE_P(\pi) - CE_Q(\pi)$ over the 120 orders, the sign agreement, and the centered and uncentered mean absolute errors of the prediction. The pair contrast is the direct non-linear counterpart of the affine obstruction: it measures exactly the consequence that scalar delay summaries omit.

Table 38: Matched-delay pair contrasts: predicted versus actual (500-episode bank; per-episode Spearman over the 120 orders, medians across episodes; sign agreement; centered and uncentered mean absolute error of the prediction; full-transport variant).

Pair	statistic	main	strong42	strong43	strong44
$[3,2]$ vs $[2,3]$	pair-contrast ρ	0.986	0.987	0.987	0.988
	sign agreement	0.965	0.965	0.966	0.969
	centered MAE	0.0017	0.0016	0.0016	0.0014
	uncentered MAE	0.0017	0.0016	0.0016	0.0013
$[4,1]$ vs $[1,4]$	pair-contrast ρ	0.938	0.979	0.978	0.980
	sign agreement	0.977	0.962	0.963	0.966
	centered MAE	0.0019	0.0028	0.0029	0.0024
	uncentered MAE	0.0019	0.0028	0.0029	0.0024
$[3,1,1]$ vs $[1,1,3]$	pair-contrast ρ	0.996	0.996	0.996	0.996
	sign agreement	0.981	0.982	0.980	0.983
	centered MAE	0.0013	0.0012	0.0014	0.0011
	uncentered MAE	0.0013	0.0012	0.0013	0.0011

Six-shape calibrated extension. Under the same joint full-support initialization with per-graph fixed step sizes, the pre-specified criteria on the primary $[3,2]$ shape are met at all four checkpoints: full-transport median ρ 0.975–0.979, and positive paired increments over both the initialization-anchored and the local-defect comparators (median positive with bootstrap CI excluding zero at 4/4 checkpoints each). The six pre-specified multi-refresh shapes range 0.885–0.993 across shapes and checkpoints (per-shape medians in the repository records). At the small calibrated steps the initialization-anchored and local-defect comparators compress toward rank saturation, so absolute levels are regime-specific and not comparable with the fixed-operator regime reported above.

Target scale and ties. The contrast target is non-trivial: the within-episode range of the actual pair contrast is 0.034–0.086 CE at the median, with at most one episode per cell in which the pair contrast is exactly zero (a degenerate target order; target-scale check). The contrast-scale and zero-contrast summaries are descriptive post-hoc diagnostics; they were not success criteria.

B-III. Exact Structural Theory

B.8 Exact Affine Placement Characterization

Setting and block map. Let each per-group update be exactly affine, $F_t(x) = A_t x + b_t$ ($t = 1, \dots, K$, executed in identity order), and let a block evaluate all its updates at the block-start state, committing at its boundary (the composition semantics of the main-text family). The block map is $C_B(x) = x + \sum_{t\in B}(F_t(x) - x) = (I + S_B)x + b_B$ with $S_B = \sum_{t\in B}(A_t - I)$ and $b_B = \sum_{t\in B} b_t$; the endpoint of a contiguous partition $P = (B_1, \dots, B_r)$ is $\theta_P = C_{B_r} \circ \cdots \circ C_{B_1}(x_0)$.

Exact two-block difference. For $P = ([4,1])$ and $Q = ([1,4])$, with $S_i = A_i - I$, $S_{234} = S_2 + S_3 + S_4$, $b_{234} = b_2 + b_3 + b_4$,

$$\theta_P - \theta_Q = \big(S_5 S_{234} - S_{234} S_1\big)x_0 + \big(S_5 b_{234} - S_{234} b_1\big),$$

exactly (no truncation at any step size). The linear part equals $[S_5, S_{234}] + S_{234}(S_5 - S_1)$ and also the difference of the included cross-block interaction products $\sum_{(s,t)} S_t(S_s x_0 + b_s)$ between the two placements; the omitted stale-pair mask is the complement of the interaction mask in the full ordered-pair universe, so the two descriptions are equivalent. For two-block partitions the difference is exactly second order in the update contrast ($A_t = I + \alpha M_t$, $b_t = \alpha c_t$ gives α^2 times the pair-interaction coefficient, with no cubic remainder); partitions with three or more blocks carry genuine higher-order nested products.

Homogeneous same-delay invariance. If every update shares one affine map ($A_t = A$, $b_t = b$), a block of size m has $C_m(x) = (I + mS)x + mb$ with $S = A - I$, and block maps commute: $(I + mS)(I + nS) = (I + nS)(I + mS)$ and the affine parts agree. The endpoint therefore depends only on the multiset of block sizes, not on their placement. Within the contiguous K=5 family the raw count $C(P) = \sum_B \binom{|B|}{2}$ takes the seven distinct values $\{10, 6, 4, 3, 2, 1, 0\}$ across the seven block-size multisets (equivalently, the normalized $S(P) = C(P)/\binom{5}{2}$ takes $\{1, 0.6, 0.4, 0.3, 0.2, 0.1, 0\}$), so equal $S(P)$ identifies equal block-size multisets, and matched-delay partitions have identical endpoints. Different block-size multisets generally differ ([5] versus fully sequential: $I + 5S \neq (I + S)^5$).

Sources of the placement difference. With heterogeneous or noncommuting maps the identity fails unless the exact ordered-product terms cancel: commuting but heterogeneous maps leave $S_{234}(S_5 - S_1)x_0$, and noncommuting maps leave $[S_5, S_{234}]x_0$ (plus the constant mixed term $S_5 b_{234} - S_{234} b_1$). The affine model is an exactly solvable structural system: placement effects arise already at this level, without higher-order nonlinear update maps; this does not speak to whether the underlying objective has Hessian–gradient structure. Every closed form above was verified numerically against direct composition on random affine maps (deterministic seed; relative error $\leq 10^{-15}$; the two-block difference confirmed exactly α^2; a three-block fixture shows the cubic residue), in a verification record of the supplementary repository.

Approximate-homogeneity bound. If, for the matched pair above, each update is near a common map, $S_t = S + E_t$ and $b_t = b + e_t$ with $\|E_t\| \leq \varepsilon$ and $\|e_t\| \leq \delta$ under a compatible submultiplicative norm, the triangle inequality applied to the exact identity gives

$$\|\theta_P - \theta_Q\| \leq (12\|S\|\varepsilon + 6\varepsilon^2)\|x_0\| + 6\|S\|\delta + 6\varepsilon\|b\| + 6\varepsilon\delta,$$

so the placement invariance at $\varepsilon = \delta = 0$ degrades continuously with update heterogeneity. Under $A_t = I + \alpha M_t$ and $b_t = \alpha c_t$, with $M_t = M + \Delta M_t$, $c_t = c + \Delta c_t$, $\|\Delta M_t\| \leq \varepsilon_M$, $\|\Delta c_t\| \leq \varepsilon_c$, the discrepancy is at most

$$\alpha^2\big[(12\|M\|\varepsilon_M + 6\varepsilon_M^2)\|x_0\| + 6\|M\|\varepsilon_c + 6\varepsilon_M\|c\| + 6\varepsilon_M\varepsilon_c\big],$$

second order in the step size with a coefficient controlled by update heterogeneity. The bound gives a sufficient stability statement around the homogeneous symmetry, not a necessary condition for small placement differences. It can be conservative because exact algebraic cancellation may make the actual discrepancy much smaller. The result is specific to this matched two-block comparison, and ε and δ are theoretical perturbation parameters rather than observed diagnostics. We therefore do not use the bound as an online reliability criterion.