Functional compatibility as a determinant of persistent neural learning

Hossein Javidnia *hossein.javidnia@dcu.ie*
School of Computing
Dublin City University
Dublin, Ireland

Abstract

Artificial neural networks can acquire new capabilities but often damage existing ones when they continue to learn.(Dohare et al., 2024; Kirkpatrick et al., 2017) This stability-plasticity problem has motivated replay, regularization and constrained-update methods, yet it remains unclear whether a property of incoming learning itself determines what can be retained without disrupting protected behaviour.(Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019; Farajtabar et al., 2020; Benjamin et al., 2019) Here we show that functional compatibility, the extent to which new learning can coexist with behaviour that must be preserved, is a causal determinant of persistent learning. To our knowledge, this is the first controlled causal demonstration in which compatibility is deliberately changed from matched neural states and persistent learning is measured under a common retention requirement.(Wang et al., 2021; Störk, 2026; Kitkana & Arora, 2026; Abbes et al., 2026) The effect generalizes across independent learning directions, convolutional and transformer architectures, vision and text, and additional seeds. Learning rules differ in how efficiently they exploit available compatibility, while retention constraints limit how much can be stored. At larger finite updates, nonlinear geometry changes the available learning opportunity and ultimately prevents the matched compatibility continuum from being realized. These results establish functional compatibility as an experimentally controllable principle of persistent neural learning, shifting the problem from preventing forgetting towards identifying which components of new learning can safely become permanent.

1 Introduction

Continual learning is commonly framed as a competition between acquiring new information and preserving what a neural network already knows. Regularization, replay and constrained-gradient methods reduce interference by limiting parameter changes, revisiting old observations or projecting updates away from protected directions.(Kirkpatrick et al., 2017; Rolnick et al., 2019; Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020; Saha et al., 2021) Function-space approaches instead protect the behaviour of the network directly.(Benjamin et al., 2019) These strategies have transformed continual learning, but they leave a more basic question unresolved: before choosing an algorithm, is there a property of the incoming learning signal itself that determines how much of it can coexist with retained behaviour?

This question emerged from Adaptive Functional Metaplasticity (AFM), a task-free framework that treats retention and learning as a measurable frontier rather than requiring either to dominate. AFM first computes the update that the same learner would accept from the identical pre-update state if no protection were imposed. This *same-state no-protection endpoint* is computed but never committed. It defines the available learning opportunity. AFM then identifies the component compatible with protected functional behaviour and stores an accepted fraction of that motion in the ordinary network parameters. For the scalar-gradient comparator used in the central guarantee, functional compatibility is

$$\kappa = \frac{\|\Pi_t g_t\|^2}{\|A_t g_t\|^2}, \tag{1}$$

where A_t selects the active learning coordinates and Π_t is the protected-compatible projector. Under the stated certificate conditions, the accepted persistent path fraction satisfies $\widehat{\lambda} \geq \eta$, and

$$\frac{\Delta_{\text{persistent}}}{\Delta_0} \geq \frac{\widehat{\lambda}\kappa}{3} \geq \frac{\eta\kappa}{3}. \tag{2}$$

A separate bounded functional completion step can reproduce the no-protection endpoint on the current finite evidence while restoring declared protected outputs. Persistent adaptation and finite deployed protection are therefore distinct: exact finite restoration does not imply that the unrestricted parameter update has been stored, and it does not by itself imply population-level retention.

AFM is more than this local projection rule. The complete framework maintains bounded sketches of whole-behaviour sensitivity, allocates protection through multiple metaplastic timescales and a rank-timescale controller, performs task-free observable routing with separately controlled consolidation and reopening, provides a guarded finite functional shield, and permits function-preserving renewal of zero-gated dormant structure. Its theory includes a deterministic streamed-memory certificate, an exact first-order rank-plasticity frontier, pathwise retention and projected-stationarity results, a convex dynamic-regret specialization, explicit resource bounds, and obstruction results that state when routing, retention, transfer or renewal cannot be guaranteed. AFM certifies persistent adaptation relative to the genuine same-state no-protection endpoint while separately providing exact finite counterfactual endpoint emulation. The complete specification and proofs are provided in the sections below.

Equation 2 also suggested a testable scientific prediction. If compatibility is more than a descriptive correlate of interference, experimentally changing κ from the same neural state should change the amount of new learning that survives a fixed retention requirement. Previous work has optimized gradient alignment, constructed null spaces, analysed interference energy and curvature, causally ablated gradient alignment in a specialized subliminal-learning setting, and demonstrated gradient-alignment methods in continual large-language-model pre-training.(Riemer et al., 2019; Farajtabar et al., 2020; Wang et al., 2021; Störk, 2026; Shu et al., 2026; Gong et al., 2026; Kitkana & Arora, 2026; Abbes et al., 2026) To our knowledge, no previous study has used a matched-state, multi-level intervention to set a quantified functional compatibility variable while controlling the unrestricted learning opportunity and retention requirement, and then measured its causal effect on persistent learning. We therefore tested compatibility as an intervened variable rather than as a post-hoc diagnostic.

2 AFM framework and theoretical foundations

The following sections give the complete AFM formulation, guarantees, implementation logic, benchmark protocols, ablations and execution analyses that support the paper.

2.1 Problem formulation and design principle

We consider supervised continual learning on a nonanticipating stream in which the learner receives observations and labels but no task, session, episode, bucket, segment, intervention, or boundary identifiers. At protected round t, the persistent model parameters are $\theta_t \in \mathbb{R}^d$ and the bounded structural state is S_t. The deployed predictor is

$$F_t(x) = G_{\theta_t}(z(x)) + S_t(z(x)), \tag{3}$$

where $z(x)$ denotes the representation available to the continual learner. For the pathwise retention and optimization results, the loss sequence may be arbitrary; stochastic assumptions are introduced only for the statistical statements that explicitly require them.

AFM separates three objects that are frequently conflated. First, *persistent adaptation* is learning stored in the ordinary model parameters and therefore affects future predictions away from any finite intervention set. Second, *finite deployed protection* refers to exact output constraints on declared observations. Third, *population behavior* concerns inputs beyond those finite protected observations and requires additional assumptions or certificates. The distinction is consequential: exact protection of a finite evidence set does not imply preservation of an unseen population, and AFM does not make that inference.

2.1.1 Same-state no-protection comparator

At a protected update, AFM first evaluates the update that the corresponding no-protection learner would actually accept from the same complete pre-step state. The comparator uses the same parameters, structural state, current minibatch, active coordinates, learning-rate and backtracking rules, mutable buffers, and declared private randomness, but no retention constraint. If the resulting endpoint is θ_t^0 with deployed current-minibatch logits U_t, its accepted prediction-space loss decrease is

$$\Delta_t^0 = \mathcal{L}_t(F_t) - \mathcal{L}_t(U_t) > 0. \tag{4}$$

This endpoint is a counterfactual reference only. During a protected transaction it is never installed as the persistent base model.

The comparator makes the plasticity cost of protection operational. Rather than comparing a protected step with a nominal gradient or an arbitrary step-size schedule, AFM asks how much of the learning achieved by the actual accepted no-protection transaction can be stored persistently without violating the declared retention geometry.

2.1.2 Persistent compatible assimilation

Let A_t denote the structural-availability projector and Π_t the selected protected projector. In the scalar-gradient comparator used by the evaluated implementation,

$$d_t^0 = -\alpha_t A_t g_t, \qquad v_t = \Pi_t d_t^0 = -\alpha_t \Pi_t g_t, \tag{5}$$

where $g_t = \nabla \ell_t^{S_t}(\theta_t)$. Write $s_t^0 = \|v_t\|$. The certified retention charge along this projected comparator path is

$$C_t(s) = E_t s + \frac{\overline{H}_t}{2} s^2, \tag{6}$$

with E_t the certified first-order leakage term and $\overline{H}_t$ the behavior-curvature bound. A predeclared assimilation coordinate $\eta_t \in [0,1]$ allocates

$$b_t^{\mathrm{norm}} = \eta_t C_t(s_t^0), \qquad \lambda_t = \max\{\lambda \in [0,1] : C_t(\lambda s_t^0) \le b_t^{\mathrm{norm}}\}. \tag{7}$$

The compatible-gradient fraction is

$$\kappa_t = \frac{\|\Pi_t g_t\|^2}{\|A_t g_t\|^2} \in [0,1]. \tag{8}$$

Thus η_t specifies how much of the projected comparator charge is requested, whereas κ_t measures how much active gradient energy is compatible with the selected protected geometry.

2.1.3 Finite endpoint completion

A persistent compatible step generally cannot reproduce the full current endpoint of the unrestricted learner. AFM therefore performs a second, explicitly separate operation in function space. After the safe persistent base move, a bounded compact-cardinal residual is constructed on the frozen representation. On the current minibatch it requests the comparator logits U_t; on active protected evidence and unselected certified candidates it requests their pre-step deployed outputs; on a selected candidate it requests the frozen transfer target or the declared contraction toward it. If these finite requirements are mutually consistent and the support, capacity, and numerical checks pass, the residual satisfies them exactly.

The residual is not treated as persistent assimilation. Away from its finite support, the deployed predictor follows the protected base endpoint. If finite requests conflict at the same observable address, compatible motion is unavailable, capacity is exhausted, or a required certificate fails, the transaction is rejected atomically. The complete specification, finite-support construction, and rollback conditions are given in Section 2.5.

2.2 Adaptive Functional Metaplasticity

The protected transaction is embedded in a bounded task-free learner that combines functional sensitivity memory, adaptive protection geometry, operational consolidation, evidence-based reopening, and function-preserving structural renewal. The full algorithm is given in Section 2.5; this subsection describes the mechanisms needed to interpret the theoretical and empirical results.

2.2.1 Whole-behavior sensitivity memory

A committed record protects a behavior map rather than a parameter vector. In the finite-evidence instantiation, a record j with evidence $x_{j,1}, \dots, x_{j,n_j}$ uses

$$\widehat{\Phi}_j(\theta) = n_j^{-1/2}\big(f_\theta(x_{j,1}), \dots, f_\theta(x_{j,n_j})\big). \tag{9}$$

All Jacobian rows of this map at the frozen anchor are streamed into a Frequent Directions sketch (Ghashami et al., 2016). The sketches define a bounded approximation to the protected sensitivity covariance. The selected rank-r protected subspace suppresses the most consequential directions, while the unprotected complement remains available for adaptation. The exact sketch certificate, its population extension under declared sampling assumptions, and the rank-r min-max frontier are given in Section 2.7.

2.2.2 Metaplastic allocation

AFM does not assign one permanent importance coefficient to every protected record. It maintains a finite bank of exponentially spaced traces and uses them to predict which protected sensitivities remain relevant at the current time. This construction is motivated by the ability of multiple timescales to represent long temporal ranges efficiently (Benna & Fusi, 2016). In AFM the traces have a specific optimization role: a finite policy family chooses both temporal weights and a spectral rank, and the policy loss charges residual protected leakage together with current gradient energy blocked by protection. Section 2.9 proves that a logarithmic trace bank approximates scale-free relevance profiles with constant distortion over an exponentially long horizon, whereas a single exponential has polynomial worst-case distortion; it also gives the allocation-transfer and online policy-regret bounds.

2.2.3 Task-free consolidation, reopening, and route refinement

Candidate consolidation is separated from both fitting and deployment. A private candidate is fitted on a finite routed block and then frozen. Later outcomes form a distinct validation sequence, so the candidate prediction precedes each validation outcome. An anytime-valid upper confidence sequence controls the first certification crossing; subsequent evidence is handled by a separate staleness process. AFM uses time-uniform inference because ordinary fixed-time tests are not generally valid under repeated optional checking (Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024).

Routing uses only observable signatures and never evaluator task identities. Outcome evidence can justify reopening or release, but a new observable route additionally requires independent signature evidence. Sequential two-sample procedures provide one admissible route-refinement construction (Jang et al., 2022). When observables do not distinguish two semantic states, AFM does not manufacture a route distinction: the resulting ambiguity appears as an obstruction. Full calibration, consolidation, staleness, reopening, and route-refinement statements are provided in Section 2.6.

2.2.4 Function-preserving structural renewal

The model contains a fixed pool of dormant zero-gated modules. Resetting internal parameters of a module whose functional gate is exactly zero leaves the deployed predictor and active protected behaviors unchanged. After a reset, AFM recomputes the complete gradient, protected projector, and leakage certificates; a slot is activated only if an accepted protected update produces nonzero motion in that slot. Renewal therefore expands usable structure without granting an exception to the retention transaction. Section 2.10 gives the exact reset result and shows that renewal success is controlled by the realized compatible richness of the trial distribution.

Fig. 6 summarizes how these components interact within one protected AFM transaction.

The full protected-update schematic is provided as Fig. 6 of the paper.

2.2.5 Protected update

A protected round can be summarized as follows:

1. route the current observations using only declared observable signatures;
2. update candidate fitting, certification, and staleness state without copying a private candidate into deployment;
3. update bounded whole-behavior sketches and metaplastic traces;
4. compute the genuine same-state no-protection endpoint and the compatible projected reference;
5. accept only a persistent base point that satisfies the retention, descent, and normalized-assimilation checks;
6. construct the finite residual that completes the current endpoint and restores the declared finite protected outputs;
7. update reopening, route-refinement, or renewal state only through their separately controlled evidence and capacity rules.

No failed certificate authorizes installation of the unrestricted base endpoint or reduction of the declared persistent-assimilation requirement.

2.3 Theoretical guarantees

The central guarantee is split into persistent and deployed components. This prevents exact finite interpolation from being misinterpreted as learning stored in the base model. Recall that $d_t^0 = -\alpha_t A_t g_t$ is the accepted same-state no-protection displacement, $v_t = \Pi_t d_t^0$ is its compatible projection, and

$$\kappa_t = \frac{\|\Pi_t g_t\|^2}{\|A_t g_t\|^2}$$

is the corresponding compatible-gradient energy fraction.

Theorem 2.1 (Counterfactual-normalized persistent assimilation)**.** *Assume the accepted no-protection comparator has the scalar-gradient form $d_t^0 = -\alpha_t A_t g_t$ with $0 < \alpha_t \le \overline{L}_t^{-1}$, A_t and Π_t are orthogonal projectors satisfying $\Pi_t A_t = \Pi_t$, the comparator decrease Δ_t^0 is positive, and the projected reference $v_t = \Pi_t d_t^0$ is nonzero. Assume the stated loss-smoothness and retention-charge certificates hold on the complete comparator and projected-reference line segments. Then the normalized construction in Eq.* (7) *satisfies*

$$\lambda_t \ge \eta_t. \tag{10}$$

For any accepted realized fraction $\widehat{\lambda}_t \ge \eta_t$,

$$\frac{\Delta_t^{\mathrm{base}}}{\Delta_t^0} \ge \widehat{\lambda}_t \kappa_t \frac{1 - \frac{1}{2}\alpha_t \overline{L}_t \widehat{\lambda}_t}{1 + \frac{1}{2}\alpha_t \overline{L}_t} \ge \frac{\widehat{\lambda}_t \kappa_t}{3} \ge \frac{\eta_t \kappa_t}{3}. \tag{11}$$

For $\eta_t = 1$, the complete projected comparator is feasible.

The proof, including the exact role of convexity of the retention charge and the reverse smoothness inequality used for the denominator, is given in Section 2.5. The guarantee is not an η_t fraction of unrestricted loss decrease in general: the compatibility factor κ_t is unavoidable in the selected protected geometry.

Theorem 2.2 (Exact finite counterfactual completion)**.** *Suppose the no-protection comparator and certified persistent base endpoint are both available, the finite requested logits are consistent at duplicate frozen addresses, the declared node and guard capacities are sufficient, support separation is certified, and the numerical endpoint checks pass. Then the joint base-plus-residual transaction satisfies: (i) the persistent base remains within its certified behavior budget and achieves protected descent; (ii) the deployed predictor equals the no-protection endpoint on every current-minibatch observation; (iii) every active finite protected output is unchanged; (iv) every unselected certified candidate is unchanged and the selected candidate reaches its declared finite target or contraction; and (v) the residual is exactly zero outside its finite support union and on the declared guards.*

Because the current deployed loss factors through the current-minibatch logits, Theorem 2.2 gives

$$\mathcal{L}_t(F_t) - \mathcal{L}_t(F_{t+1}) = \Delta_t^0, \qquad \frac{\mathcal{L}_t(F_t) - \mathcal{L}_t(F_{t+1})}{\Delta_t^0} = 1. \tag{12}$$

The complete finite construction, support bound, transfer result, and finite-precision realization are developed in Section 2.5.

2.3.1 Retention, stationarity, and local optimality

Combining the persistent and finite legs gives a one-step guarantee. For the accepted persistent displacement d_t^{safe} with $s_t^{\mathrm{safe}} = \|d_t^{\mathrm{safe}}\|$,

$$\|\Phi_t^{[S_t]}(\theta_t + d_t^{\mathrm{safe}}) - \Phi_t^{[S_t]}(\theta_t)\| \le (\varepsilon_t + e_t^{\mathrm{pop}}) s_t^{\mathrm{safe}} + \frac{\overline{H}_t}{2}(s_t^{\mathrm{safe}})^2 \le b_t, \tag{13}$$

$$\ell_t^{S_t}(\theta_t) - \ell_t^{S_t}(\theta_t + d_t^{\mathrm{safe}}) \ge \frac{1}{2} s_t^{\mathrm{safe}} \|\Pi_t g_t\|. \tag{14}$$

On the declared finite protected evidence, the stronger deployed statement is exact equality across accepted protected updates. On broader behavior maps, the theorem retains explicit structural-shield and routing-mismatch charges rather than asserting population preservation from finite evidence. Section 2.8 gives the active-interval bounds, release decomposition, arbitrary-stream projected-stationarity identity, and lifelong compatible-plasticity corollary.

The local protection geometry is also sharp at first order. For a behavior Jacobian J with singular values $\sigma_1 \ge \cdots \ge \sigma_d$, the smallest worst-case first-order leakage achievable by any $(d-r)$-dimensional plastic subspace is $\sigma_{r+1}(J)$. Thus finite-rank protection has an explicit spectral price; the result and the streamed-memory certificate are proved in Section 2.7.

2.3.2 Integrated frontier and unavoidable obstructions

The full AFM theorem combines pathwise retention, endpoint emulation, compatible stationarity, task-free routing, operational consolidation, metaplastic allocation, reopening, structural renewal, bounded structural resources, and a convex dynamic-regret specialization. Its deterministic conclusions hold on every realized trajectory whenever the corresponding certificates accept. Statistical conclusions, such as candidate-risk, route-refinement, or renewal-probability statements, additionally require only the assumptions stated for those conclusions and share a summable failure budget. The complete theorem and dependency proof are given in Section 2.11.

The result is obstruction-aware. Its scope includes explicit failure conditions for instances in which the required observable information, compatible motion, or bounded capacity is unavailable. Exact semantic routing cannot be guaranteed when observable laws are identical; future risk cannot be certified distribution-free from an arbitrary finite past; finite-rank protection cannot eliminate all functional leakage under nonzero movement; direct output contradictions cannot be simultaneously satisfied; reopening requires observational information; renewal requires compatible richness; and fixed finite memory cannot encode an unlimited number of independent facts. Section 2.12 states and proves these lower bounds. These lower bounds define the scope of the frontier.

The nonlinear and convex optimization statements have different scope. For smooth nonlinear models, the corresponding global-in-time optimization statement is the pathwise projected-stationarity identity of

Theorem 2.29, under its stated certificate conditions. The dynamic-regret guarantee in Section 2.11 requires the separate convex affine specialization and is not inherited by the neural experiments below. The experiments instead evaluate the nonlinear AFM transaction and its theorem-aligned quantities empirically.

2.4 AFM framework synthesis

Adaptive Functional Metaplasticity treats continual learning as a certified frontier between persistent compatible adaptation and exact finite deployed completion. Each protected update is referenced to the genuine same-state no-protection endpoint. Compatible motion is stored persistently under an explicit retention charge, with a quantitative loss guarantee that exposes the local compatibility factor; a separate bounded functional residual completes the finite current endpoint and restores declared protected outputs. The framework also integrates task-free routing, whole-behavior sensitivity memory, multiscale metaplastic allocation, operational consolidation, evidence-based reopening, structural renewal, bounded structural resources, and explicit obstruction conditions.

Across CORe50, CLEAR-10, and CLAD-C, the evaluated AFM frontier improves on the strongest tested classical non-oracle comparison family under the frozen five-seed protocols and remains competitive against six modern challengers under the stated common interface. The execution audit confirms that the finite endpoint transaction is non-vacuous and that the persistent component respects the theorem-aligned assimilation reference on the accepted nonzero controlled interventions. The causal experiments then isolate the compatibility quantity predicted by the framework and show that deliberately changing it changes retention-constrained persistent learning from matched neural states. The remaining limitations are explicit: finite evidence does not imply population retention, observable routing requires observable information, compatibility is not sufficient independently of retention allowance and finite curvature, and the present implementation incurs substantial computational overhead. Within this stated scope, the combined theory and experiments show that persistent learning and deployed endpoint preservation can be separated, quantified, causally interrogated, and coupled within one task-free continual-learning framework.

The sections that follow provide the full AFM mathematical specification, proofs, obstruction results, complete benchmark analyses, controlled causal study, and follow-up experiments supporting the paper.

2.5 Full AFM specification and finite endpoint construction

This section gives the complete mathematical state, endpoint construction, finite-support residual, and protected-update algorithm used by the results in the main text. It also records the precise conditions under which a protected transaction is rejected.

2.5.1 Admissible specification

Definition 2.3 (Admissible AFM specification)**.** Before the protected horizon begins, an AFM specification fixes all objects that can affect a protected decision. These comprise:

(i) a finite family of observable context-signature maps, their mixture weights, calibration prefix, requested coverage, calibration ceiling, routing thresholds, registry size, and deterministic tie rules;

(ii) a bounded outcome loss, a deterministic same-state no-protection update operator, the normalized coordinate η_t, the protected safe-base operator, the activation-gap threshold, and all trust, smoothness, curvature, and numerical-certification procedures;

(iii) a finite candidate-fitting procedure, validation horizon, first-crossing certification rule, post-certification staleness test, candidate-service priority, and service horizon;

(iv) the frozen-representation address map, finite shield support rule, duplicate-consistency convention, current-minibatch bound, shield-node capacity, guard capacity, selected-transfer fraction, and atomic rollback rule;

(v) the behavior-map constructors, protected-record and sketch capacities, finite rank-policy family, metaplastic trace bank, and all associated thresholds;

(vi) the shadow challenger, outcome-reopening and observable-signature route-refinement procedures, minimum effect and distinct-block requirements, diagnostic grace window, and capacity policy; and

(vii) a fixed pool of functionally zero-gated renewal modules, the renewal trial distribution, the certified numerical precision or finite-horizon precision policy, and all remaining resource budgets.

An optional finite unprotected initialization prefix may produce the initial representation but creates no protected record. After that prefix, the representation is frozen and any bounded prefix references used for signature calibration are replayed through the final representation before protected decisions begin. Every component must be measurable with respect to the declared learner history. Failure of a component to satisfy its own conditions yields the corresponding calibration, routing, statistical, functional, capacity, numerical, retention, compatibility, or renewal obstruction; the declared algorithm is unchanged.

2.5.2 Losses and retained behaviors

At round t, the learner has parameters $\theta_t \in \mathbb{R}^d$, observes an arbitrary data item, and incurs a differentiable loss $\ell_t : \mathbb{R}^d \to \mathbb{R}$ with gradient $g_t = \nabla \ell_t(\theta_t)$. A data item may be a declared finite ordered minibatch. In that case every member is routed in the fixed within-batch order, every consolidation and outcome-reopening statistic is updated at outcome resolution, every route-refinement statistic is updated once per distinct observable signature block, and ℓ_t is the declared differentiable minibatch loss used for the single safe parameter update. Each fixed signature map may assign one image-only signature to every member of a predeclared observable microblock, for example a normalized mean of frozen features, because the whole microblock is observed before its outcomes; labels, task identities, and evaluator metadata may not enter that signature. Repeating one block signature across several labeled members counts as one signature observation and several outcome observations. A semantic boundary inside such a block is not assumed away and contributes to the routing-mismatch terms. This is not an iid assumption: the ordered minibatch itself is one arbitrary stream element, and its size is bounded by the fixed specification. No stochastic or stationarity assumption is imposed on the loss stream. For the randomized allocation theorem, the current data and loss function are fixed before the controller draws p_t; they may otherwise depend arbitrarily on the past. Thus the allocation losses form a nonanticipating adaptive sequence rather than an adversary reacting to the current private coin.

Before any nonzero update, AFM must possess a trust radius $R_t^{\mathrm{cap}} \geq 0$ and finite upper certificates $\overline{L}_t, \overline{H}_t$ such that for every $d \in \mathrm{range}(\Pi_t)$ with $\|d\| \leq R_t^{\mathrm{cap}}$,

$$\ell_t(\theta_t + d) \leq \ell_t(\theta_t) + g_t^T d + \frac{\overline{L}_t}{2} \|d\|^2 , \tag{15}$$

$$\|\Phi_t(\theta_t + d) - \Phi_t(\theta_t) - J_t d\| \leq \frac{\overline{H}_t}{2} \|d\|^2 . \tag{16}$$

These are direct quadratic upper-model certificates; they do not require a globally existing Hessian. If no valid certificate is available, the algorithm sets $R_t^{\mathrm{cap}} = 0$ and makes no update. Hence the proof never defines smoothness or curvature after choosing the step. Losses are bounded below by $\ell_{\inf}$ on the realized iterates.

Proposition 2.4 (Constructible trust certificates for finite computation graphs)**.** *Suppose the current loss and every active empirical behavior map are represented by finite computation graphs whose primitive operations have computable interval enclosures for their first and second derivatives on a compact parameter box $\mathcal{B}_t$. Then interval automatic differentiation and the chain, product, and composition rules produce finite certified values $\overline{L}_t$, $\overline{H}_t$, and $L_{J,j,t}$ valid on every Euclidean ball contained in $\mathcal{B}_t$. For piecewise-affine primitives, any ball certified not to cross a switching surface has zero second-order remainder for that primitive. If interval propagation is unbounded or inconclusive, choosing $R_t^{\mathrm{cap}} = 0$ remains valid.*

Proof. Proceed in topological order through the finite graph. Interval arithmetic encloses every intermediate value. The derivative of each primitive is enclosed by hypothesis; applying the chain and product rules

recursively encloses the graph Jacobian and all directional second derivatives on the box. Taking operator-norm upper bounds gives the displayed certificates. A piecewise-affine primitive is affine on a switch-stable box, so its second derivative there is zero. □

The active protected records are indexed by $j \in \mathcal{A}_t$. Record j has a Fréchet-differentiable behavior map

$$\Phi_j : \mathbb{R}^d \to \mathcal{H}_j,$$

where $\mathcal{H}_j$ is a Hilbert space. It may represent an entire predictor in $L^2(P)$, a policy, value function, representation, or an empirical behavior over every observation in a frozen routed segment. In the empirical finite-evidence instantiation below, Φ_j is the fixed empirical commit-block map $\widehat{\Phi}_j$; a population map is substituted only when a valid population certificate is supplied. The stacked active map is

$$\Phi_t = \bigoplus_{j \in \mathcal{A}_t} \Phi_j, \qquad J_t = D\Phi_t(\theta_t).$$

2.5.3 Task-free routing and semantic mismatch

The algorithm receives a context signature Z_t but no task label. It uses the predictable bounded-registry router defined in Algorithm 2.12. The internal theorem protects the behaviors assigned by this router. To compare with any external semantic partition, embed the internal and evaluator behaviors in a common Hilbert space. For the joint base-shield state, let $\widetilde{\Psi}_t(\theta, S)$ be the evaluator's desired behavior map, let $\widetilde{\Phi}_t(\theta, S)$ be the router-selected map, and define the exact mismatch

$$\widetilde{\Delta}_t(\theta, S) = \widetilde{\Psi}_t(\theta, S) - \widetilde{\Phi}_t(\theta, S).$$

For the fixed-pre-shield safe-base leg define the realized routing derivatives

$$r_t = \left\| D_\theta \widetilde{\Delta}_t(\theta_t, S_t) \right\|_{\mathrm{op}}, \qquad k_t = \sup_{0 \le s \le 1} \left\| D_\theta^2 \widetilde{\Delta}_t(\theta_t + s d_t^{\mathrm{safe}}, S_t) \right\|_{\mathrm{op}}. \tag{17}$$

For the structural replacement define the separate exact mismatch charge

$$\omega_t^{\Delta} := \left\| \widetilde{\Delta}_t(\bar{\theta}_{t+1}, S_{t+1}) - \widetilde{\Delta}_t(\bar{\theta}_{t+1}, S_t) \right\|. \tag{18}$$

All three terms are zero when the routing semantics agree. They are not assumed small, and ω_t^{Δ} prevents a shield replacement from being hidden inside a parameter-only routing bound.

2.5.4 Streamed whole-behavior sensitivity memory

A protected segment j has a frozen anchor $\bar{\theta}_j$ and a frozen commit evidence block

$$\mathcal{D}_j = (x_{j,1}, \ldots, x_{j,n_j}).$$

Its empirical whole-behavior map is

$$\widehat{\Phi}_j(\theta) = n_j^{-1/2} \big(f_\theta(x_{j,1}), \ldots, f_\theta(x_{j,n_j}) \big).$$

This is the complete behavior on every observation in the commit block, not a hand-picked probe list. During the atomic commit, all anchor Jacobian blocks

$$A_j(x_{j,i}) = D_\theta f_\theta(x_{j,i}) \big|_{\theta = \bar{\theta}_j}$$

are streamed into a Frequent Directions sketch. At atomic commit, the sketch and its shrinkage certificate are rescaled by $n_j^{-1/2}$ and n_j^{-1} respectively, producing $B_j \in \mathbb{R}^{\ell_j \times d}$. The conceptual scaled row stack is

$$C_j = n_j^{-1/2} [A_j(x_{j,1}); \cdots ; A_j(x_{j,n_j})].$$

It is then fixed and need not be stored by the optimizer. In a certified population-map instantiation, raw observations may be discarded after the sketch and commit certificate are formed. In the empirical endpoint-check instantiation, the specification instead declares a fixed evidence capacity $n_{\max}$ and retains at most $n_{\max}$ fixed-size observation references (or the corresponding bounded tensors), transforms, labels, and anchor outputs per active segment so that the complete commit-block map can be reevaluated exactly. Later observations may create a new append-only segment, but never alter or replace the map protected by this segment. Explicit release deletes that segment's evidence and ends its future guarantee. For a segment arising from a certified candidate, let F_j^{snap} denote the complete frozen validated candidate predictor on its commit evidence block, including the structural state frozen with the candidate snapshot; in the base-shield notation introduced below this state is $(\bar{\theta}_j, S_j^{\mathrm{snap}})$. For an activated candidate the commit evidence is the frozen transfer evidence, so $\mathcal{D}_j = \mathcal{S}_j$ and $n_j = |\mathcal{S}_j|$. Because validation may take time while the global learner continues moving, let c_j be the atomic activation round and define the exactly observed *joint deployed activation-transfer gap*

$$a_j^{\mathrm{act}} = n_j^{-1/2} \left(\sum_{i=1}^{n_j} \left\| F_{c_j}(x_{j,i}) - F_j^{\mathrm{snap}}(x_{j,i}) \right\|^2 \right)^{1/2}. \tag{19}$$

Equivalently, this is the empirical Hilbert norm between the complete deployed state at activation and the complete frozen validated snapshot, not a base-parameter-only distance. The validation certificate belongs to the frozen snapshot $(\bar{\theta}_j, S_j^{\mathrm{snap}})$; the active-interval budget begins at the actually deployed state (θ_{c_j}, S_{c_j}). The transfer cost is charged explicitly by a_j^{act} plus subsequent certified motion, rather than by an unrelated current-batch risk gate. Merely pausing deployed learning after an off-path private fit does not make this gap zero; it is zero only if the validated behavior is already deployed. The transfer rule below therefore attempts to reduce the gap through ordinary protected updates and commits only after the measured gap lies below a threshold fixed before the stream. If no compatible transfer direction is found, the corresponding transfer obstruction is returned and the deployed predictor is unchanged.

2.5.5 Guarded compact-cardinal shielding of certified candidates

A candidate becomes a transfer target only at the first anytime-valid certification crossing defined in Section 2.6.2. At that stopping time, freeze the complete bounded commit evidence block $\mathcal{S}_j$, the final-frozen representation $z(x)$ of every member, the deployed logits $F_{\tau_j}(x)$, the frozen candidate structural state S_j^{snap}, and the complete candidate-snapshot logits $Y_j(x) = F_j^{\mathrm{snap}}(x) = G_{\bar{\theta}_j}(z(x)) + S_j^{\mathrm{snap}}(z(x))$. The fixed empirical transfer objective remains

$$D_j(F) = \frac{1}{2|\mathcal{S}_j|} \sum_{x \in \mathcal{S}_j} \|F(x) - Y_j(x)\|^2. \tag{20}$$

The private parameter vector is never copied into deployment. A predictable bounded least-recently-served rule selects one certified candidate s_t for service; every other certified candidate remains a safeguard. While certified, each candidate reserves its bounded route slot and frozen evidence. It leaves the certified set only by atomic commitment, a separately controlled staleness crossing, or an explicitly declared experiment termination, which carries no completion claim.

The deployed logits are decomposed as

$$F_t(x) = G_{\theta_t}(z(x)) + S_t(z(x)), \tag{21}$$

where G_θ is the ordinary trainable base map and S_t is bounded structural state over the final frozen representation z. For any behavior constructor $\mathcal{B}_j$, write

$$\Phi_j^{[S]}(\theta) = \mathcal{B}_j(G_\theta + S), \qquad \widetilde{\Phi}_j(\theta, S) = \mathcal{B}_j(G_\theta + S).$$

The bracket in $\Phi_j^{[S]}$ means that the structural shield is held fixed while the persistent parameters move. The streamed Jacobian memory, compatible projector, safe radius, and endpoint check govern $\Phi_j^{[S_t]}$ during the safe-base proposal; the compact-cardinal transaction then replaces S_t by S_{t+1}. The joint theorem accounts for both legs rather than treating the shield replacement as invisible structural state. For the current minibatch

define analogously $\ell_t^S(\theta)$ as the declared loss of $G_\theta + S$ with S held fixed, and set $g_t = \nabla \ell_t^{S_t}(\theta_t)$. In the joint exact-endpoint-emulation mode, this current-batch loss is required to factor through the finite deployed minibatch logits: there is a bounded loss functional $\mathcal{L}_t$ such that

$$\ell_t^S(\theta) = \mathcal{L}_t((G_\theta(z(x)) + S(z(x)))_{x\in\mathcal{B}_t}). \tag{22}$$

For any deployed predictor F, write $\mathcal{L}_t(F) := \mathcal{L}_t((F(x))_{x\in\mathcal{B}_t})$. Hence equality of all current-minibatch logits implies equality of the deployed current loss. A direct parameter regularizer, if used in some separate optimization objective, is not part of the exact endpoint-emulation ratio unless it too factors through these declared minibatch outputs.

Genuine ordinary counterfactual. Let U_t^0 be the complete deterministic update operator used by the declared no-protection learner: it receives the same pre-step deployed state, current minibatch $\mathcal{B}_t$, active-coordinate mask, learning-rate and backtracking rules, renewal trial, and private random choices, but no protected basis, behavior budget, candidate constraint, or retention radius. Applied to the current state, it either abstains or returns base parameters θ_t^0 and endpoint logits

$$U_t(x) = G_{\theta_t^0}(z(x)) + S_t(z(x)), \qquad x \in \mathcal{B}_t. \tag{23}$$

Its exact endpoint decrease is

$$\Delta_t^0 = \mathcal{L}_t(F_t) - \mathcal{L}_t(U_t) = \ell_t^{S_t}(\theta_t) - \ell_t^{S_t}(\theta_t^0). \tag{24}$$

This is the comparator used by the counterfactual-normalized mode. It is not projected into the protected nullspace and is not restricted by the retention-safe radius. If U_t^0 has no accepted nonzero endpoint, AFM reports that the exact counterfactual endpoint is unavailable; it does not substitute a weaker denominator.

Persistent metaplastic safe endpoint. Using the selected policy basis Q_t, compatible projector Π_t, behavior budget b_t, leakage certificate E_t, curvature certificate $\overline{H}_t$, trust cap, and the same declared backtracking rule, let $\mathsf{U}_t^{\mathrm{safe}}$ return either an accepted base endpoint

$$\bar\theta_{t+1} = \theta_t + d_t^{\mathrm{safe}}, \qquad d_t^{\mathrm{safe}} \in \mathrm{range}(\Pi_t), \qquad \left\|d_t^{\mathrm{safe}}\right\| \le R_t,$$

or reports that a certified safe-base endpoint is unavailable. This is the only persistent base endpoint in the joint mode. The unrestricted endpoint θ_t^0 is never installed. Every backtracking candidate for both operators is evaluated from the identical pre-step predictive transaction state; parameters, mutable buffers, current shield, module modes, and declared private randomness. Endpoint probing restores that state before every candidate factor and after each operator; only the complete accepted predictive state of the safe-base operator is eligible for commitment.

Counterfactual-normalized persistent assimilation. The general endpoint-emulation oracle below remains valid for any deterministic comparator. The stronger persistent-assimilation mode considered here is an auditable scalar-gradient specialization. Let A_t be the active-coordinate projector already contained in Π_t, put

$$a_t = A_t g_t, \qquad q_t = \Pi_t g_t,$$

and suppose the accepted ordinary comparator displacement is

$$d_t^0 = -\alpha_t a_t, \qquad 0 < \alpha_t \le \overline{L}_t^{-1}. \tag{25}$$

Its feasible persistent reference is

$$v_t = \Pi_t d_t^0 = -\alpha_t q_t, \qquad s_t^0 = \|v_t\|. \tag{26}$$

Define the certified retention charge

$$C_t(s) = E_t s + \frac{\overline{H}_t}{2} s^2.$$

For a predeclared charge coordinate $\eta_t \in [0,1]$, the normalized round budget is

$$b_t^{\mathrm{norm}} = \eta_t C_t(s_t^0), \tag{27}$$

and the maximal certified reference-path fraction is

$$\lambda_t = \max\{\lambda \in [0,1] : C_t(\lambda s_t^0) \le b_t^{\mathrm{norm}}\}. \tag{28}$$

If there is no active protected behavior, set $\lambda_t = 1$. An endpoint backtrack may reduce this to $\widehat{\lambda}_t$; the normalized mode accepts only when $\widehat{\lambda}_t \ge \eta_t$. Comparator collinearity, projection idempotence, trust-cap membership, persistent descent, and the realized charge are checked explicitly. Failure returns a typed obstruction, with the declared path and denominator unchanged.

Theorem 2.5 (Counterfactual-normalized persistent assimilation)**.** *Assume* (25)*, A_t and Π_t are orthogonal projectors with $\Pi_t A_t = \Pi_t$, the accepted comparator has $\Delta_t^0 > 0$, and $s_t^0 > 0$. Assume further that the same fixed-shield current loss admits the stated $\overline{L}_t$ quadratic smoothness bound on both complete line segments $\{\theta_t + s d_t^0 : 0 \le s \le 1\}$ and $\{\theta_t + s v_t : 0 \le s \le 1\}$, and that the retention charge $C_t(s)$ is certified for every projected-reference displacement $s v_t / s_t^0$ with $0 \le s \le s_t^0$. In particular, the full projected reference and every accepted realized backtrack lie inside their respective certified domains. Then the normalized construction satisfies*

$$\lambda_t \ge \eta_t, \tag{29}$$

$$C_t(\lambda_t s_t^0) \le b_t^{\mathrm{norm}}, \tag{30}$$

and $\eta_t = 1$ gives $\lambda_t = 1$, hence the complete projected comparator v_t. For any accepted realized fraction $\widehat{\lambda}_t \ge \eta_t$, define

$$\kappa_t = \frac{\|q_t\|^2}{\|a_t\|^2} \in [0,1].$$

Writing

$$\Delta_t^{\mathrm{base}} = \ell_t^{S_t}(\theta_t) - \ell_t^{S_t}(\theta_t + \widehat{\lambda}_t v_t),$$

one has

$$\frac{\Delta_t^{\mathrm{base}}}{\Delta_t^0} \ge \widehat{\lambda}_t \kappa_t \frac{1 - \frac{1}{2}\alpha_t \overline{L}_t \widehat{\lambda}_t}{1 + \frac{1}{2}\alpha_t \overline{L}_t} \ge \frac{\widehat{\lambda}_t \kappa_t}{3} \ge \frac{\eta_t \kappa_t}{3}. \tag{31}$$

After the compact-cardinal residual is installed, the deployed prediction-loss current-batch ratio remains exactly one by (38)*. Thus η_t governs persistent assimilation, κ_t displays the unavoidable local compatibility price, and the shield supplies only the residual endpoint correction. If $s_t^0 = 0$, $\kappa_t = 0$, the comparator is not scalar-gradient aligned, or any required certificate or endpoint inequality fails, AFM returns the corresponding compatibility, alignment, retention, descent, or numerical obstruction atomically.*

Proof. The function C_t is nonnegative, convex, nondecreasing on $[0,\infty)$, and satisfies $C_t(0) = 0$. Hence $C_t(\eta_t s_t^0) \le \eta_t C_t(s_t^0) = b_t^{\mathrm{norm}}$, proving (29); maximality gives (30), and $\eta_t = 1$ makes the endpoint feasible with equality. Smoothness, $\Pi_t A_t = \Pi_t$, and (25) give

$$\Delta_t^{\mathrm{base}} \ge \widehat{\lambda}_t \alpha_t \|q_t\|^2 - \frac{\overline{L}_t}{2} \widehat{\lambda}_t^2 \alpha_t^2 \|q_t\|^2 .$$

The reverse smoothness inequality gives

$$\Delta_t^0 \le \alpha_t \|a_t\|^2 + \frac{\overline{L}_t}{2} \alpha_t^2 \|a_t\|^2 .$$

Dividing yields the first quotient in (31); since $0 < \alpha_t \overline{L}_t \le 1$ and $0 < \widehat{\lambda}_t \le 1$, its scalar factor is at least $1/3$. The compact-cardinal theorem then restores the exact comparator logits on the current finite minibatch without changing the committed base displacement. □

Provisionally move the complete accepted safe-base predictive state to $\bar{\theta}_{t+1}$. Form one finite desired-output multiset. On the current minibatch require the exact counterfactual logits U_t; on every active protected evidence block and every unselected certified candidate block require the pre-update deployed logits; and on the selected candidate block require its frozen target or the predeclared contraction toward it:

$$q_t(x) = \begin{cases} U_t(x), & x \in \mathcal{B}_t, \\ F_t(x), & x \in \mathcal{P}_t, \\ F_t(x), & x \in \mathcal{S}_j,\ j \in \mathcal{C}_t \setminus \{s_t\}, \\ (1-\xi_t)F_t(x) + \xi_t Y_{s_t}(x), & x \in \mathcal{S}_{s_t}, \end{cases} \qquad \xi_t \in (0,1]. \tag{32}$$

Here $\mathcal{P}_t$ is the union of the complete bounded active empirical evidence blocks; the evaluated construction uses $\xi_t = 1$. The same restoration system is executed on every round with an active protected record, whether or not a candidate is currently served.

To remove arbitrary feature scale, use the fixed injective address map

$$\iota(z) = \frac{z}{\sqrt{1 + \|z\|^2}}, \qquad \iota : \mathbb{R}^{d_Z} \longrightarrow \{u : \|u\| < 1\}. \tag{33}$$

If one addressed feature occurs in several constraint rows, those rows are merged only when their requested logits agree under the declared exact or outward-certified equality convention. Distinct requests at the same address are a functional inconsistency: no deterministic predictor depending only on the frozen representation can satisfy them simultaneously.

The specification declares a bounded label-free guard bank $\mathcal{Q}_t$ of at most $Q_{\max}$ frozen-backbone addresses. Guards use no labels, evaluator data, task identity, session identity, or boundary metadata. Let $z_1, \dots, z_N$ be the merged distinct raw frozen-representation constraint nodes, let

$$u_i = \iota(z_i)$$

be their injective compact addresses, and let q_i be the corresponding requested logits from (32). Define

$$R_i = q_i - G_{\bar{\theta}_{t+1}}(z_i)$$

to be the required residual logits after the certified metaplastic safe-base update. The base map is always evaluated at the raw frozen representation z_i; u_i is used only as the compact support address. A predeclared replay envelope $\varepsilon_a > 0$ and multiplier $\kappa > 1$ determine $r_i = \kappa\varepsilon_a$; the evaluated construction uses $\kappa = 4$. Deployment requires every nonmatching center or guard to be farther than $2r_i$. Define the C^1 plateau-cardinal bump

$$\beta_{r,\varepsilon_a}(u;c) = \begin{cases} 1, & \|u-c\| \le \varepsilon_a, \\ \left(1 - \left(\dfrac{\|u-c\| - \varepsilon_a}{r - \varepsilon_a}\right)^2\right)^2, & \varepsilon_a < \|u-c\| < r, \\ 0, & \|u-c\| \ge r, \end{cases}$$

and atomically replace the shield by

$$S_{t+1}(z) = \sum_{i=1}^{N} \beta_{r_i,\varepsilon_a}(\iota(z); u_i) R_i. \tag{34}$$

There is no fitted interpolation coefficient or linear solve. The node count obeys the fixed capacity

$$N \le B_{\max} + J_{\max} n_{\max} + C_{\max} n_{\max}. \tag{35}$$

The projected safe-radius update is the persistent base update in counterfactual-normalized mode. If either the safe-base or shield transaction fails, the update is rejected; the unrestricted endpoint is not substituted for the protected update.

Theorem 2.6 (Joint safe-base counterfactual-emulation oracle and sharp obstruction)**.** *Assume that the genuine comparator U_t^0 returns an accepted endpoint with $\Delta_t^0 > 0$, the metaplastic safe-base operator returns $\bar{\theta}_{t+1} = \theta_t + d_t^{\mathrm{safe}}$ with*

$$d_t^{\mathrm{safe}} \in \mathrm{range}(\Pi_t), \qquad \left\|d_t^{\mathrm{safe}}\right\| \le R_t,$$

and its complete bounded base-behavior and loss endpoint checks pass. Assume also that the requested logits in (32) *are consistent on duplicate frozen addresses, the node capacity is sufficient, pairwise support and guard separation are outward-certified, and endpoint arithmetic is certified. Then the atomic safe-base-plus-shield update exists and satisfies:*

(i) certified persistent base retention and descent,

$$\left\|\Phi_t^{[S_t]}(\bar{\theta}_{t+1}) - \Phi_t^{[S_t]}(\theta_t)\right\| \le b_t, \tag{36}$$

$$\ell_t^{S_t}(\bar{\theta}_{t+1}) \le \ell_t^{S_t}(\theta_t) - \frac{1}{2} s_t^{\mathrm{safe}} \left\|\Pi_t g_t\right\|; \tag{37}$$

(ii) exact current counterfactual equality, $F_{t+1}(x) = U_t(x)$ for every $x \in \mathcal{B}_t$;

(iii) exact finite protected retention, $F_{t+1}(x) = F_t(x)$ for every $x \in \mathcal{P}_t$;

(iv) exact unselected-candidate safeguards and the selected contraction in (32)*;*

(v) pairwise-disjoint support, exact guard silence, and exact zero residual outside the finite support union;

(vi) the global pointwise bound $\|S_{t+1}(z)\| \le \max_i \|R_i\|$.

Here $\ell_t^{S_t}(\theta)$ is the current loss with the pre-step shield held fixed. Consequently

$$\mathcal{L}_t(F_t) - \mathcal{L}_t(F_{t+1}) = \Delta_t^0, \qquad \frac{\mathcal{L}_t(F_t) - \mathcal{L}_t(F_{t+1})}{\Delta_t^0} = 1 \ge \rho_{\mathrm{tr}}. \tag{38}$$

For any broader deployed behavior map, define the exact structural charge

$$\omega_t^S := \left\|\widetilde{\Phi}_t(\bar{\theta}_{t+1}, S_{t+1}) - \widetilde{\Phi}_t(\bar{\theta}_{t+1}, S_t)\right\|. \tag{39}$$

Then

$$\left\|\widetilde{\Phi}_t(\bar{\theta}_{t+1}, S_{t+1}) - \widetilde{\Phi}_t(\theta_t, S_t)\right\| \le b_t + \omega_t^S. \tag{40}$$

For the declared finite protected-evidence map, item (iii) gives the stronger exact value zero rather than this generic upper bound. If any comparator, safe-base, consistency, capacity, address-resolution, or numerical assumption fails, AFM restores every predictive component touched by the transaction; parameters, mutable model buffers, shield state, module modes, and declared private randomness; and emits the corresponding typed obstruction. Structural route and renewal proposals are committed only after this transaction accepts. Monotone audit and service counters retain the failed attempt. A rejected protected transaction does not install the unrestricted endpoint.

Proof. The safe-base endpoint is an accepted projected step in $\mathrm{range}(\Pi_t)$ with norm at most the retention-safe radius. The streamed-memory certificate and the declared curvature upper model therefore prove (36); the safe-base current-loss endpoint check proves (37). At each merged raw node z_i, its compact address is $u_i = \iota(z_i)$, so compact cardinality and support separation give $\beta_i(\iota(z_i)) = 1$ and $\beta_h(\iota(z_i)) = 0$ for $h \ne i$. Hence $S_{t+1}(z_i) = R_i$, and because the residual coefficient is computed relative to the base evaluated at the same raw node,

$$G_{\bar{\theta}_{t+1}}(z_i) + S_{t+1}(z_i) = G_{\bar{\theta}_{t+1}}(z_i) + R_i = q_i.$$

The four choices of q_i prove current equality, protected equality, candidate safeguards, and selected transfer. By (22), current-minibatch logit equality with U_t proves (38). Compact support, guard exclusion, and the pointwise bound follow because at most one bump is active at an address. Equation (40) is the triangle inequality applied first to the persistent base move with S_t fixed and then to the structural shield replacement. Duplicate contradiction is logically impossible for a deterministic address map; every other failed certificate triggers the declared full-state rollback. □

Proposition 2.7 (Conservative finite-precision realization)**.** *A theorem-certified finite-precision implementation deploys exact-counterfactual restoration only when outward-certified arithmetic proves: availability and acceptance of the counterfactual endpoint; duplicate consistency; node and guard capacities; a certified replay envelope; unique replay-address ownership; pairwise center and guard separation; bounded cardinal residual arithmetic; and the current, protected, candidate, selected, and guard endpoint inequalities after the declared two-sided numerical envelope is added. Empirical endpoint mode may use floating-point calculations and complete reevaluation on all bounded evidence and guards, but it is labeled a falsifiable empirical instantiation rather than pre-step theorem certification. Any failed check restores both the base parameters and shield.*

Proposition 2.8 (Full genuine ordinary progress and zero candidate damage)**.** *On every accepted joint safe-base counterfactual-emulation step,*

$$\mathcal{L}_t(F_t) - \mathcal{L}_t(F_{t+1}) = \Delta_t^0 \geq \rho_{\mathrm{tr}} \Delta_t^0, \tag{41}$$

$$D_j(F_t) - D_j(F_{t+1}) = 0 \quad \forall j \in \mathcal{C}_t \setminus \{s_t\}, \tag{42}$$

$$D_{s_t}(F_{t+1}) = (1-\xi_t)^2 D_{s_t}(F_t), \tag{43}$$

the persistent base endpoint satisfies its declared behavior budget, and every declared active finite deployed protected behavior is unchanged exactly. With $\xi_t = 1$, the selected finite transfer objective is zero after one accepted service. Thus the unchanged numerical value $\rho_{\mathrm{tr}} = 0.25$ is measured against the genuine no-protection endpoint, and the construction actually attains ratio one.

Corollary 2.9 (Budget-controlled persistence off support)**.** *At every address outside the new compact support union,*

$$F_{t+1}(x) = G_{\bar{\theta}_{t+1}}(z(x)).$$

Hence future off-support behavior follows the certified metaplastic base endpoint, not the unrestricted no-protection endpoint. Different rank or behavior-budget policies may therefore induce different persistent trajectories even though all accepted variants reproduce the same genuine no-protection logits on the finite current minibatch. Equality with the no-protection base trajectory occurs only in the special case $\bar{\theta}_{t+1} = \theta_t^0$.

For every certified candidate, define the exact deployed ledger increment

$$\Delta_{j,t}^D = D_j(F_t) - D_j(F_{t+1}).$$

Accepted compact-cardinal steps and all other certified-candidate-safe updates satisfy $\Delta_{j,t}^D \geq 0$ under the single declared numerical envelope. Hence

$$D_j(F_T) = D_j(F_{\tau_j}) - \sum_{t=\tau_j}^{T-1} \Delta_{j,t}^D, \qquad \Delta_{j,t}^D \geq 0. \tag{44}$$

An implementation must snapshot and roll back the base parameters, deployed shield, support radii, guard state, and frozen candidate and record shield states atomically. The activation requirement $a_j^{\mathrm{act}} \leq \tau_j^{\mathrm{act}}$ is equivalent on the same block to $D_j(F) \leq (\tau_j^{\mathrm{act}})^2/2$.

Proposition 2.10 (Predictable non-starving transfer service)**.** *Suppose at most $C_{\max}$ certified candidates coexist. Initialize each candidate's service time to its certification round and update it at every actual service round. If one candidate is selected at each available transfer round by the least-recently-served rule, every continuously eligible candidate is selected at least once in every $C_{\max}$ such rounds, even when later candidates arrive. The rule is predictable and uses no task, route, episode, boundary, or evaluator metadata.*

Proof. Fix a continuously eligible candidate j and its current timestamp. A later arrival receives a later initial timestamp. Each competing candidate selected ahead of j receives the current, hence larger, timestamp. At most $C_{\max} - 1$ competitors coexist, so deterministic tie breaking selects j after at most that many other selections. □

Theorem 2.11 (Finite compact-cardinal transfer completion without a compatibility-rate assumption)**.** *Let candidate j remain certified and non-stale until its next available service round. If the finite constraints for*

that service are functionally consistent, lie within the declared node and guard capacities, and the address and endpoint arithmetic are certified, then the choice $\xi = 1$ gives

$$D_j(F_{t+1}) = 0 \tag{45}$$

and therefore passes every nonnegative predeclared activation-gap threshold on the same frozen evidence block after that one accepted service. Under Proposition 2.10, this occurs within at most $C_{\max}$ available service rounds. For a predeclared $\xi \in (0,1)$, after N accepted services,

$$D_{j,N} \le (1-\xi)^{2N} D_{j,0}, \qquad N \ge \left\lceil \frac{\log(D_j^{\mathrm{act}}/D_{j,0})}{2\log(1-\xi)} \right\rceil \tag{46}$$

suffices whenever $D_{j,0} > D_j^{\mathrm{act}}$. There is no projected-PL, positive-cosine, inter-service-damage, or interpolation-conditioning assumption. Failure can occur only through a declared statistical, exact functional-consistency, capacity, address-certification, endpoint-certification, or service-availability obstruction.

Proof. The selected identity in Proposition 2.8 gives one-step completion for $\xi = 1$ and geometric contraction otherwise. Proposition 2.10 supplies the service bound. The ledger (44) prevents intervening accepted updates from undoing the progress. □

This theorem concerns finite empirical function-space constraints; population retention requires the additional assumptions stated separately below. In contrast to a globally supported residual construction, the compact-cardinal oracle proves exactly where the structural residual can act, proves silence on a bounded label-free guard set, and prevents accumulation from multiple centers. Behavior inside an unguarded compact support but away from its center remains governed by the displayed pointwise residual bound and the separate routing, Jacobian, curvature, and evaluator statements below. If a current, protected, or candidate requirement requests contradictory logits at exactly the same observable address, no deterministic continuation of that representation can satisfy all of them; AFM must abstain or invoke a separately proved representation-renewal mechanism; no accepted transaction is formed from inconsistent requests.

At operational commitment AFM additionally requires the observed gap (19) to satisfy $a_j^{\mathrm{act}} \le \tau_j^{\mathrm{act}}$. This requirement does not replace the validation certificate or enlarge a retention budget. The private vector is not copied into deployment.

Frequent Directions maintains an observable shrinkage certificate Δ_j^{FD} such that, deterministically for every $v \in \mathbb{R}^d$,

$$0 \le \|C_j v\|^2 - \|B_j v\|^2 \le \Delta_j^{\mathrm{FD}} \|v\|^2 . \tag{47}$$

Moreover, for every $k < \ell_j$,

$$\Delta_j^{\mathrm{FD}} \le \frac{\|C_j - (C_j)_k\|_F^2}{\ell_j - k}, \tag{48}$$

which is the standard deterministic Frequent Directions guarantee Ghashami et al. (2016).

At the start of round t, before the current gradient is revealed, the admissible specification also fixes a predictable orthogonal *structural-availability projector* A_t. Its range contains exactly the parameter directions available to the optimizer in that round: ordinary active coordinates and, during a renewal trial, the named reset slot; dormant nontrial coordinates are excluded. Define

$$g_t^A = A_t g_t, \qquad \widetilde{B}_{j,t} = B_j A_t. \tag{49}$$

This projector preserves the declared committed protection geometry within the bounded architecture. It makes that architecture part of the mathematical algorithm: every leakage calculation and every policy comparison is performed on the same directions in which an update is actually permitted.

Each allocation policy is a fixed pair $p = (\beta^p, r^p)$ with $\beta_k^p \ge 0$, $\sum_k \beta_k^p \le 1$, and $0 \le r^p \le r_{\max}$. The specification also supplies a predictable frontier price $\zeta_t \in [0, \zeta_{\max}]$ that values current plasticity relative to

retention leakage. The declared family may include the normalized scale-free choices

$$\beta_k^{(\alpha)} = \frac{\rho_k^{\alpha-1}}{\sum_{h=0}^{K-1} \rho_h^{\alpha-1}}, \qquad 0 < \alpha \le 1,$$

for a fixed finite grid of α values and ranks. Before the current importance signal is revealed a policy predicts

$$\widehat{u}_{j,t}^{(p)} = 1 + \sum_{k=0}^{K-1} \beta_k^p z_{j,t}^{(k)} \in [1, 2]$$

and forms

$$\widehat{S}_{p,t} = \sum_{j \in \mathcal{A}_t} \widehat{u}_{j,t}^{(p)} \widetilde{B}_{j,t}^T \widetilde{B}_{j,t}. \tag{50}$$

It chooses $Q_{p,t}$ as the top-r^p nonzero eigenspace within range(A_t), with the fixed deterministic eigenvector/tie convention from the admissible specification; if the available covariance has rank below r^p, all of its nonzero eigendirections are used. After g_t is observed, define the explicit realized importance

$$\chi_{j,t} = \min\left\{1, \frac{\left\|\widetilde{B}_{j,t} g_t^A\right\|^2}{1 + \left\|g_t^A\right\|^2}\right\}, \qquad w_{j,t} = 1 + \chi_{j,t} \in [1, 2], \tag{51}$$

and the realized weighted covariance

$$S_t^* = \sum_{j \in \mathcal{A}_t} w_{j,t} \widetilde{B}_{j,t}^T \widetilde{B}_{j,t}. \tag{52}$$

For every counterfactual policy define its current blocked-gradient fraction

$$c_{p,t}^{\text{block}} = \frac{\left\|Q_{p,t} g_t^A\right\|^2}{1 + \left\|g_t^A\right\|^2} \in [0, 1]. \tag{53}$$

The controller has already selected p_t and uses $Q_t = Q_{p_t,t}$. Because range$(Q_t) \subseteq$ range(A_t), the effective compatible projector

$$\Pi_t = A_t - Q_t \tag{54}$$

is the orthogonal projector onto the structurally available directions orthogonal to the selected protected subspace. The observable spectral-allocation residual used by the optimizer is exactly

$$a_t^{\text{spec}} = \operatorname{tr}\big(\Pi_t S_t^*\big) = \operatorname{tr}\big((I - Q_t) S_t^*\big). \tag{55}$$

The pair $(a_t^{\text{spec}}, c_{p_t,t}^{\text{block}})$ is the realized local stability-plasticity frontier: the first measures protected sensitivity left in plastic directions, while the second measures current descent energy removed by protection. The anchor-drift term is

$$d_t^{\text{anc}} = \left(\sum_{j \in \mathcal{A}_t} w_{j,t} L_{J,j,t}^2 \left\|\theta_t - \bar{\theta}_j\right\|^2\right)^{1/2}, \tag{56}$$

where $L_{J,j,t}$ is any valid Lipschitz bound for the fixed segment map's stacked Jacobian between $\bar{\theta}_j$ and θ_t. The certified empirical leakage is

$$\varepsilon_t = \sqrt{a_t^{\text{spec}} + \sum_{j \in \mathcal{A}_t} w_{j,t} \Delta_j^{\text{FD}}} + d_t^{\text{anc}}. \tag{57}$$

If a population map Φ_j is desired instead of $\widehat{\Phi}_j$, let e_t^{pop} be a valid upper certificate for the population-empirical derivative discrepancy on range(Π_t). The empirical finite-evidence instantiation is empirical protection, for which $e_t^{\text{pop}} = 0$; a population guarantee requires a separately justified certificate, for example from matrix concentration under declared sampling conditions.

Given $E_t = \varepsilon_t + e_t^{\mathrm{pop}}$, behavior budget $b_t \geq 0$, certified curvature $\overline{H}_t \geq 0$, and trust cap R_t^{cap}, define the retention-safe radius

$$\widetilde{R}_t = \begin{cases} +\infty, & E_t = \overline{H}_t = 0, \\ 0, & b_t = 0,\ E_t + \overline{H}_t > 0, \\ \dfrac{2b_t}{E_t + \sqrt{E_t^2 + 2\overline{H}_t b_t}}, & b_t > 0,\ \overline{H}_t > 0, \\ \dfrac{b_t}{E_t}, & b_t > 0,\ \overline{H}_t = 0,\ E_t > 0, \end{cases} \qquad R_t = \min\{R_t^{\mathrm{cap}}, \widetilde{R}_t\}. \tag{58}$$

It is the largest radius certified by the quadratic upper model $E_t R + \overline{H}_t R^2/2 \leq b_t$, before the independent trust cap is applied.

2.5.6 Conflict, variation, evidence, and renewal richness

The compatible gradient is

$$q_t = \Pi_t g_t. \tag{59}$$

The discarded component $(I - \Pi_t)g_t$ measures local first-order conflict with the structurally available protected complement, including both protection and declared structural unavailability.

For the non-convex pathwise theorem, define realized one-sided variation

$$\nu_t = [\ell_{t+1}(\theta_{t+1}) - \ell_t(\theta_{t+1})]_+, \qquad V_T = \sum_{t=1}^{T-1} \nu_t. \tag{60}$$

For reopening record j, let the fixed predictable challenger procedure output a shadow prediction before the nth routed outcome. Let $Y_{j,n}^{\mathrm{rec}}, Y_{j,n}^{\mathrm{chal}} \in [0,1]$ be the subsequently observed losses of the committed snapshot and challenger, and fix a hysteresis margin $\kappa_j \in [0,1]$. AFM uses the concrete observable increment

$$X_{j,n} = Y_{j,n}^{\mathrm{rec}} - Y_{j,n}^{\mathrm{chal}} - \kappa_j. \tag{61}$$

The operational validity null is

$$\mathbb{E}[X_{j,n} \mid \mathcal{F}_{n-1}] \leq 0. \tag{62}$$

Because $X_{j,n}$ lies in an interval of length two, its centered noise is conditionally 1-sub-Gaussian by Hoeffding's lemma; a smaller certified parameter may be used. Under change, define the actual observable challenger advantage

$$\mu_{j,n} = \mathbb{E}[X_{j,n} \mid \mathcal{F}_{n-1}]. \tag{63}$$

It may be zero, positive, negative, or time-varying. Thus reopening uses raw routed outcomes and a predictable comparator, not a semantic oracle. A real semantic change that no declared challenger can expose has zero observable information and is correctly covered by the impossibility term.

A renewal trial resets a zero-gated batch and recomputes the *full* current loss gradient $g_{e,n}^{\mathrm{reset}}$ and the same global compatible direction used by the ordinary optimizer,

$$q_{e,n}^{\mathrm{reset}} = \Pi_{e,n} g_{e,n}^{\mathrm{reset}}.$$

Let $M_{e,n}$ be the predictable orthogonal coordinate projector onto the named trial slot, with $M_{e,n}A_{e,n} = M_{e,n}$. An eligible dormant slot satisfies the declared zero-gate invariant: while its functional gate is zero, the current loss and every active protected behavior have zero derivative with respect to its internal nongate coordinates; a previously activated slot is returned to the renewal pool only by a separately certified exact-dormancy operation that re-establishes this invariant. Because the protected covariance has zero columns in those dormant internal coordinates and the current gradient has zero components there, the compatible projected gradient has zero dormant-internal component as well. Thus any nonzero component selected by $M_{e,n}$

contains nonzero gate-coordinate motion and is functionally activating after an accepted step. The trial-slot component of the global compatible gradient is

$$r_{e,n} = M_{e,n} q_{e,n}^{\mathrm{reset}}.$$

A trial is γ-useful when $\|r_{e,n}\| \geq \gamma$. Its exact conditional richness is

$$\rho_{e,n}(\gamma) = \mathbb{P}\big(\big\|M_{e,n}\Pi_{e,n} g_{e,n}^{\mathrm{reset}}\big\| \geq \gamma \mid \mathcal{F}_{e,n-1}\big)\,. \tag{64}$$

No positivity is assumed. The safe update is always computed from the full direction $q_{e,n}^{\mathrm{reset}}$; $r_{e,n}$ is the exact diagnostic for whether that accepted update moves the renewed slot. Reset-induced changes in the current functional Jacobian are charged through the recomputed anchor-drift and population-discrepancy terms in the same leakage certificate (57).

2.5.7 Complete AFM algorithm

Algorithm 2.12 (Adaptive Functional Metaplasticity)**.** Fix the finite structural capacities and summable statistical error allocations in Definition 2.3. If an unprotected initialization prefix is used, complete it, freeze the representation, replay the bounded calibration references through that representation, and either certify the requested routing calibration or record that no positive routing conclusion is available. For each protected round t:

1. **Route.** Compute only the predeclared observable signatures and assign the current observations to the bounded registry using the fixed threshold and tie rule. Capacity overflow follows the predeclared release or unprotected-overflow policy; no evaluator task identity is used.

2. **Fit, certify, and service candidates.** Update finite private fitting blocks. Once a candidate is frozen, validate it only on later outcomes with a fresh summable error allocation. At the first certification crossing, freeze its evidence, outputs, and transfer target; thereafter update a separately allocated staleness process. Certified candidates are serviced by the predictable non-starving rule, and private parameters are never copied directly into deployment.

3. **Maintain append-only protected records.** At commitment, freeze the record's evidence, anchor, behavior map, and bounded sensitivity sketch. Re-anchoring creates a new segment rather than rewriting an active one. When structural capacity is exhausted, only the predeclared explicit release or overflow rule may remove a guarantee.

4. **Allocate protection.** Before the current importance signal is disclosed to the controller, select a policy from the finite policy family, form the predicted protected covariance, and choose its declared rank. After the current gradient is observed, update the realized importance signals, frontier losses, policy weights, and the metaplastic traces

$$z_{j,t+1}^{(k)} = (1-\rho_k) z_{j,t}^{(k)} + \rho_k \chi_{j,t}, \qquad \rho_k = 2^{-(k+1)}. \tag{65}$$

   New traces are initialized at zero.

5. **Compute the same-state comparator and persistent base.** Evaluate the genuine no-protection endpoint without installing it. Construct the protected projected reference, the normalized charge budget, and the maximal certified path fraction. Accept a persistent base endpoint only if the declared path-fraction, retention, descent, projection, trust-region, and numerical checks all pass.

6. **Complete the finite deployed transaction.** From the accepted persistent base, form the finite requested-output set for the current minibatch, active protected evidence, certified candidates, selected transfer target, and guards. Construct the compact-cardinal residual and accept only if all finite endpoint, capacity, separation, and numerical checks pass. Otherwise restore the entire predictive state. The unrestricted base endpoint is never a fallback for a failed protected transaction.

7. **Update reopening and route refinement.** Outcome evidence may justify reopening or release. A new observable route additionally requires the separately controlled signature evidence and effect-size conditions. Lack of signature information cannot be replaced by semantic labels.

8. **Renew only through zero-gated structure.** A dormant module may be reset only while its functional gate makes that reset exactly function preserving. Recompute the full protected geometry after reset and activate the module only through an accepted protected update with nonzero trial-slot motion; otherwise roll the reset back.

2.6 Task-free routing, consolidation, and reopening

This section collects the statistical and routing results used by the task-free controller. Positive route identification is conditional on observable separation; failure of that condition does not invalidate the pathwise retention results for the actually routed sequence.

Theorem 2.13 (Exact task-free routing under separated observable contexts)**.** *Suppose that at most $C_{\max}$ contexts are active and context c emits signatures only in the closed ball $B(\mu_c, r_Z)$. Assume*

$$\|\mu_c - \mu_{c'}\| > 4r_Z \quad (c \neq c'), \qquad h_Z = 2r_Z,$$

and no capacity release occurs. Initialize a new centroid at the first signature assigned to its empty slot and update it by the convex EMA rule in Algorithm 2.12. Then every subsequent signature is routed to the unique correct context slot, without task labels.

Proof. A centroid initialized from context c lies in $B(\mu_c, r_Z)$. Because the ball is convex, every EMA update using another signature from c keeps the centroid in the same ball. Hence a signature from c lies at distance at most $2r_Z = h_Z$ from its own centroid. Its distance from any centroid of $c' \neq c$ is greater than $4r_Z - 2r_Z = h_Z$. Thus nearest-threshold routing selects the unique correct slot. Before context c has a slot, its first signature lies beyond threshold from every existing other-context centroid and therefore opens an empty candidate slot. Induction over arrivals completes the proof. □

2.6.1 Frozen-representation calibration and observable families

A finite unprotected initialization prefix may train the backbone, but calibration features produced while that backbone is changing do not constitute one fixed signature map. The admissible sequence is therefore: retain bounded learner-visible prefix references and transform seeds; finish bootstrap; freeze the backbone; replay the declared prefix through that final frozen representation; and only then fix all signature centers, scales, radii, temporal state, and routing thresholds.

Proposition 2.14 (Honest frozen-signature calibration)**.** *Let $R_1, \dots, R_m$ be the declared learner-visible prefix references and let $Z^\star(R_i)$ be their signatures under the final frozen representation. Any deterministic calibration statistic computed from $(Z^\star(R_i))_{i \le m}$ is measurable before the protected horizon and may initialize the fixed router. If the requested calibrated threshold h_Z^{req} exceeds a predeclared operational ceiling $h_Z^{\max}$, applying the ceiling is permitted only while withholding the positive routing conclusion, with no invocation of Theorem 2.13 for that calibration. The universal retention, consolidation, reopening, and obstruction statements remain valid for the actual routed sequence.*

Proof. The replay uses only pre-horizon learner-visible information and the final frozen map, so all calibrated quantities are fixed before protected decisions. A clipped threshold generally does not satisfy the coverage or separation premise that produced h_Z^{req}; withholding the positive routing corollary is therefore necessary. The arbitrary-stream results condition on the realized routing and already charge mismatch explicitly, so they do not require a positive calibration claim. □

The specification may predeclare a finite family of bounded task-free signatures $Z^{(1)}, \dots, Z^{(M)}$, such as named frozen-layer or deterministic-augmentation statistics. No family member may be selected with evaluator contexts or test-set purity. For route r, let $E_{r,b}^{(m)}$ be a valid e-process for the stability null of expert m on distinct observable blocks, and fix $w_m > 0$ with $\sum_m w_m = 1$.

Proposition 2.15 (Error-controlled finite signature family)**.** *The mixture*

$$E_{r,b}^{\mathrm{mix}} = \sum_{m=1}^{M} w_m E_{r,b}^{(m)} \tag{66}$$

is an e-process under the intersection stability null and therefore

$$\mathbb{P}\left(\sup_b E_{r,b}^{\mathrm{mix}} \geq 1/\alpha_r^{\mathrm{sig}} \right) \leq \alpha_r^{\mathrm{sig}}.$$

If one predeclared expert $m^\star$ has positive information, its crossing analysis incurs only the fixed evidence overhead $\log(1/w_{m^\star})$. If no observable expert has positive information, AFM makes no finite route-identification claim.

Proof. A nonnegative fixed weighted sum of e-processes is an e-process. Ville's inequality gives the first statement, and $E_{r,b}^{\mathrm{mix}} \geq w_{m^\star} E_{r,b}^{(m^\star)}$ gives the fixed log-evidence overhead. Sequential classifier two-sample processes are one possible construction, but the theorem requires only valid predictable e-processes Jang et al. (2022); Ramdas et al. (2023). □

The evaluated implementation uses $M = 1$ with a predeclared hybrid signature; the theorem records the stronger finite-family option without assuming that an informative observable signature exists.

2.6.2 Operational consolidation and evidence-based reopening

The bounded-state instantiation separates candidate fitting from certification. For one routing slot, collect a finite fitting block $\mathcal{T} = (z_1, \ldots, z_{m_{\mathrm{fit}}})$, apply the fixed deterministic candidate optimizer to a private vector initialized before that block, and freeze its output $\bar{\theta}$. The outcomes in $\mathcal{T}$ are training outcomes only. The validation sequence starts strictly afterwards, so the snapshot prediction for every validation outcome is fixed before that outcome is observed. Candidate fitting may succeed or fail arbitrarily; no retention or risk claim is attached to it until the separate validation test commits.

During consolidation the privately fitted candidate snapshot $\bar{\theta}$ is frozen, while the deployed learner may continue ordinary protected AFM updates and may attempt the safe transfer rule above. Let $Y_n \in [0, 1]$ be the frozen candidate's predictable validation loss on the nth routed outcome and let $\mu_n = \mathbb{E}[Y_n \mid \mathcal{F}_{n-1}]$. Define

$$U_n = \bar{Y}_n + \sqrt{\frac{\log(\pi^2 n^2/(6\alpha))}{2n}}. \tag{67}$$

Theorem 2.16 (Anytime-valid operational consolidation)**.** *With probability at least $1 - \alpha$, simultaneously for every $n \geq 1$,*

$$\frac{1}{n} \sum_{i=1}^{n} \mu_i \leq U_n. \tag{68}$$

Therefore a record committed only when $U_n \leq \tau$ has certified average conditional validation risk at most τ on its evidence sequence. No stationarity or iid assumption is needed.

Proof. For fixed n, conditional Hoeffding-Azuma gives

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (\mu_i - Y_i) > r_n \right) \leq e^{-2nr_n^2} = \frac{6\alpha}{\pi^2 n^2}.$$

Sum over n. □

A bounded candidate test also fixes a maximum validation count N_{val}. Let $S_n = \sum_{i=1}^{n} Y_i$. Because every future loss is nonnegative, the smallest terminal upper bound attainable after the current history, even if all remaining losses are zero, is

$$U^{\mathrm{best}}_{n \to N_{\mathrm{val}}} = \frac{S_n}{N_{\mathrm{val}}} + \sqrt{\frac{\log(\pi^2 N_{\mathrm{val}}^2/(6\alpha))}{2N_{\mathrm{val}}}}. \tag{69}$$

If this quantity exceeds τ, rejection is algebraically forced and may occur immediately; otherwise the candidate is rejected without commitment at N_{val} if its UCB has never crossed the threshold. Rejection makes no statistical claim. Starting another candidate with a fresh summably allocated α^{cert} preserves Theorem 2.16 and lifetime error control.

Define the first certification crossing

$$\tau_j^{\mathrm{cert}} = \inf\{n \geq n_{\min} : U_{j,n} \leq \tau_{\mathrm{risk}}\}. \tag{70}$$

When this stopping time is finite, AFM freezes the count, loss sum, UCB, bounded commit evidence, candidate outputs, streamed sketch, and transfer objective. It stops appending outcomes to that certification sequence. The transfer target is therefore immutable while the deployed learner approaches it.

For later routed outcomes, let $Y^{\mathrm{stale}}_{j,m} \in [0,1]$ be the frozen candidate's predictable error and define

$$X^{\mathrm{stale}}_{j,m} = Y^{\mathrm{stale}}_{j,m} - \tau_{\mathrm{risk}}.$$

A fresh summably allocated half-normal-mixture e-process, with certified sub-Gaussian scale at least $1/2$, tests the post-certification null

$$\mathbb{E}[X^{\mathrm{stale}}_{j,m} \mid \mathcal{F}_{m-1}] \leq 0.$$

Proposition 2.17 (Frozen first-crossing certificate and separately controlled staleness)**.** *With probability at least $1 - \alpha_j^{\mathrm{cert}}$, the certificate in Theorem 2.16 holds at the stopping time τ_j^{cert}. Conditional on the separately allocated staleness null, the probability that the post-certification e-process ever falsely cancels the candidate is at most $\alpha_j^{\mathrm{stale}}$. The two conclusions hold simultaneously over the learner's lifetime under the declared summable allocation.*

Proof. Theorem 2.16 is simultaneous in n and therefore valid at the first crossing. The staleness process uses only subsequent predictable outcomes and a fresh allocation; Ville's inequality controls its lifetime crossing. A union bound over all instantiated processes gives the lifetime statement Waudby-Smith & Ramdas (2024); Ramdas et al. (2023). □

A certified candidate whose activation gap exceeds τ^{act} enters the predictable bounded transfer queue. It is committed only after the exact gap closes, canceled if the staleness process crosses, or rejected after its fixed number of service opportunities. These outcomes respectively record successful safe transfer, observable post-certification invalidation, or a quantified transfer-service obstruction. No alternative deployed predictor is protected in their place.

2.6.3 Separately controlled reopening and observable route refinement

Outcome contradiction and observable context displacement are different statistical questions. For record j, let $E^{\mathrm{out}}_{j,n}$ denote the outcome e-process driven by (61), where n counts routed outcomes. Independently, let b count *distinct observable signature blocks*. A signature repeated for several outcome-resolved members contributes once to b and once to the route-refinement statistic.

At commitment, freeze the source centroid $m_{c(j)}$ and a predictable reference radius $r_j \in [0,2]$ computed from pre-crossing information by the declared specification. Assume signatures are normalized so $\|Z_{j,b}\| \leq 1$ and $\left\|m_{c(j)}\right\| \leq 1$. Define

$$D_{j,b} = \left\|Z_{j,b} - m_{c(j)}\right\|, \qquad W_{j,b} = \frac{D_{j,b} - r_j}{2}. \tag{71}$$

For fixed r_j, $W_{j,b}$ lies in an interval of length one, so its centered noise is conditionally 1/2-sub-Gaussian. The operational observable-stability null is

$$\mathbb{E}[W_{j,b} \mid \mathcal{G}_{j,b-1}] \le 0, \tag{72}$$

where $\mathcal{G}_{j,b-1}$ contains the history before the bth distinct signature block. This is an operational null, not an assumption that semantic contexts are identifiable. If r_j is estimated from a stochastic recurrence model, any claim that (72) holds must be justified by that model and, when statistical, by a separately allocated calibration event; absent such a model, the universal theorem reports violation of the null as observable-shift evidence without treating recurrence as an assumed property.

Using a fixed half-normal mixing distribution Π_Z and certified scale $\sigma_Z = 1/2$ (or any larger certified scale), define

$$E_{j,b}^{\mathrm{sig}} = \int_0^\infty \exp\left(\lambda \sum_{i=1}^{b} W_{j,i} - \frac{\lambda^2 \sigma_Z^2 b}{2}\right) \Pi_Z(d\lambda). \tag{73}$$

It has the same fixed-state half-normal evaluation as (77), with the outcome sufficient statistics replaced by the signature-block sufficient statistics.

Let $\bar{Z}_{j,b} = b^{-1} \sum_{i=1}^{b} Z_{j,i}$. With separately allocated levels α_j^{open} and $\alpha_j^{\mathrm{split}}$, route refinement is permitted only if

$$E_{j,n}^{\mathrm{out}} \ge \frac{1}{\alpha_j^{\mathrm{open}}}, \qquad E_{j,b}^{\mathrm{sig}} \ge \frac{1}{\alpha_j^{\mathrm{split}}}, \qquad b \ge b_{\mathrm{split}}, \qquad \left\|\bar{Z}_{j,b} - m_{c(j)}\right\| \ge h_{\mathrm{split}}. \tag{74}$$

Then the bounded route-refinement rule may preserve all active source segments and allocate one distinct centroid at $\bar{Z}_{j,b}$, subject to the declared context-capacity policy. The split changes only future routing and alters no protected anchor, evidence block, sketch, active budget, or current parameter. If only the outcome process crosses, the record remains protected during a fixed finite diagnostic grace window and is then reopened or released without allocating a route unless (74) becomes true.

Proposition 2.18 (Independent lifetime control and retention invariance)**.** *For a record satisfying the outcome null* (62)*,*

$$\mathbb{P}\left(\sup_n E_{j,n}^{\mathrm{out}} \ge 1/\alpha_j^{\mathrm{open}}\right) \le \alpha_j^{\mathrm{open}}.$$

For a record satisfying the observable-stability null (72)*, even when the outcome null is false,*

$$\mathbb{P}(\textit{record } j \textit{ ever creates a route split}) \le \mathbb{P}\left(\sup_b E_{j,b}^{\mathrm{sig}} \ge 1/\alpha_j^{\mathrm{split}}\right) \le \alpha_j^{\mathrm{split}}.$$

If construction of the frozen reference set or radius is covered by a separate event of failure probability η_j, the unconditional false-split bound is $\eta_j + \alpha_j^{\mathrm{split}}$. Conditional on any permitted split, every source record's existing retention guarantee is unchanged until explicit release.

Proof. Each outcome and signature exponential component is a nonnegative supermartingale under its own null, and fixed mixing preserves that property. Ville's inequality gives the two bounds separately. A split is a subset of the signature crossing event regardless of whether semantic or predictive contradiction makes the outcome process cross. The split performs no parameter update and modifies none of the source records' protected state, so their retention proofs are unchanged. The optional calibration term follows by a union bound. □

Theorem 2.19 (Joint delay under outcome and observable information)**.** *Suppose after a change the outcome score has conditional mean at least $\Delta_{\mathrm{out}} > 0$, the signature score has conditional mean at least $\Delta_{\mathrm{sig}} > 0$, and after some finite block index the vector effect condition in* (74) *holds. For every $0 < \beta < 1$, a permitted split occurs, subject to available bounded capacity, after at most*

$$C \max\left\{ \frac{\sigma_{\mathrm{out}}^2}{\Delta_{\mathrm{out}}^2} \left[\log \frac{1}{\alpha_j^{\mathrm{open}}} + \log \frac{2}{\beta} + \omega_{\mathrm{out}} \right], \frac{\sigma_Z^2}{\Delta_{\mathrm{sig}}^2} \left[\log \frac{1}{\alpha_j^{\mathrm{split}}} + \log \frac{2}{\beta} + \omega_{\mathrm{sig}} \right] \right\} \tag{75}$$

corresponding outcome and distinct-block observations, up to the declared minimum-block and effect-size gates. Here ω_{out} and ω_{sig} are the fixed mixture-mass overheads from Theorem 2.21. If $\Delta_{\mathrm{sig}} = 0$, no finite route-splitting delay is claimed, even when semantic conflict makes $\Delta_{\mathrm{out}} > 0$; ordinary reopening or release remains available.

Proof. Apply Theorem 2.21 separately to the outcome and signature scores with failure probability $\beta/2$, then take a union bound and wait for the slower crossing together with the deterministic gates. Zero signature drift removes the information needed to justify a new observable route. □

This refinement does not assert that an external semantic context has been identified. Semantic-only change can justify reopening from outcomes but cannot justify a new route when the declared signature law is stable. Observational indistinguishability therefore remains an explicit routing-mismatch obstruction rather than being converted into a noisy semantic oracle.

For reopening fix a probability measure Π on $(0, \infty)$ that assigns positive mass to every nonempty interval; a half-normal density is one concrete choice. Use any certified sub-Gaussian scale σ for (61) (the universal bounded-loss choice is $\sigma = 1$) and define the continuous-mixture e-process

$$E_n = \int_0^\infty \exp\left(\lambda \sum_{i=1}^n X_i - \frac{\lambda^2 \sigma^2 n}{2}\right) \Pi(d\lambda). \tag{76}$$

It requires only the sufficient state $S_n = \sum_{i=1}^n X_i$ and n. For example, with the half-normal prior of scale $\tau > 0$, set $A_n = \sigma^2 n + \tau^{-2}$; then

$$E_n = \frac{2}{\tau\sqrt{A_n}} \exp\left(\frac{S_n^2}{2A_n}\right) \Phi_{\mathrm{N}}\left(\frac{S_n}{\sqrt{A_n}}\right), \tag{77}$$

where Φ_{N} is the standard-normal distribution function. Thus adaptation to unknown positive drift has exact fixed-state evaluation and does not require an ever-growing grid.

Theorem 2.20 (Lifetime false-reopening control)**.** *Under* (62)*, E_n is a nonnegative supermartingale with $E_0 = 1$. Therefore*

$$\mathbb{P}\left(\sup_{n \geq 1} E_n \geq 1/\alpha\right) \leq \alpha. \tag{78}$$

Summable allocation of α over records controls every false reopening over an unbounded lifetime.

Proof. Each exponential component is a supermartingale by conditional sub-Gaussianity. Tonelli's theorem shows that integration against the fixed probability measure Π preserves this property. Apply Ville's inequality; see Ramdas et al. (2023). □

Theorem 2.21 (Delay under actual observable contradiction)**.** *Suppose after a change $\mu_i \geq \Delta > 0$. Let*

$$I_\Delta = \left[\frac{\Delta}{2\sigma^2}, \frac{3\Delta}{4\sigma^2}\right], \qquad \pi_\Delta = \Pi(I_\Delta) > 0.$$

For every $0 < \beta < 1$, threshold crossing occurs by

$$N \leq C \frac{\sigma^2}{\Delta^2}\left[\log\frac{1}{\alpha} + \log\frac{1}{\beta} + \log\frac{1}{\pi_\Delta} + 1\right] \tag{79}$$

with probability at least $1 - \beta$, for a universal numerical constant C. If no score has positive observable drift, no finite delay is claimed.

Proof. Sub-Gaussian concentration gives, with probability $1 - \beta$, $S_N := \sum_{i=1}^N X_i \geq N\Delta - \sigma\sqrt{2N\log(1/\beta)}$. Uniformly for $\lambda \in I_\Delta$,

$$\lambda N\Delta - \frac{\lambda^2\sigma^2 N}{2} \geq \frac{3N\Delta^2}{8\sigma^2},$$

while the concentration penalty $\lambda\sigma\sqrt{2N\log(1/\beta)}$ is at most $3\Delta\sqrt{2N\log(1/\beta)}/(4\sigma)$. Hence

$$E_N \geq \pi_\Delta \exp\left(\frac{3N\Delta^2}{8\sigma^2} - \frac{3\Delta}{4\sigma}\sqrt{2N\log(1/\beta)}\right).$$

A sufficiently large universal C in (79) makes this at least $1/\alpha$. □

2.7 Whole-behavior memory and spectral protection

The results below quantify the leakage represented by the bounded sensitivity sketches and characterize the exact first-order rank-plasticity frontier.

2.7.1 Whole-behavior memory and the local frontier

Theorem 2.22 (Deterministic streamed-memory certificate)**.** *For every round and every unit vector $v \in \mathrm{range}(\Pi_t)$,*

$$\|J_t v\| \leq \varepsilon_t + e_t^{\mathrm{pop}}. \tag{80}$$

For the routed empirical behavior, $e_t^{\mathrm{pop}} = 0$ and the result is deterministic.

Proof. At the anchors, stack the realized weighted conceptual matrices $\sqrt{w_{j,t}}C_j$ and sketches $\sqrt{w_{j,t}}B_j$. By (47),

$$\sum_j w_{j,t}\|C_j v\|^2 \leq v^T S_t^* v + \left(\sum_j w_{j,t}\Delta_j^{\mathrm{FD}}\right)\|v\|^2.$$

Since $v \in \mathrm{range}(\Pi_t)$ and Q_t is an orthogonal projector,

$$v^T S_t^* v \leq \lambda_{\max}((I-Q_t)S_t^*(I-Q_t)) \leq \mathrm{tr}((I-Q_t)S_t^*) = a_t^{\mathrm{spec}}.$$

The difference between the current and anchor derivatives is at most $d_t^{\mathrm{anc}}\|v\|$ by the definition of the Lipschitz constants and Minkowski's inequality. Since $w_{j,t} \geq 1$, the unweighted stacked norm is no larger than the weighted norm. Add the valid population-empirical discrepancy certificate e_t^{pop}. □

Corollary 2.23 (Constructive population certificate)**.** *Suppose the commit inputs of record j are iid from a distribution P_j and $\|A_j(x)\|_{\mathrm{op}} \leq R_j$ almost surely. Let*

$$G_j = \mathbb{E}_{P_j}[A_j(x)^T A_j(x)], \qquad \widehat{G}_{j,n} = C_{j,n}^T C_{j,n}$$

for the scaled conceptual row stack formed from the first n commit inputs. If n_j is fixed before those inputs are observed, then with probability at least $1-\delta_j$,

$$\left\|\widehat{G}_{j,n_j} - G_j\right\|_{\mathrm{op}} \leq \xi_j := R_j^2\left[\sqrt{\frac{2\log(2d/\delta_j)}{n_j}} + \frac{2\log(2d/\delta_j)}{3n_j}\right]. \tag{81}$$

If instead the commitment count is selected adaptively from the same iid sequence, predeclare numbers $\delta_{j,n} > 0$ with $\sum_{n\geq 1}\delta_{j,n} \leq \delta_j$ and define

$$\xi_{j,n} := R_j^2\left[\sqrt{\frac{2\log(2d/\delta_{j,n})}{n}} + \frac{2\log(2d/\delta_{j,n})}{3n}\right]. \tag{82}$$

Then with probability at least $1-\delta_j$, simultaneously for every $n \geq 1$, $\left\|\widehat{G}_{j,n} - G_j\right\|_{\mathrm{op}} \leq \xi_{j,n}$; in particular the bound is valid at any data-dependent commit time. In the formulas below, ξ_j denotes the applicable fixed-n_j value or the simultaneous value ξ_{j,n_j} at commitment. Assume additionally that the pointwise Jacobian is

$L_{J,j}$-Lipschitz in θ. On the intersection of these events, a valid current-parameter population certificate in (80) *is*

$$e_t^{\mathrm{pop}} \le \left(\sum_{j \in \mathcal{A}_t} w_{j,t} \xi_j \right)^{1/2} + \left(\sum_{j \in \mathcal{A}_t} w_{j,t} L_{J,j}^2 \left\| \theta_t - \bar{\theta}_j \right\|^2 \right)^{1/2}.$$

The second term transports the population derivative from the anchor to the current parameter; the empirical transport is already present in d_t^{anc}. Thus the population term is constructible under declared sampling conditions rather than assumed to vanish, including at adaptive commitment times when the simultaneous allocation is used.

Proof. For each fixed n, apply self-adjoint matrix Bernstein to $A_j(x_i)^T A_j(x_i) - G_j$ and divide by n Tropp (2012). This gives (81) at a fixed n_j. For adaptive commitment, apply the same fixed-n bound with level $\delta_{j,n}$ and take a union bound over n; the resulting event is simultaneous and hence remains valid at a stopping time. For any unit v, the population quadratic form is at most the empirical form plus the applicable ξ_j. Stack the weighted records and take square roots. □

Theorem 2.24 (Exact rank-r local frontier)**.** *Let $J : \mathbb{R}^d \to \mathcal{H}$ be compact with singular values $\sigma_1 \ge \cdots \ge \sigma_d \ge 0$, and let $0 \le r < d$. Among all plastic subspaces $S \subseteq \mathbb{R}^d$ of dimension $d - r$,*

$$\inf_{\dim S = d-r} \sup_{v \in S,\ \|v\|=1} \|Jv\| = \sigma_{r+1}(J). \tag{83}$$

The optimum is the span of the right singular vectors associated with $\sigma_{r+1}, \ldots, \sigma_d$.

Proof. This is the Courant-Fischer min-max characterization applied to J^*J. □

Corollary 2.25 (Unavoidable first-order forgetting price)**.** *No rank-r linear protection rule can guarantee first-order leakage smaller than $\sigma_{r+1}(J)$ in every unit plastic direction. Thus the spectral tail in the AFM certificate is not an artifact of the proof.*

2.8 Retention and persistent adaptation

This section derives the one-step and active-interval retention results, the projected-stationarity identity, and sufficient conditions for continued compatible persistent movement.

2.8.1 Retention and compatible adaptation

Theorem 2.26 (One-step joint persistent retention and exact deployed progress)**.** *For every accepted protected update, let $\bar{\theta}_{t+1} = \theta_t + d_t^{\mathrm{safe}}$ be the persistent safe-base endpoint and let $s_t^{\mathrm{safe}} = \left\| d_t^{\mathrm{safe}} \right\|$. Then*

$$\left\| \Phi_t^{[S_t]}(\bar{\theta}_{t+1}) - \Phi_t^{[S_t]}(\theta_t) \right\| \le (\varepsilon_t + e_t^{\mathrm{pop}}) s_t^{\mathrm{safe}} + \frac{\overline{H}_t}{2} (s_t^{\mathrm{safe}})^2 \le b_t, \tag{84}$$

$$\ell_t^{S_t}(\bar{\theta}_{t+1}) \le \ell_t^{S_t}(\theta_t) - \frac{1}{2} s_t^{\mathrm{safe}} \left\| \Pi_t \nabla \ell_t^{S_t}(\theta_t) \right\|, \tag{85}$$

$$\mathcal{L}_t(F_{t+1}) = \mathcal{L}_t(U_t) = \mathcal{L}_t(F_t) - \Delta_t^0, \tag{86}$$

$$\frac{\ell_t^{S_t}(\theta_t) - \ell_t^{S_t}(\bar{\theta}_{t+1})}{\Delta_t^0} \ge \frac{\widehat{\lambda}_t \kappa_t}{3} \quad \textit{in counterfactual-normalized mode.} \tag{87}$$

For any deployed behavior constructor,

$$\left\| \widetilde{\Phi}_t(\bar{\theta}_{t+1}, S_{t+1}) - \widetilde{\Phi}_t(\theta_t, S_t) \right\| \le b_t + \omega_t^S. \tag{88}$$

For an external evaluator map $\widetilde{\Psi}_t = \widetilde{\Phi}_t + \Delta_t$ defined on the joint state,

$$\left\| \widetilde{\Psi}_t(\bar{\theta}_{t+1}, S_{t+1}) - \widetilde{\Psi}_t(\theta_t, S_t) \right\| \le b_t + \omega_t^S + \omega_t^\Delta + r_t s_t^{\mathrm{safe}} + \frac{k_t}{2} (s_t^{\mathrm{safe}})^2. \tag{89}$$

On every declared finite protected-evidence block, the deployed drift in (88) *is exactly zero by Theorem 2.2. No routing correctness is assumed.*

Proof. The safe-base displacement lies in range(Π_t) and is clipped by the root of the certified quadratic retention upper model, proving (84). The loss upper model and the accepted safe-base endpoint check prove (85); Theorem 2.1 gives (87) in the normalized mode. The shield construction reproduces the genuine no-protection logits on the complete current minibatch, proving (86). Equation (88) is the joint-state bound (40); applying the same fixed-S_t Taylor bound to $\widetilde{\Delta}_t$ and then charging its exact shield-replacement change by (18) gives (89). Exact finite deployed retention is item (iii) of Theorem 2.2. □

Corollary 2.27 (Active-interval joint retention and exact finite deployed retention)**.** *If record j remains active on rounds $\tau, \dots, T-1$, then every broader deployed behavior map obeys*

$$\left\|\widetilde{\Phi}_j(\theta_T, S_T) - \widetilde{\Phi}_j(\theta_\tau, S_\tau)\right\| \le \sum_{t=\tau}^{T-1} \left(b_t + \omega_{j,t}^S + \omega_{j,t}^\Delta + r_{j,t} s_t^{\mathrm{safe}} + \frac{k_{j,t}}{2} (s_t^{\mathrm{safe}})^2 \right). \tag{90}$$

For the declared finite protected-evidence map, every accepted joint step restores the complete pre-step deployed value, so the stronger identity

$$\widetilde{\Phi}_j(\theta_T, S_T) = \widetilde{\Phi}_j(\theta_\tau, S_\tau) \tag{91}$$

holds up to the single declared outward numerical envelope. If the record was activated at c_j from validated snapshot $(\bar{\theta}_j, S_j^{\mathrm{snap}})$, then

$$\left\|\widetilde{\Phi}_j(\theta_T, S_T) - \widetilde{\Phi}_j(\bar{\theta}_j, S_j^{\mathrm{snap}})\right\| \le a_j^{\mathrm{act}}, \tag{92}$$

with no cumulative active-interval finite-evidence drift term. Thus the budget controls the persistent safe-base leg, the explicit ω^S ledger charges the structural replacement for broader maps, and the finite protected map is restored exactly.

Proof. Telescope the one-step joint bound (89) to obtain (90). Equation (91) follows by induction from exact pre-step restoration on the complete finite evidence block. Bridge once from the frozen validated snapshot to the activation state to obtain (92). □

Proposition 2.28 (Explicit activation and release decomposition)**.** *Let record j be activated at c_j from snapshot $\bar{\theta}_j$ and released at $r_j \le T$, with the convention $r_j = T$ if it remains active. Then*

$$\left\|\widetilde{\Phi}_j(\theta_T, S_T) - \widetilde{\Phi}_j(\bar{\theta}_j, S_j^{\mathrm{snap}})\right\| \le a_j^{\mathrm{act}} + \left\|\widetilde{\Phi}_j(\theta_T, S_T) - \widetilde{\Phi}_j(\theta_{r_j}, S_{r_j})\right\|. \tag{93}$$

The first term is the observed transfer from the validated snapshot to the deployed activation state, the active finite-evidence interval is exact, and the final term is exact post-release drift intentionally uncontrolled by a bounded registry. Thus delayed activation and capacity release are reported separately from protected forgetting.

Proof. Insert and subtract $\widetilde{\Phi}_j(\theta_{r_j}, S_{r_j})$, apply the triangle inequality, and use exact finite deployed retention from Corollary 2.27 on the active interval. □

Theorem 2.29 (Arbitrary-stream safe-base projected-stationarity identity)**.** *Let*

$$\nu_t^{\mathrm{joint}} := \max\{\ell_{t+1}^{S_{t+1}}(\theta_{t+1}) - \ell_t^{S_t}(\theta_{t+1}), 0\}$$

be the realized variation of the pre-step-shield loss sequence. For every horizon T,

$$\sum_{t=1}^{T-1} s_t^{\mathrm{safe}} \left\|\Pi_t \nabla \ell_t^{S_t}(\theta_t)\right\| \le 2\big(\ell_1^{S_1}(\theta_1) - \ell_{\inf} + V_T^{\mathrm{joint}}\big), \qquad V_T^{\mathrm{joint}} = \sum_{t=1}^{T-1} \nu_t^{\mathrm{joint}}. \tag{94}$$

This holds pathwise for smooth non-convex losses. Independently, every accepted deployed update has exact current-batch ratio one by (38)*.*

Proof. Apply (85), add the realized joint-loss variation, telescope, and use the lower bound on the final loss. The shield-emulated deployed progress is a separate exact ledger and is not substituted into the base stationarity proof. □

Corollary 2.30 (Realized lifelong compatible base plasticity plus exact current emulation)**.** *On any infinite tail for which*

$$\sum_t \nu_t^{\mathrm{joint}} < \infty, \qquad \sum_t s_t^{\mathrm{safe}} = \infty,$$

one has

$$\liminf_{t\to\infty} \left\| \Pi_t \nabla \ell_t^{S_t}(\theta_t) \right\| = 0.$$

If additionally $\sum_t b_t < \infty$, the cumulative certified fixed-shield behavior charge of the safe-base legs is finite, while every declared finite deployed evidence block has exact active-interval retention. A broader joint behavior has finite total drift whenever its structural and routing charges are also summable. Thus indefinitely available compatible persistent movement, exact finite protection, and ratio-one current endpoint emulation coexist on the realized accepted tail.

Proof. If the liminf were bounded below by $\gamma > 0$, the left side of (94) would dominate $\gamma \sum_t s_t^{\mathrm{safe}} = \infty$. The safe-base behavior-charge and exact finite deployed-retention statements follow from Theorem 2.26 and Corollary 2.27. □

Proposition 2.31 (A checkable sufficient safe-base movement condition)**.** *Suppose an infinite non-barrier tail satisfies $\sum_t \nu_t^{\mathrm{joint}} < \infty$, $\sum_t b_t < \infty$, and $\sum_t \sqrt{b_t} = \infty$. If there is a constant $c_s > 0$ such that every accepted nonzero safe-base update on that tail has*

$$s_t^{\mathrm{safe}} \geq c_s \sqrt{b_t},$$

and, writing $\mathcal{I} = \{t : \text{the safe-base update is accepted and nonzero}\}$, the accepted rounds retain divergent square-root budget mass,

$$\sum_{t\in\mathcal{I}} \sqrt{b_t} = \infty,$$

then Corollary 2.30 applies. Equivalently, it suffices that the omitted rounds have finite square-root budget mass $\sum_{t\notin\mathcal{I}} \sqrt{b_t} < \infty$. The lower bound is an observable certificate condition on the realized safe radius, projected gradient, and endpoint backtracking factor; it is not inferred merely from first-order compatibility or from the shield update.

Proof. The assumptions give $\sum_t s_t^{\mathrm{safe}} \geq \sum_{t\in\mathcal{I}} s_t^{\mathrm{safe}} \geq c_s \sum_{t\in\mathcal{I}} \sqrt{b_t} = \infty$. Apply Corollary 2.30. □

2.9 Multi-timescale metaplasticity and allocation

The following results justify the logarithmic trace bank and the adaptive rank/timescale controller. The controller is evaluated on a frontier cost that contains both residual protected leakage and gradient energy blocked by protection.

2.9.1 Multi-timescale metaplasticity and allocation

For $\rho_k = 2^{-(k+1)}$, the impulse response of trace k at age $\tau \geq 1$ is

$$h_k(\tau) = \rho_k(1-\rho_k)^{\tau-1}.$$

For $0 < \alpha \leq 1$, define the unnormalized scale-free policy kernel

$$M_{\alpha,K}(\tau) = \sum_{k=0}^{K-1} \rho_k^{\alpha-1} h_k(\tau) = \sum_{k=0}^{K-1} \rho_k^{\alpha}(1-\rho_k)^{\tau-1}. \tag{95}$$

A policy rescales this known finite kernel into its declared bounded weight range.

Theorem 2.32 (Logarithmic EMA bank approximates scale-free relevance)**.** *For every $0 < \alpha \le 1$ and $K \ge 2$, there are constants $0 < c_\alpha \le C_\alpha < \infty$, independent of K and τ, such that*

$$c_\alpha \tau^{-\alpha} \le M_{\alpha,K}(\tau) \le C_\alpha \tau^{-\alpha}, \qquad 1 \le \tau \le 2^{K-2}. \tag{96}$$

Consequently, $K = O(\log H)$ actual EMA traces approximate a power-law relevance profile over ages $1, \dots, H$ with constant multiplicative distortion.

Proof. Choose k_* so that $\rho_{k_*}\tau \in [1/4, 1/2]$, which exists in the stated range. Since $\rho_{k_*} \le 1/2$,

$$(1-\rho_{k_*})^{\tau-1} \ge \exp(-2\rho_{k_*}(\tau-1)) \ge e^{-1}.$$

The k_* term is therefore at least $e^{-1}4^{-\alpha}\tau^{-\alpha}$.

For $\tau = 1$, the upper bound is the finite geometric sum $\sum_k \rho_k^\alpha$. For $\tau \ge 2$, split the sum into $\rho_k\tau < 1$ and $\rho_k\tau \ge 1$. The first part is a geometric sum bounded by a constant times $\tau^{-\alpha}$. For the second, $\tau - 1 \ge \tau/2$ and $(1-\rho_k)^{\tau-1} \le e^{-\rho_k(\tau-1)}$; grouping dyadic values of $\rho_k\tau$ shows that the sum is at most $\tau^{-\alpha}\sum_{m\ge 0} 2^{\alpha(m+1)}e^{-2^{m-1}}$, which is finite. □

Theorem 2.33 (Every single exponential has polynomial worst-case distortion)**.** *Let $H \ge 3$ be odd and $f(\tau) = \tau^{-\alpha}$. If $g(\tau) = aq^{\tau-1}$, $a > 0$, $0 < q \le 1$, satisfies*

$$C^{-1} f(\tau) \le g(\tau) \le C f(\tau), \qquad 1 \le \tau \le H,$$

then

$$C \ge 2^{-\alpha/2} H^{\alpha/4}. \tag{97}$$

Proof. Let $m = (H+1)/2$. Exponential sequences satisfy $g(m)^2 = g(1)g(H)$. The approximation inequalities imply

$$C^2 f(m)^2 \ge g(m)^2 \ge C^{-2} f(1) f(H),$$

so $C^4 \ge f(1)f(H)/f(m)^2 = m^{2\alpha}H^{-\alpha} \ge H^\alpha/4^\alpha$. □

For each allocation policy $p \in \mathcal{P}$, let $Q_{p,t}$ be the projector it would select before the current importance is revealed. After the current signal, define its realized frontier cost and bounded expert loss

$$F_t(p) = \operatorname{tr}((I - Q_{p,t})S_t^*) + \zeta_t c_{p,t}^{\text{block}}, \tag{98}$$

$$\mathcal{L}_t(p) = \Lambda^{-1} F_t(p) \in [0,1]. \tag{99}$$

The design uses declared bounds $\|B_j\|_F^2 \le M_j$ and

$$\Lambda = 2\sum_{j=1}^{J_{\max}} M_j + \zeta_{\max},$$

so the loss is bounded for every registry state. The first term penalises retention leakage; the second penalises plastic gradient energy blocked by the protected subspace. Thus larger rank is not free and cannot dominate merely by freezing more coordinates.

Theorem 2.34 (Prediction error to functional-leakage transfer)**.** *Fix a rank r. Let $Q_{p,t}$ be a top-r projector of $\widehat{S}_{p,t}$ and let Q_t^* be a top-r projector of S_t^*. Define*

$$a_{p,t} = \operatorname{tr}((I - Q_{p,t})S_t^*), \qquad a_t^* = \operatorname{tr}((I - Q_t^*)S_t^*) = \sum_{i>r} \lambda_i(S_t^*).$$

Then

$$0 \le a_{p,t} - a_t^* \le 2r \left\| \widehat{S}_{p,t} - S_t^* \right\|_{\text{op}} \le 2r \sum_{j \in \mathcal{A}_t} |\widehat{u}_{j,t}^{(p)} - w_{j,t}| \, \|B_j\|_{\text{op}}^2 . \tag{100}$$

Thus a metaplastic policy that predicts the realized retention weights accurately attains correspondingly near-oracle functional leakage. If its prediction is exact, its rank-r allocation is oracle optimal for that round.

Proof. By Ky Fan's maximum principle, Q_t^* maximises $\operatorname{tr}(QS_t^*)$ and $Q_{p,t}$ maximises $\operatorname{tr}(Q\widehat{S}_{p,t})$ over rank-r orthogonal projectors. Hence

$$\begin{aligned} a_{p,t} - a_t^* &= \operatorname{tr}(Q_t^* S_t^*) - \operatorname{tr}(Q_{p,t} S_t^*) \\ &\le |\operatorname{tr}(Q_t^*(S_t^* - \widehat{S}_{p,t}))| + |\operatorname{tr}(Q_{p,t}(S_t^* - \widehat{S}_{p,t}))| \\ &\le 2r \left\| S_t^* - \widehat{S}_{p,t} \right\|_{\mathrm{op}}. \end{aligned}$$

The final inequality follows from the covariance definitions and the triangle inequality. □

Theorem 2.35 (High-probability metaplastic allocation regret)**.** *For $P \ge 2$, run exponential weights over the declared policies with learning rate $\eta_H = \sqrt{8\log(P)/T}$ and sample p_t from the current weights before observing $\mathcal{L}_t$. For $P = 1$ the conclusion below holds trivially with both overhead terms involving $\log P$ equal to zero. For every $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \mathcal{L}_t(p_t) \le \min_{p \in \mathcal{P}} \sum_{t=1}^{T} \mathcal{L}_t(p) + \sqrt{\frac{T \log P}{2}} + \sqrt{\frac{T \log(1/\delta)}{2}}. \tag{101}$$

Equivalently, multiplying by Λ,

$$\sum_{t=1}^{T} F_t(p_t) \le \min_{p \in \mathcal{P}} \sum_{t=1}^{T} F_t(p) + \Lambda\sqrt{\frac{T \log P}{2}} + \Lambda\sqrt{\frac{T \log(1/\delta)}{2}}.$$

The selected frontier costs contain exactly the spectral residual entering the memory certificate and exactly the current gradient fraction blocked by protection. Hence the timescale/rank controller is coupled simultaneously to retention and compatible adaptation rather than analyzed separately.

Proof. The exponential-weights potential argument bounds the cumulative conditional mean loss by the best expert plus $\log P/\eta_H + \eta_H T/8$. The difference between sampled and conditional-mean loss is a bounded martingale difference; Azuma-Hoeffding adds the final term. Simplifying the chosen learning rate gives (101). □

Proposition 2.36 (Metaplastic atomization invariance)**.** *The trace, policy, spectral-loss, allocation-transfer, and Hedge results remain valid if each record contribution $B_j^T B_j$ is decomposed into any finite PSD atomization*

$$B_j^T B_j = \sum_{a \in \mathcal{I}_j} M_a, \qquad M_a \succeq 0.$$

Give atom a the explicit realized signal

$$\chi_{a,t} = \min\left\{1, \frac{g_t^T M_a g_t}{1 + \|g_t\|^2}\right\},$$

its own K traces, and a policy weight, and form the predicted and realized covariances from the corresponding nonnegative weighted sums $\sum_a \widehat{u}_{a,t} A_t M_a A_t$ and $\sum_a w_{a,t} A_t M_a A_t$. The memory-certificate result also remains valid under atom-specific weights provided each atom carries a declared certified PSD sensitivity/error decomposition whose weighted sum upper-bounds the same conceptual whole-behavior derivative energy, Frequent-Directions error, and anchor-transport terms used in Theorem 2.22. A purely algebraic split of $B_j^T B_j$ does not by itself create a stronger independently weighted derivative certificate; when no atom-specific certificate is supplied, the original record-level memory certificate is retained while atomization is used only by the allocation controller. In particular, certified atoms may be individual sketch eigen-directions, parameter groups, or rank-one individual synaptic sensitivities. Persistent state grows with the fixed number of retained atoms, not with elapsed lifetime.

Proof. For the controller results, every relevant argument uses only nonnegative weighted sums of PSD contributions, their trace residual under $I - Q_t$, operator-norm perturbations of those sums, and bounded expert losses. Replacing one PSD term by a finite certified sum therefore reproduces the trace, spectral-cost, allocation-transfer, and Hedge arguments verbatim after relabeling records by atoms. For the derivative-memory statement, Theorem 2.22 additionally uses a certified upper bound on conceptual derivative energy plus sketch and anchor-transport errors. If these quantities are decomposed atomwise with a weighted sum that upper-bounds the same whole-behavior quantities, the identical Minkowski and trace argument applies; otherwise the record-level certificate is left unchanged. This is precisely the stated condition. □

2.10 Function-preserving structural renewal

Structural renewal is restricted to modules whose zero functional gate makes internal reset exactly function preserving. Any subsequent activation must pass the same protected transaction as ordinary learning.

2.10.1 Function-preserving structural renewal

A renewal batch consists of reserved modules $a_i\varphi_{w_i}$ whose gates initially satisfy $a_i = 0$. The architecture applies the same gate to the module's contribution to the realized predictor and to every active protected behavior map. Equivalently, for every active j, Φ_j is independent of w_i whenever $a_i = 0$. This is the functional, rather than merely output-level, definition of a dormant renewable module.

Theorem 2.37 (Exact reset and certified activation)**.** *Resetting any internal module parameter while its functional gate is zero changes the realized predictor, current loss value, and every active protected behavior by exactly zero. After reset, recompute A_t, Π_t, ε_t, e_t^{pop}, and the full current gradient g_t^{reset}. Let*

$$q_t^{\mathrm{reset}} = \Pi_t g_t^{\mathrm{reset}}, \qquad r_t = M_t q_t^{\mathrm{reset}},$$

where M_t projects onto an eligible named trial slot satisfying the zero-gate invariant above. The ordinary AFM update in direction $-q_t^{\mathrm{reset}}$ satisfies the behavioral bound (84) *and loss decrease* (85)*. Moreover, whenever that update is accepted with step length $s_t > 0$ and $r_t \neq 0$, its trial-slot displacement is*

$$M_t d_t = -\frac{s_t}{\|q_t^{\mathrm{reset}}\|} r_t \neq 0,$$

and, because the dormant internal nongate component is identically zero under the invariant, this nonzero slot displacement contains nonzero gate motion. Activation is therefore functionally nonvacuous and covered by the same certificate. If no such accepted motion occurs, rolling back the reset restores the exact pretrial parameter state.

Proof. By construction, a zero functional gate removes the module from both the realized predictor and every active behavior map, so reset has exactly zero immediate functional effect. The post-reset proposal is the ordinary global compatible direction in range(Π_t), not a separately projected partial gradient; Theorem 2.26 therefore applies directly after recomputing the certificates. Applying M_t to the displayed update gives the trial-slot displacement formula. The dormant-slot invariant eliminates nongate internal motion while the gate is zero, so $r_t \neq 0$ implies nonzero gate displacement and therefore a nonzero module contribution after the accepted step. Transactional rollback is exact because the pretrial slot state and zero gate are retained until acceptance. □

Theorem 2.38 (Renewal success is exactly controlled by actual richness)**.** *During stall episode e, use fresh conditional randomness at each trial. Let $I_{e,n}$ indicate a γ-useful batch and let the predictable conditional richness be $\rho_{e,n}(\gamma) = \mathbb{E}[I_{e,n} \mid \mathcal{F}_{e,n-1}]$. For every $x > 0$ and every N,*

$$\mathbb{P}\left(I_{e,1} = \cdots = I_{e,N} = 0, \quad \sum_{n=1}^{N} \rho_{e,n}(\gamma) \geq x \,\middle|\, \mathcal{F}_{e,0} \right) \leq e^{-x}. \tag{102}$$

Consequently, whenever the realized cumulative conditional richness reaches $\log(1/\delta_e)$, failure beyond that point has probability at most δ_e. No known lower bound is required by the algorithm. If every richness value is zero, no sampling algorithm using that renewal family can succeed.

Proof. Define

$$M_N = \mathbf{1}\{I_{e,1} = \cdots = I_{e,N} = 0\} \exp\left(\sum_{n=1}^{N} \rho_{e,n}(\gamma)\right).$$

On the event of no previous success,

$$\mathbb{E}[(1 - I_{e,N})e^{\rho_{e,N}} \mid \mathcal{F}_{e,N-1}] = (1 - \rho_{e,N})e^{\rho_{e,N}} \leq 1.$$

Hence (M_N) is a nonnegative supermartingale with $M_0 = 1$. Markov's inequality gives (102) for each fixed N. Moreover, if $\tau_x = \inf\{N : \sum_{n=1}^{N} \rho_{e,n}(\gamma) \geq x\}$, then failure through τ_x implies $M_{\tau_x} \geq e^x$; Ville's inequality therefore gives $\mathbb{P}(\tau_x < \infty,\ I_{e,1} = \cdots = I_{e,\tau_x} = 0 \mid \mathcal{F}_{e,0}) \leq e^{-x}$, which proves the stated random-hitting-time consequence. The zero-richness statement follows from the definition. □

Corollary 2.39 (Isotropic trial-slot compatible-gradient model)**.** *Condition on the composed linear map $P = M\Pi$ and suppose it is an orthogonal projector of rank h on the trial coordinates. If the post-reset full gradient is $g^{\text{reset}} = \omega z$ with $z \sim \mathcal{N}(0, I_d)$ independent of P, then*

$$\frac{\|Pg^{\text{reset}}\|^2}{\omega^2} \sim \chi^2_h. \tag{103}$$

Hence $h > 0$ and $\omega > 0$ imply explicit positive richness. This corollary is optional; the universal theorem uses the actual richness (64) *and does not assume isotropy or commutation of M and Π.*

Proof. Under the stated projector condition, rotational invariance of the standard Gaussian makes its projection onto the h-dimensional trial-compatible subspace a standard h-variate Gaussian. □

A countable sequence of episodes uses summable δ_e and a fixed structural pool. Failed or unactivated trials reuse their zero-gated slots exactly. A successful activation consumes its slot until a separately certified structural move returns that module to an exactly zero-gated dormant state. If no dormant slot remains, the renewal family's realized richness is zero and the algorithm reports a structural-capacity obstruction and does not allocate additional parameters. Thus lifetime resource use remains fixed; computation and delay are the sums of the trial budgets, with the horizon-dependent precision stated in (106).

2.11 Integrated frontier theorem and convex specialization

The convex branch supplies a global dynamic-regret statement when an exact projection oracle is available. The integrated theorem then combines the pathwise, statistical, resource, and obstruction-aware components of AFM.

2.11.1 Convex affine-behavior dynamic regret

The nonlinear theorem above is pathwise and stationarity-based. A stronger global regret statement requires convexity.

Assumption 2.40 (Convex affine branch)**.** *There is a compact convex set $\mathcal{K} \subset \mathbb{R}^d$ of diameter D. Every ℓ_t is convex and differentiable on a neighborhood of $\mathcal{K}$, with $\|\nabla \ell_t(\theta)\| \leq G$ for every $\theta \in \mathcal{K}$ (and hence is G-Lipschitz on $\mathcal{K}$). The exact retention-compatible set $\mathcal{C}_t \subseteq \mathcal{K}$ is nonempty, closed, and convex; this holds, for example, for affine behavior maps and convex norm tolerances. The initial point satisfies $\theta_1 \in \mathcal{C}_1$, so every projected iterate is retention feasible.*

In AFM's exact-convex solver mode use the exact projection:

$$\theta_{t+1} = \Pi_{\mathcal{C}_{t+1}}(\theta_t - \eta g_t). \tag{104}$$

Let

$$\theta_t^* \in \arg\min_{\theta \in \mathcal{K}} \ell_t(\theta), \qquad \theta_t^\dagger \in \arg\min_{\theta \in \mathcal{C}_t} \ell_t(\theta),$$

and define

$$\Gamma_t = \ell_t(\theta_t^\dagger) - \ell_t(\theta_t^*) \geq 0, \qquad P_T^\dagger = \sum_{t=1}^{T-1} \left\| \theta_{t+1}^\dagger - \theta_t^\dagger \right\|.$$

For the final projection inequality use the harmless convention $\mathcal{C}_{T+1} = \mathcal{C}_T$ and $\theta_{T+1}^\dagger = \theta_T^\dagger$.

Theorem 2.41 (Dynamic regret with unavoidable compatibility price)**.** *Under Assumption 2.40, for every $\eta > 0$,*

$$\sum_{t=1}^{T} \left[\ell_t(\theta_t) - \ell_t(\theta_t^*) \right] \leq \sum_{t=1}^{T} \Gamma_t + \frac{D^2}{2\eta} + \left(\frac{D}{\eta} + G \right) P_T^\dagger + \frac{\eta G^2 T}{2}. \tag{105}$$

The first term cannot be removed by any retention-feasible learner.

Proof. Convexity gives $\ell_t(\theta_t) - \ell_t(\theta_t^\dagger) \leq g_t^T(\theta_t - \theta_t^\dagger)$. Projection nonexpansiveness with comparator $\theta_{t+1}^\dagger \in \mathcal{C}_{t+1}$ yields

$$2\eta g_t^T(\theta_t - \theta_{t+1}^\dagger) \leq \left\| \theta_t - \theta_{t+1}^\dagger \right\|^2 - \left\| \theta_{t+1} - \theta_{t+1}^\dagger \right\|^2 + \eta^2 G^2.$$

The diameter bound implies

$$\left\| \theta_t - \theta_{t+1}^\dagger \right\|^2 \leq \left\| \theta_t - \theta_t^\dagger \right\|^2 + 2D \left\| \theta_{t+1}^\dagger - \theta_t^\dagger \right\|,$$

and $g_t^T(\theta_{t+1}^\dagger - \theta_t^\dagger) \leq G \left\| \theta_{t+1}^\dagger - \theta_t^\dagger \right\|$. Sum and telescope. Finally add Γ_t. □

2.11.2 Integrated frontier theorem

Theorem 2.42 (Maximal Adaptive Functional Metaplasticity Frontier)**.** *Fix any finite horizon T and any nonanticipating stream-generating law, with the convention that at each round the current data item and loss function are fixed before the controller's current private policy or renewal draw. Conditional on every realized stream and learner history, the deterministic retention, endpoint-emulation, stationarity, obstruction, and resource statements below hold pathwise whenever their declared certificates accept. For those clauses that invoke conditional-risk, null, iid population, policy-randomization, or renewal-richness assumptions, suppose the corresponding stated conditions hold under the stream-generating law. Run AFM with fixed structural budgets and summable statistical error allocations. Allocate the total failure budget δ across consolidation, outcome reopening, observable route splitting, optional reference calibration, optional population certificates, policy randomization, refresh or append events, and renewal episodes; for an unbounded lifetime, use any summable allocation across doubling epochs and event indices. AFM executes a nonzero step only when the pre-step certificates* (15)-(16) *and every invoked population certificate are valid; Proposition 2.4 supplies them constructively for a broad finite-graph class; otherwise it sets the trust radius to zero. Then, jointly over the stochastic observations covered by the invoked statistical assumptions and AFM's private randomization, with probability at least $1 - \delta$ all probabilistic conclusions below hold simultaneously, while the pathwise conclusions hold on every realized trajectory.*

(i) ***Operational task-free routing, constructive safe transfer, and consolidation.*** *Routing is defined on every stream and exact under the non-oracular separation conditions of Theorem 2.13. Frozen-representation calibration is honest in the sense of Proposition 2.14: inadequate calibration produces an explicit obstruction and no positive routing claim. A finite predeclared signature family retains lifetime false-alarm control by Proposition 2.15. Dual-evidence route refinement has lifetime false-split probability at most its independently allocated signature budget by Proposition 2.18, even when predictive contradiction is real, and it never deletes source protection. Every committed record satisfies its first-crossing time-uniform average conditional risk certificate* (68)*, the separately controlled staleness condition of Proposition 2.17, and its predeclared activation-gap admission bound. Every accepted shielded candidate-transfer step satisfies Proposition 2.8: it reproduces the full ordinary current endpoint, exactly preserves active empirical behaviors and unselected candidates, and completes the selected empirical transfer in one service under $\xi = 1$. Theorem 2.2 gives the declared deterministic*

construction for consistent finite constraints and a sharp functional obstruction for contradictory duplicate features. Proposition 2.10 is predictable and non-starving, and Theorem 2.11 gives the displayed finite service bound without a projected-compatibility assumption. No uncommitted candidate receives a retention claim.

(ii) ***Whole-behavior retention.*** *On every accepted protected round the persistent safe-base leg obeys the budget-controlled fixed-shield bound* (84)*. The complete declared finite deployed evidence is stronger: it is restored exactly on every accepted protected step, so its active-interval drift is zero up to the declared arithmetic envelope and its total drift from the validated snapshot is only the observed activation-transfer gap. Any broader deployed or evaluator map carries the explicit internal structural-shield charge ω_t^S, the structural routing-mismatch charge ω_t^{Δ}, and the fixed-shield routing derivative terms in* (89)-(90)*. Proposition 2.28 separates activation, exact active finite retention, and post-release drift.*

(iii) ***Arbitrary non-convex adaptation.*** *The persistent metaplastic base satisfies the exact pathwise projected-stationarity inequality* (94)*. In counterfactual-normalized mode, every accepted protected round additionally stores at least the persistent current-loss fraction* (31)*, with the compatible-gradient fraction κ_t displaying the explicit local compatibility price for the selected protected projector; $\eta_t = 1$ recovers the full projected comparator. Independently, every accepted deployed update reproduces one hundred percent of the genuine no-protection current endpoint by* (38)*. On tails satisfying Corollary 2.30, finite cumulative certified safe-base behavior charge, exact finite deployed protection, and realized lifelong compatible base plasticity coexist; a broader joint map additionally pays its explicit structural-shield charge.*

(iv) ***Metaplastic timescale optimality within the declared family.*** *Whenever the declared finite family contains the normalized scale-free coefficient choices above, the logarithmic EMA bank contains constant-distortion scale-free policies by Theorem 2.32, whereas one exponential has polynomial distortion by Theorem 2.33. The deterministic transfer bound* (100) *converts temporal prediction error into excess leakage, while* (101) *competes on the realized composite frontier cost* (98)*, including the gradient energy blocked by protection.*

(v) ***Evidence-based reopening and bounded route refinement.*** *False outcome reopening and false observable splitting are controlled by separate lifetime allocations. Under* (62)*, false reopening is bounded by α_j^{open}; under* (72)*, false splitting is bounded by $\alpha_j^{\mathrm{split}}$ even when a real semantic conflict violates the outcome null. Theorem 2.21 gives finite reopening delay under positive outcome information, and Theorem 2.19 gives finite split delay only when both outcome and distinct-block signature information are positive and the effect/capacity gates hold. With semantic-only change, ordinary reopening or release remains valid but no observable route separation is claimed.*

(vi) ***Structural renewal.*** *Every reset is exactly preserving for the realized predictor and all active protected behaviors. The probability and delay of finding nonzero trial-slot motion inside the global compatible descent direction are controlled by the realized richness sequence through* (102)*. Failed trials reuse fixed slots; successful activations consume them until separately certified exact dormancy is restored. Exhausted structural capacity and zero richness are reported as obstructions, not replaced by assumptions.*

(vii) ***Bounded structural resources and explicit precision.*** *The number of persistent scalar or fixed-size vector states is*

$$\begin{aligned} W = {} & O(d) + O(L_{\max} d) + O(C_{\max} \ell_{\mathrm{cand}} d) + O((J_{\max} + C_{\max}) d) + O(C_{\max} d_Z) \\ & + O(C_{\max} m_{\mathrm{fit}} d_{\mathrm{ref}}) + O((J_{\max} + C_{\max}) n_{\max} (d_{\mathrm{ref}} + d_\Phi)) \\ & + O(K A_{\max}) + O(PK) \\ & + O(J_{\max} + C_{\max}) + O(N_{\mathrm{shield}} (d_Z + d_\Phi + 2)) + O(Q_{\max} d_Z) + O(\textit{renewal slots}), \end{aligned} \tag{106}$$

independent of elapsed lifetime. Here $B_{\max}$ is the declared current-minibatch cardinality bound and $N_{\mathrm{shield}} \le B_{\max} + J_{\max} n_{\max} + C_{\max} n_{\max}$ is the fixed cardinal-node cap and $Q_{\max}$ is the fixed label-free

guard-address cap. The $C_{\max}$ terms include one private candidate initialization or frozen snapshot, the finite m_{fit} fitting-reference block before freezing, bounded staging sketches, frozen first-crossing validation sufficient statistics and evidence, fixed candidate-transfer ledger and queue state, and fixed sufficient state for the separately allocated staleness, outcome, and distinct-block signature e-processes. The fitting block is deleted from decision state after the frozen snapshot is produced. The empirical endpoint-check mode stores at most $n_{\max}$ fixed-size observation references and behavior outputs per candidate or active segment; a certified population-map mode may omit the active-segment evidence term after commitment. Referenced observations belong to the external input source, not to learner-owned adaptive state. Likewise, an optional append-only execution log is external output never read by the algorithm and is not counted as persistent decision state; retaining such a log is an experimental reproducibility choice; because the algorithm never reads it, it is not part of persistent decision state. For an exact guarantee through horizon T, cumulative counters, log-wealth, and error-allocation indices require $O(\log T + \log(1/\delta))$ bits per such scalar, in addition to the declared numerical precision for model and sketch entries; certified rounding errors are included in the same obstruction certificates. A fixed-bit implementation therefore declares a maximum certified horizon or an explicit restart/precision policy. New anchors are append-only while active; when the registry is full, any deletion is an explicit capacity release whose guarantee ends and whose heavy segment state is removed.

(viii) ***Convex global performance.*** *Under Assumption 2.40, an admissible specification that supplies a deterministic exact Euclidean projection oracle for each closed convex retention set instantiates the exact-convex safe-base mode and satisfies* (105)*. The compatibility price $\sum_t \Gamma_t$ is unavoidable. Oracle failure or uncertified feasibility is an explicit convex-oracle obstruction; no uncertified approximate solver is substituted.*

The theorem is universal in the following precise sense: it does not require routing mismatch, spectral tails, anchor drift, curvature, conflict, variation, test delay, or renewal delay to be small. When they are small, the displayed inequalities yield strong continual learning. When they are large, the lower bounds below show why no bounded learner can generally remove them.

Proof. Item (i) combines the routing algorithm, Theorem 2.13, Propositions 2.14, 2.15, 2.18, 2.17, 2.10, and 2.8, and Theorems 2.16 and 2.11, with a union bound over summably allocated records and statistical processes. The optional finite initialization prefix supplies the arbitrary initial state; its bounded references are replayed through the final frozen representation before signature calibration. Item (ii) follows from Theorems 2.22, 2.2, and 2.26, Corollary 2.27, and Proposition 2.28. Item (iii) combines Theorems 2.1, 2.2, 2.26, and 2.29 with Corollary 2.30. Item (iv) intersects Theorems 2.32, 2.33, 2.34, and 2.35. Item (v) is Proposition 2.18 and Theorems 2.20, 2.21, and 2.19. Item (vi) is Theorems 2.37 and 2.38. Item (vii) follows by counting fixed arrays, bounded empirical evidence, the declared $A_{\max}$ trace bank, and sketches; Frequent Directions does not store the conceptual row stacks. Exact counters up to T and confidence levels down to δ have the stated logarithmic bit length. Item (viii) is Theorem 2.41 under the supplied exact projection-oracle condition. The deterministic implications in items (i)-(viii) are pathwise once their premises and certificates are fixed. The union of all invoked statistical, policy-randomization, and renewal failure events has probability at most δ under the stated stream-generating law by the summable allocation. □

2.12 Obstructions and maximality

The main theorem does not assume away incompatibility, observational ambiguity, finite memory, or finite structural capacity. The following constructions show why the corresponding terms or abstentions are necessary.

2.12.1 Matching obstructions and maximality

Proposition 2.43 (Routing unidentifiability)**.** *If two semantic contexts induce the same law for every observable context signature and outcome available to the router, but require different retained behaviors, then no task-free router can identify the correct context with error probability below $1/2$ under an equal prior.*

More specifically, if outcome laws differ but the declared signature law is identical, outcome evidence can justify reopening while no task-free test can justify a new observable route with both nontrivial power and distribution-free false-split control. Consequently, a nonzero routing-mismatch term or outcome-only release is unavoidable.

Proof. For the fully identical case the two hypotheses induce identical observation distributions, so the total variation distance is zero and binary testing cannot beat random guessing. In the semantic-only case, condition on the outcome evidence that reveals invalidity: the marginal law of the declared signatures remains identical under "same route" and "new route" hypotheses, so any signature-based split test faces the same zero-total-variation obstruction. □

Proposition 2.44 (No distribution-free future-risk certificate from a finite commit history)**.** *For every consolidation rule that commits after a finite observed history, there exist two data streams that are identical through the commit time but have different future outcome laws, one on which the committed predictor remains valid and one on which its future risk is arbitrarily poor. Therefore Theorem 2.16 can certify its observed conditional evidence sequence, while a future-risk guarantee requires an explicit recurrence, exchangeability, drift, or other predictive assumption.*

Proof. Fix the realized finite history that triggers commitment. Extend it in one world with outcomes predicted perfectly by the snapshot and in the other with outcomes chosen to maximize its bounded loss. The learner has identical information at commitment in both worlds. □

Proposition 2.45 (Curvature obstruction)**.** *For every linear operator J with nontrivial nullspace and every $H > 0$, there is a smooth behavior map into an enlarged Hilbert space whose derivative at zero has first component J and whose change along every $d \in \mathrm{null}(J)$ equals $H \|d\|^2/2$ in an orthogonal output coordinate.*

Proof. Use $\widetilde{\Phi}(\theta) = (J\theta, (H/2)\|\theta\|^2)$. □

Proposition 2.46 (Anchor-drift obstruction)**.** *A certificate based only on the derivative at a frozen anchor cannot generally control current leakage without a term proportional to representation drift. Specifically, for every $L > 0$ and displacement a, the one-dimensional map*

$$\Phi(\theta) = \frac{L}{2}\theta^2$$

has $D\Phi(0) = 0$ but $|D\Phi(a)| = L|a|$. Hence any universally valid anchor-based derivative certificate must charge an order-$L\left\|\theta - \bar{\theta}\right\|$ term or refresh the anchor.

Proof. Differentiate the displayed map. At the anchor the sensitivity is zero, while at displacement a it is La. This saturates the Lipschitz-Jacobian transport order used in (56). □

Proposition 2.47 (Compatibility lower bound)**.** *Every learner constrained to $\mathcal{C}_t$ has regret relative to the unconstrained optimum at least Γ_t on round t. Hence $\sum_t \Gamma_t$ in (105) cannot be removed.*

Proof. By definition, $\ell_t(\theta) \geq \ell_t(\theta_t^\dagger) = \ell_t(\theta_t^*) + \Gamma_t$ for every $\theta \in \mathcal{C}_t$. □

Proposition 2.48 (Reopening information lower bound)**.** *Let validity and obsolescence generate iid laws P_0, P_1 with $I = D_{\mathrm{KL}}(P_1\|P_0)$, where $0 < \alpha < 1 - \beta < 1$. If a rule stops by time N with probability at least $1 - \beta$ under P_1 and at most α under P_0, then*

$$NI \geq \mathrm{kl}(1 - \beta, \alpha).$$

If $I = 0$, reliable finite-delay reopening is impossible.

Proof. Apply data processing for KL divergence to the event that the rule stops by time N. □

Proposition 2.49 (Renewal zero-richness obstruction)**.** *If $\rho_{e,n}(\gamma) = 0$ for every trial, no number of samples from that renewal family can produce a γ-useful compatible direction.*

Proof. The union of countably many probability-zero success events has probability zero. □

Proposition 2.50 (Finite-memory obstruction)**.** *A learner with at most B persistent bits cannot exactly recall every sequence of $T > B$ independent unbiased bits.*

Proof. There are 2^T histories and at most 2^B persistent states, so two histories collide and require different answers to some recall query. □

Proposition 2.51 (Finite-state lifetime-testing obstruction)**.** *Consider a detector with finitely many nonalarm states under an iid null on a finite alphabet with full support. Suppose that from every nonalarm state there exists an observation word that triggers alarm. Then the detector eventually false alarms with probability one. Consequently, a truly finite-state detector cannot have both false-alarm probability below one over an unbounded lifetime and nonzero eventual sensitivity from every state; exact anytime guarantees require growing precision, an externally supplied clock/epoch policy, or sacrificed delayed sensitivity.*

Proof. Because the nonalarm state set is finite, choose for each state a triggering word and let L be their maximum length. Full support and finiteness give a uniform lower bound $p > 0$ on the probability of the selected word from any state. Conditional on survival at the start of each successive block of length L, alarm occurs within the block with probability at least p. Hence survival through m blocks is at most $(1-p)^m \to 0$. □

The operational scope of consolidation is forced by Proposition 2.44; the anchor-drift order is forced by Proposition 2.46; the exact spectral lower bound is Theorem 2.24; the single-timescale obstruction is Theorem 2.33; dynamic-regret lower bounds of matching path-variation order are known in online convex optimization Yang et al. (2016); Zhang et al. (2018).

Corollary 2.52 (Componentwise maximality of the universal claim)**.** *No bounded task-free learner can, over every realized stream, simultaneously guarantee all of the following with the corresponding obstruction term deleted: always-correct semantic routing; future-valid consolidation from an arbitrary finite past; zero finite-rank functional leakage under nonzero plastic movement; zero nonlinear or anchor-drift damage; exact adaptation to directly incompatible objectives; finite-delay reopening at zero observational information; successful renewal at zero compatible richness; exact recall of unbounded independent information with fixed bounded-bit persistent state; and exact infinite-horizon testing with both finite state and sensitivity after arbitrary delay. Therefore every universal theorem retaining all AFM mechanisms must either display these terms, abstain, release information, or impose assumptions that make them small.*

Proof. Apply, respectively, routing unidentifiability, Proposition 2.44, Theorem 2.24, the curvature and anchor-drift obstructions, the compatibility lower bound, the reopening information lower bound, the zero-richness obstruction, the finite-memory counting argument, and Proposition 2.51. Each construction is a member of the universal stream class, so deleting any associated term falsifies the corresponding universal guarantee. □

These results establish maximality of the theorem's structure and quantifiers, not sharpness of every numerical constant or minimax optimality for every restricted subclass.

2.12.2 Guarantee and obstruction summary

The integrated result couples eight mechanism families. Task-free routing is guaranteed under observable separation and otherwise retains an explicit routing-mismatch term. Consolidation couples first-crossing validation, staleness control, finite transfer service, and atomic rollback; contradictory output requests or unavailable service remain explicit obstructions. Whole-behavior memory provides deterministic derivative-leakage control subject to spectral tail, finite sampling, anchor drift, curvature, and deliberate release.

Table 1: Fidelity of the controlled compatibility intervention

System	Mean κ at request 0	Nonzero MAE	Maximum error	Mean Δ_0 spread	Maximum spread
CNN	0.010681	1.1×10^{-5}	1.16×10^{-4}	0.283%	0.391%
ViT	0.041266	4.0×10^{-6}	8.9×10^{-5}	0.282%	0.396%
Text transformer	0.000083	5.0×10^{-6}	1.30×10^{-3}	0.268%	0.390%

Compatibility errors exclude the requested-zero condition. Within-state unrestricted-progress spread is $(\max_\kappa \Delta_0 - \min_\kappa \Delta_0)/\mathrm{mean}_\kappa \Delta_0$. The zero request is a minimum-compatibility boundary condition in the two vision systems.

Metaplastic allocation competes within the declared timescale/rank family while finite rank leaves unavoidable residual leakage. Compatible adaptation gives certified descent and projected stationarity, with direct conflict outside the protected tangent space. Reopening and route refinement require outcome and, for splitting, observable-signature information. Structural renewal is exactly function preserving at zero gate and succeeds only when compatible richness is present. Finally, the fixed structural resources require explicit finite precision over a certified horizon, and the convex branch pays the unavoidable compatibility cost in dynamic regret. These are the same mechanism-guarantee-obstruction dependencies used in the proof of the integrated theorem.

3 Controlled causal compatibility experiment

3.1 Compatibility as a causal variable

We reconstructed 750 fully unfrozen pre-update states from three systems: a CIFAR-10(Krizhevsky, 2009) convolutional network, a CIFAR-10 vision transformer(Dosovitskiy et al., 2021) and a character-level text transformer(Vaswani et al., 2017) trained on WikiText-2.(Merity et al., 2017) Every intervention at a state began from the same model, optimizer, stream position, replay reservoir and random-number-generator state. We constructed six learning signals spanning minimum to near-complete compatibility through a generalized function-space eigenproblem, recomputed realized κ, and matched the genuine unrestricted finite decrease Δ_0 across compatibility conditions within each state. The mean within-state spread in Δ_0 was only 0.27-0.28% and never exceeded 0.40% (Table 1). Each signal was then evaluated under the same absolute retention budget.

Changing compatibility changed persistent learning. At the intermediate retention budget $\beta = 0.5$, the slope of the persistent-progress ratio $\rho = \Delta_{\mathrm{persistent}}/\Delta_0$ on realized κ for the projected AFM-compatible proposal was 1.014 in the convolutional network, 1.001 in the vision transformer and 0.965 in the text transformer (Fig. 1b). All seed-level slopes were positive. The effect was not a universal property of every update rule: unrestricted, distillation and EWC(Kirkpatrick et al., 2017) branches converted compatibility less efficiently, while replay and DER++(Buzzega et al., 2020) showed little seed-resolved compatibility slope under this matched test (Fig. 1d). Nor was κ sufficient by itself. Tight retention budgets reduced the exploitable path length, particularly in the text system; as the budget relaxed, the compatibility slope approached one (Fig. 1c). Thus the experiment separates *available compatibility* from the ability of a learning rule and retention budget to exploit it.

The following sections document the matched-state causal intervention, statistical analysis, natural-state validation and theorem-aligned mechanism audit used to test functional compatibility as an intervened variable.

3.2 Purpose and scope

The purpose of this experiment is not to add another benchmark comparison to a continual-learning method paper. It is to isolate a more general scientific question:

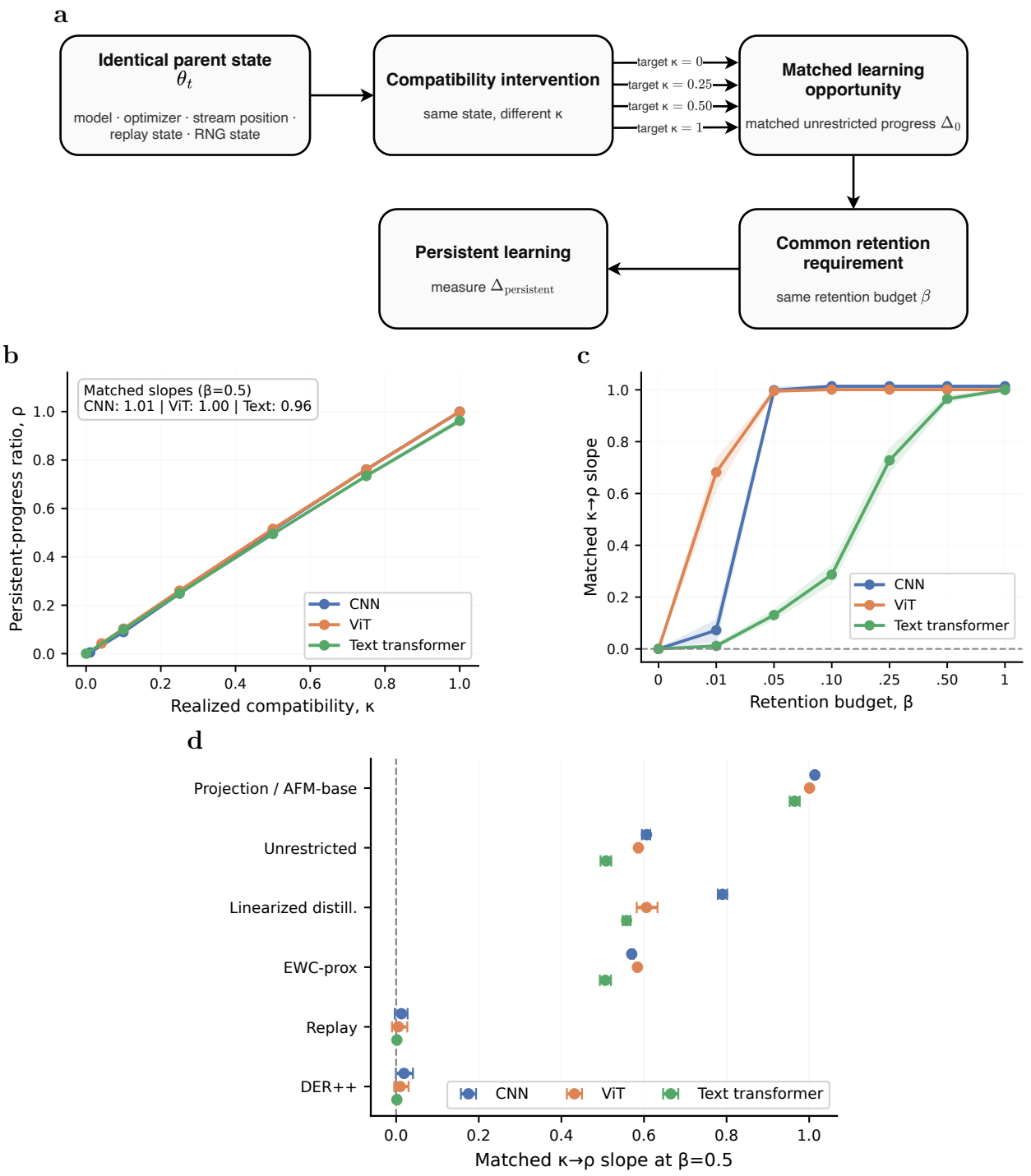

Figure 1: **Functional compatibility causally controls persistent learning. a**, Experimental logic. All interventions at a causal state start from the identical parent neural state, including the same model, optimizer, stream position, replay state and random-number-generator state. Functional compatibility is manipulated while unrestricted learning progress Δ_0 is matched, after which the same retention requirement is applied and persistent learning is measured. **b**, Persistent-progress ratio versus realized functional compatibility at $\beta = 0.5$ in three fully unfrozen systems. **c**, Compatibility slope as a function of retention budget. **d**, Matched compatibility slopes for distinct learning rules at $\beta = 0.5$. For **b**–**d**, $n = 5$ independent seed trajectories per system and 50 reconstructed causal states per seed; state-level slopes are aggregated within seed, and error bars show two-sided 95% Student-t intervals across the five seed effects.

Table 2: Full compatibility slopes across retention budgets

β	CNN	ViT	Text transformer
0	0.000 [0.000,0.000]	0.000 [0.000,0.000]	0.000 [0.000,0.000]
0.01	0.072 [0.008,0.136]	0.682 [0.592,0.772]	0.012 [0.002,0.021]
0.05	0.999 [0.981,1.016]	0.996 [0.987,1.005]	0.131 [0.103,0.158]
0.10	1.014 [1.010,1.017]	1.001 [1.000,1.002]	0.287 [0.233,0.341]
0.25	1.014 [1.010,1.017]	1.001 [1.000,1.002]	0.728 [0.660,0.795]
0.50	1.014 [1.010,1.017]	1.001 [1.000,1.002]	0.965 [0.945,0.985]
1.00	1.014 [1.010,1.017]	1.001 [1.000,1.002]	1.000 [1.000,1.000]

Entries are mean seed-level slopes of persistent-progress ratio on realized compatibility for the projected AFM-compatible proposal with 95% Student-t intervals.

> When the pre-update neural state, protected evidence, current inputs, and learning scale are controlled, does changing the functional compatibility of the incoming learning signal change the amount of current learning that can be stored persistently while satisfying a fixed retention criterion?

The controlled experiment addresses this question causally by intervening on compatibility. The natural-state validation addresses a complementary observational question by measuring compatibility as it arises during ordinary continual learning without constructing compatibility-targeted teacher signals.

The causal compatibility study is distinct from the chronological AFM benchmark programme in both design and statistical treatment. It is not an independent laboratory replication; it is a controlled intervention study built from reconstructed states of the evaluated systems. Numerical claims are limited to procedures and quantities that were explicitly recorded or recomputed.

3.3 Experimental systems and state reconstruction

Three fully trainable systems are used:

Table 3: Experimental matrix. All neural representations remain unfrozen during the reported causal and natural-state experiments.

System	Modality	Seeds	Causal states/seed	Natural states/seed
CIFAR-10 CNN	vision	11, 29, 47, 71, 101	50	50
CIFAR-10 ViT	vision	11, 29, 47, 71, 101	50	50
Text transformer	text	11, 29, 47, 71, 101	50	50

The two vision systems use CIFAR-10 (Krizhevsky, 2009); the vision-transformer system follows the Vision Transformer architecture (Dosovitskiy et al., 2021). The text system uses the WikiText-2 training corpus (Merity et al., 2017) obtained from Salesforce's published dataset distribution and is evaluated as a character-level transformer (Vaswani et al., 2017).

For each system-seed trajectory, the experiment resumes from a saved pre-probe parent checkpoint containing the complete model, optimizer, stream position, replay reservoir and random-number-generator state. This is important for causal matching: every method and compatibility intervention at a given causal state starts from the same pre-update neural parameters and the same stored history rather than from separately evolved trajectories.

The design contains 750 causal states:

$$3 \text{ systems} \times 5 \text{ seeds} \times 50 \text{ states} = 750.$$

Each state is evaluated under six requested compatibility conditions and seven method branches:

$$750 \times 6 \times 7 = 31{,}500$$

method-level causal outcomes. These 31,500 rows are repeated measurements nested within 750 causal states and 15 system-seed trajectories. They must not be treated as 31,500 statistically independent experimental units.

3.4 Controlled causal compatibility intervention

3.4.1 Compatibility construction

At a fixed causal state, the current-learning signal is altered in function space rather than by injecting an arbitrary parameter-space gradient. The implementation constructs residual modes through the generalized compatibility eigenproblem

$$J_c P J_c^\top r = \lambda J_c J_c^\top r, \tag{107}$$

and mixes low- and high-compatibility residual modes to target

$$\kappa_{\text{request}} \in \{0,\ 0.1,\ 0.25,\ 0.5,\ 0.75,\ 1\}.$$

The actual compatibility is recomputed after construction:

$$\kappa = \frac{\|Pq\|^2}{\|q\|^2}. \tag{108}$$

All statistical analyses use this realized value rather than assuming that the requested value was achieved exactly.

The six interventions at a given state therefore differ in the direction of the current learning signal while sharing the same pre-update state, protected evidence, and current inputs. This within-state construction is the central causal control.

3.4.2 Matching the genuine unrestricted decrease

Changing a learning direction can also change the magnitude of the finite current-task improvement. To avoid interpreting a trivial change in unrestricted learning difficulty as a compatibility effect, the experiment matches the genuine unrestricted finite decrease across the six compatibility conditions.

For each causal state, the weakest attainable unrestricted endpoint defines a common target decrease. A one-dimensional bisection is then performed along each original direction to match that finite decrease without rotating the direction. The resulting common denominator is

$$\Delta_0 = \mathcal{L}_{\text{current}}(\theta_{\text{pre}}) - \mathcal{L}_{\text{current}}(\theta_{\text{unrestricted}}). \tag{109}$$

The intervention protocol additionally specifies gradient-norm matching. The causal summary used for the present numerical validation does not record the causal gradient norm, so the validation reported here verifies finite-Δ_0 matching directly and does not claim a separate numerical revalidation of gradient-norm matching.

3.4.3 Common retention budget

For each causal state, the reference protected drift is

$$D_{\text{ref}} = \max_{\kappa} D_{\text{unrestricted}}(\kappa). \tag{110}$$

Every method and every compatibility condition at that state is evaluated under the same absolute budget

$$D \le \max\left(10^{-8}, \beta D_{\text{ref}}\right), \tag{111}$$

with

$$\beta \in \{0,\ 0.01,\ 0.05,\ 0.1,\ 0.25,\ 0.5,\ 1\}.$$

Thus β controls retention strictness independently of the requested compatibility level.

Protected drift need not be monotone in nonlinear step scale. The implementation therefore evaluates all 33 candidate endpoint scales from zero to the full proposal and selects the best feasible persistent endpoint rather than assuming that a one-dimensional monotone backtracking search is valid for every generic method.

3.4.4 Proposal mechanisms

Seven method labels are evaluated from the identical pre-update state:

1. unrestricted learning;
2. replay (Rolnick et al., 2019);
3. projection;
4. linearized distillation;
5. EWC-proximal updating (Kirkpatrick et al., 2017);
6. DER++ (Buzzega et al., 2020);
7. the AFM-compatible projected proposal.

In this causal comparison, the generic AFM-compatible proposal is identical to the projection proposal family. Cross-system slope summaries therefore collapse these duplicate branches into a single projection/AFM-base family to prevent double counting.

Native AFM is evaluated separately. It uses its own accepted backtracking fraction $\widehat{\lambda}$, finite functional completion, protected endpoint restoration, and deployed-update checks. Native AFM is therefore not conflated with the generic 33-scale projected frontier.

3.5 Natural-state validation

The natural-state validation removes the compatibility intervention entirely. It reuses the same three architectures, five seeds, and parent checkpoints, but follows the ordinary deterministic continual-learning stream with the true current labels.

Starting from each parent checkpoint, 50 pre-update states are sampled by a fixed schedule, every tenth parent step over the following 500 steps. No state is selected or rejected according to its measured compatibility. At each sampled state the experiment records:

- the supervised current gradient and its norm;
- naturally occurring functional compatibility κ;
- the genuine unrestricted decrease Δ_0;
- method-specific persistent decrease $\Delta_{\text{persistent}}$;
- the normalized persistent progress

$$\rho_{\text{persistent}} = \frac{\Delta_{\text{persistent}}}{\Delta_0}; \tag{112}$$

- protected retention drift and retention-pass status.

Protected examples use a separate deterministic probe RNG. The validation branches are discarded after measurement, and only the ordinary supervised parent update is committed before continuing along the stream. Consequently, measurement does not alter the subsequent parent trajectory through branch-specific parameter updates or reservoir RNG consumption.

The natural validation contains

$$3 \times 5 \times 50 = 750$$

ordinary states and

$$750 \times 7 = 5{,}250$$

method outcomes. The stored retention tolerance is 0.005 for every natural-state row, and the recorded pass indicator agrees exactly with the condition $D \leq 0.005$.

3.6 Independent statistical analysis

3.6.1 Inferential unit

The experimental rows are strongly matched. Six compatibility levels share one causal state, 50 states share one seed trajectory, and five seed trajectories are available per system. Statistical significance should therefore not be calculated by treating all method-level rows as independent.

For the independent causal analysis, a compatibility slope is first computed inside every causal state:

$$b_s = \frac{\sum_j (\kappa_{sj} - \bar{\kappa}_s)(\rho_{sj} - \bar{\rho}_s)}{\sum_j (\kappa_{sj} - \bar{\kappa}_s)^2}, \tag{113}$$

where j indexes the six compatibility interventions. The 50 state slopes are averaged within each seed. The five resulting seed-level slopes are the inferential replicates for each system. The report gives the mean seed slope and a two-sided 95% Student-t interval with four degrees of freedom.

This procedure preserves the matched intervention structure and avoids pseudo-replication. The explicitly seed-resolved intervals are used for the primary causal inference throughout the paper. Aggregate cross-system summaries are reported separately as descriptive results.

3.6.2 Interpretation of the normalized ratio

The ratio $\rho_{\text{persistent}}$ is useful because it normalizes persistent progress by the genuine same-state unrestricted decrease. It is not, however, mathematically bounded above by one for arbitrary proposal mechanisms. A method-specific direction can occasionally produce a larger finite decrease than the reference unrestricted direction.

The ratio is also numerically sensitive when Δ_0 is very small. This matters particularly in the natural CNN data, where the minimum natural Δ_0 is 3.89×10^{-4} and some normalized ratios become very large. Natural-state analysis therefore emphasizes medians, retention status, and AFM's pointwise theorem-aligned margin rather than interpreting the raw mean ratio alone.

3.6.3 AFM theorem-aligned diagnostic

For native AFM, the stored theorem-aligned reference is

$$\rho_{\text{ref}} = \frac{\widehat{\lambda}\kappa}{3}. \tag{114}$$

The empirical margin is

$$M = \rho_{\text{persistent}} - \frac{\widehat{\lambda}\kappa}{3}. \tag{115}$$

This is an empirical theorem-alignment diagnostic. It is not presented as a proof that a nonlinear neural execution satisfies every outward-certified assumption of the analytical theorem.

3.7 Results

3.7.1 Intervention fidelity

Table 4 audits the two most important controls. For all nonzero requested compatibility levels, realized compatibility tracks the target with extremely small mean absolute error. The requested-zero condition is different: the lowest attainable mean realized compatibility is approximately 0.0107 for the CNN, 0.0413 for the ViT, and 8.33×10^{-5} for the text transformer. The zero request should therefore be described as a minimum-compatibility boundary condition, not as an exact $\kappa = 0$ intervention in the two vision systems.

The second control is the finite unrestricted decrease. Within each causal state, the relative spread

$$\frac{\max_\kappa \Delta_0 - \min_\kappa \Delta_0}{\mathrm{mean}_\kappa \Delta_0}$$

averages only 0.27% to 0.28% and never exceeds 0.40% in any system. This is strong evidence that the six causal directions were compared at essentially matched finite unrestricted progress.

Table 4: Audit of causal intervention fidelity. Compatibility errors exclude the requested-zero condition. Relative Δ_0 spread is computed within each matched causal state.

System	Mean κ at request 0	Nonzero MAE	Nonzero max error	Mean Δ_0 spread	Max spread
CIFAR-10 CNN	0.010681	1.1×10^{-5}	1.16×10^{-4}	0.283%	0.391%
CIFAR-10 ViT	0.041266	4.0×10^{-6}	8.9×10^{-5}	0.282%	0.396%
Text transformer	0.000083	5.0×10^{-6}	1.30×10^{-3}	0.268%	0.390%

3.7.2 Causal effect of compatibility under an intermediate retention budget

Table 5 reports the independent matched seed-level slope analysis at $\beta = 0.5$. This value is used as an illustrative intermediate retention budget, not as a retrospectively declared primary endpoint.

The projected AFM-compatible proposal shows a slope very close to one in every system:

$$1.014 \quad \text{CNN}, \qquad 1.001 \quad \text{ViT}, \qquad 0.965 \quad \text{text}.$$

All five seed-level slopes are positive in all three systems, and each system-level 95% interval lies well above zero. The cross-system descriptive mean over the 15 system-seed slopes is 0.993 with a 95% t interval of $[0.980, 1.006]$.

This means that, for the projected compatible proposal under this retention budget, a one-unit increase in realized functional compatibility is accompanied by an approximately one-unit increase in the normalized amount of current learning retained persistently. The result replicates across a convolutional vision model, a vision transformer, and a text transformer with all representations trainable.

Table 5: Matched causal slopes of $\rho_{\text{persistent}}$ on realized κ at $\beta = 0.5$. Each system interval uses five seed-level matched slopes. The final column is a descriptive mean over 15 system-seed slopes and is not a claim about an unobserved population of architectures.

Method	CIFAR-10 CNN	CIFAR-10 ViT	Text transformer	Cross-system
Projection / AFM-base	1.014 [1.010, 1.017]	1.001 [1.000, 1.002]	0.965 [0.945, 0.985]	0.993 [0.980, 1.006]
Unrestricted	0.606 [0.591, 0.622]	0.586 [0.581, 0.592]	0.509 [0.488, 0.529]	0.567 [0.542, 0.592]
Linearized distillation	0.790 [0.772, 0.809]	0.606 [0.567, 0.645]	0.558 [0.542, 0.573]	0.651 [0.593, 0.710]
EWC-prox	0.570 [0.564, 0.577]	0.584 [0.580, 0.589]	0.506 [0.485, 0.528]	0.554 [0.533, 0.574]
Replay	0.013 [-0.013, 0.039]	0.005 [-0.026, 0.036]	0.002 [-0.003, 0.006]	0.007 [-0.004, 0.017]
DER++	0.019 [-0.015, 0.053]	0.010 [-0.019, 0.038]	0.002 [-0.004, 0.008]	0.010 [-0.001, 0.021]

The cross-method result is informative in two directions. Unrestricted, linearized-distillation, and EWC-proximal branches also show clearly positive compatibility slopes under the common $\beta = 0.5$ retention constraint. Replay and DER++ do not show a seed-resolved positive slope in any individual system under this independent analysis. Consequently, the data do not support the statement that every continual-learning algorithm converts available compatibility into persistent progress at the same rate. They support the more precise statement that compatibility changes the available retention-constrained geometry, while proposal mechanisms differ strongly in how efficiently they exploit that geometry.

3.7.3 Retention strength changes the observable compatibility slope

Table 6 shows the projected AFM-compatible proposal across all stored retention budgets. The zero-budget condition is a sanity check: persistent progress is zero and the compatibility slope is zero. For every positive β, the matched slope is positive in all three systems. The two CIFAR-10 systems reach an approximately unit slope by $\beta = 0.05$ to 0.10, whereas the text transformer requires a more permissive retention budget before approaching unit conversion.

Table 6: Retention-budget sensitivity for the projected AFM-compatible proposal. Entries are mean matched seed slopes with 95% Student-t intervals.

β	CIFAR-10 CNN	CIFAR-10 ViT	Text transformer
0	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
0.01	0.072 [0.008, 0.136]	0.682 [0.592, 0.772]	0.012 [0.002, 0.021]
0.05	0.999 [0.981, 1.016]	0.996 [0.987, 1.005]	0.131 [0.103, 0.158]
0.10	1.014 [1.010, 1.017]	1.001 [1.000, 1.002]	0.287 [0.233, 0.341]
0.25	1.014 [1.010, 1.017]	1.001 [1.000, 1.002]	0.728 [0.660, 0.795]
0.50	1.014 [1.010, 1.017]	1.001 [1.000, 1.002]	0.965 [0.945, 0.985]
1.00	1.014 [1.010, 1.017]	1.001 [1.000, 1.002]	1.000 [1.000, 1.000]

This interaction is scientifically important. Compatibility is not asserted to be the only determinant of persistent learning. The retention budget controls how much of an available compatible direction can be accepted. In AFM notation, this is the distinction between the compatibility factor κ and the accepted path fraction $\widehat{\lambda}$.

3.7.4 Native AFM mechanism audit

Native AFM is evaluated independently of the generic 33-scale projection frontier. Across all 3,750 controlled observations with requested $\kappa \in \{0.1, 0.25, 0.5, 0.75, 1\}$, persistent acceptance is 3,750/3,750. The theorem-aligned empirical margin in Eq. (115) is nonnegative in 3,750/3,750 of these accepted observations. Finite functional completion succeeds in 3,745/3,750 cases, or 99.87%. The five unsuccessful finite completions occur in the CNN and are recorded as functional-constraint inconsistencies. No other obstruction type is recorded for these nonzero conditions.

For all successful finite completions, the stored finite current error, finite endpoint error, and protected endpoint error are 0.0 at the recorded precision, and the stored deployed progress ratio is exactly 1.0. These observations verify execution of the finite endpoint-completion mechanism on the declared finite constraints; they must not be interpreted as evidence of population retention away from the finite support.

Table 7 gives the detailed native AFM summary.

The requested-zero boundary is reported separately. Among accepted requested-zero native AFM rows, 65 have a negative theorem-aligned empirical margin: 52 CNN rows, one ViT row, and 12 text rows. The negative margins are small in absolute value, with the most negative value -0.003857. None of these 65 rows had certification of the relevant finite curvature or step condition, so they are not treated as certified theorem violations. The nonzero intervention levels have a uniformly nonnegative theorem-aligned empirical margin.

Table 7: Native AFM controlled-intervention results. "Margin pass" is the fraction of accepted rows with $\rho_{\text{persistent}} - \widehat{\lambda}\kappa/3 \geq 0$. "Finite" is successful finite completion divided by all rows at that condition.

System	Req. κ	Realized κ	Mean $\rho_{\text{persistent}}$	Mean $\widehat{\lambda}$	Min. margin	Accept	Margin pass	Finite
CIFAR-10 CNN	0.00	0.010681	0.00524	0.9519	-0.003857	97.2%	78.6%	97.2%
CIFAR-10 CNN	0.10	0.100000	0.08937	1.0000	0.008058	100%	100%	99.6%
CIFAR-10 CNN	0.25	0.250000	0.24836	1.0000	0.083615	100%	100%	99.6%
CIFAR-10 CNN	0.50	0.499998	0.51097	1.0000	0.241686	100%	100%	99.6%
CIFAR-10 CNN	0.75	0.749999	0.76107	1.0000	0.407596	100%	100%	99.6%
CIFAR-10 CNN	1.00	0.999986	0.99974	1.0000	0.593211	100%	100%	99.6%
CIFAR-10 ViT	0.00	0.041266	0.04206	1.0000	-0.001655	100%	99.6%	100%
CIFAR-10 ViT	0.10	0.100000	0.10326	1.0000	0.025761	100%	100%	100%
CIFAR-10 ViT	0.25	0.250000	0.26034	1.0000	0.089506	100%	100%	100%
CIFAR-10 ViT	0.50	0.500000	0.51547	1.0000	0.274879	100%	100%	100%
CIFAR-10 ViT	0.75	0.750001	0.76052	1.0000	0.418082	100%	100%	100%
CIFAR-10 ViT	1.00	0.999990	0.99964	1.0000	0.649724	100%	100%	100%
Text transformer	0.00	0.000083	0.00054	0.6902	-0.000032	96.8%	95.0%	96.8%
Text transformer	0.10	0.100000	0.09963	1.0000	0.050312	100%	100%	100%
Text transformer	0.25	0.250000	0.24924	1.0000	0.144862	100%	100%	100%
Text transformer	0.50	0.500000	0.49905	1.0000	0.304095	100%	100%	100%
Text transformer	0.75	0.750000	0.74915	1.0000	0.474841	100%	100%	100%
Text transformer	1.00	0.999985	0.99991	1.0000	0.654466	100%	100%	100%

3.7.5 Natural-state validation

The natural-state experiment is intentionally not another causal sweep. Natural compatibility occupies different ranges in the three systems:

$$\begin{aligned} \text{CNN: } 0.067 \text{ to } 0.443, &\quad \text{mean} = 0.228, \\ \text{ViT: } 0.278 \text{ to } 0.618, &\quad \text{mean} = 0.422, \\ \text{text: } 0.903 \text{ to } 0.976, &\quad \text{mean} = 0.950. \end{aligned}$$

This is itself useful evidence that the compatibility regimes are architecture- and stream-dependent.

Native AFM satisfies the stored $D \leq 0.005$ retention criterion on all 750 natural states. Its theorem-aligned margin is positive on all 750 observations. Table 8 reports the natural AFM summary.

Table 8: Native AFM on unmanipulated natural states. Median $\rho_{\text{persistent}}$ is emphasized because the ratio can be heavy-tailed when Δ_0 is small.

System	n	Mean κ	κ range	Mean $\widehat{\lambda}$	Median $\rho_{\text{persistent}}$	Min. margin
CIFAR-10 CNN	250	0.228	[0.067, 0.443]	0.552	0.348	0.0995
CIFAR-10 ViT	250	0.422	[0.278, 0.618]	0.670	0.277	0.0848
Text transformer	250	0.950	[0.903, 0.976]	0.107	0.119	0.0199

The text system illustrates why compatibility alone is not a complete predictor of realized persistent progress. Its natural compatibility is very high, approximately 0.95, but its mean native AFM path fraction is only 0.107. The resulting median persistent ratio is approximately 0.119. The causal and natural experiments therefore support the joint interpretation

$$\text{persistent progress} \quad \text{depends on} \quad \text{available compatibility} \times \text{accepted retention-constrained path},$$

rather than the stronger and unsupported claim that κ alone determines $\rho_{\text{persistent}}$ in every natural state.

The common retention criterion also distinguishes the method branches sharply in the natural experiment.

Table 9: Retention-pass rate on the 250 natural states per system under the stored common tolerance $D \leq 0.005$. This table measures retention feasibility under this specific criterion, not general method quality.

Method	CIFAR-10 CNN	CIFAR-10 ViT	Text transformer
AFM	100.0%	100.0%	100.0%
Projection	12.0%	34.8%	0.0%
Linearized distillation	11.6%	32.0%	0.0%
EWC-prox	0.0%	0.0%	0.0%
Replay	0.0%	0.0%	0.0%
DER++	0.0%	0.0%	0.0%
Unrestricted	0.0%	0.0%	0.0%

These rates should not be converted into a claim that AFM universally dominates the other algorithms. The experiment applies one common strict retention criterion to branch updates generated from a common state. It shows that native AFM's own acceptance mechanism consistently produces a retention-feasible update in these sampled natural states, whereas the other stored branch proposals often do not satisfy that same absolute threshold.

3.7.6 Local causal scale and ordinary natural scale

The controlled intervention deliberately chooses the weakest of six unrestricted endpoints as the common finite-decrease target. This makes the causal probe local, especially in the two vision systems. Table 10 makes this scale difference explicit.

Table 10: Finite unrestricted decrease in the controlled causal experiment and in the ordinary natural-state validation.

System	Causal mean Δ_0	Causal median	Natural mean Δ_0	Natural median
CIFAR-10 CNN	1.64×10^{-6}	1.48×10^{-6}	0.1199	0.1204
CIFAR-10 ViT	8.85×10^{-6}	7.01×10^{-6}	0.1582	0.1535
Text transformer	2.32×10^{-4}	1.67×10^{-4}	0.0270	0.0266

This distinction is central to interpretation. The causal experiment is strong for isolating local geometry because Δ_0 is deliberately matched. By itself it does not establish that the same quantitative slope describes large unrestricted updates. The natural-state validation operates at ordinary stream-generated decreases that are much larger, particularly for vision. Its role is therefore complementary rather than redundant.

3.8 Statistical and scientific significance

The strongest result is not a comparison of average test accuracy. It is the combination of intervention control, replication, and mechanism alignment.

First, realized compatibility is experimentally controllable over nearly the full $[0, 1]$ range, with very small nonzero target error. Second, the finite unrestricted denominator is matched within state to substantially better than 1% relative spread. Third, under a nontrivial common retention constraint, changing realized compatibility produces a large and precisely replicated change in persistent progress for the projected compatible learner. At $\beta = 0.5$, the slope is approximately one in all three fully unfrozen systems, including a text transformer. Fourth, the effect is not a universal property of every proposal rule: methods differ substantially in how efficiently they convert the changed geometry into a retention-feasible persistent endpoint. Fifth, native AFM's own mechanism produces a nonnegative theorem-aligned empirical margin on every nonzero controlled intervention and on every natural-state validation observation.

The most defensible central scientific statement is therefore:

> Functional compatibility is a causal control variable for the retention-constrained persistent-learning frontier. The amount of compatible geometry available to the current update changes how much current progress can be retained persistently, while the realized progress also depends on the accepted path fraction and on how effectively the learning rule exploits the available compatible direction.

The experiment does not support several stronger formulations. It does not establish a universal numerical law for every architecture, modality, loss, or continual-learning algorithm. It does not show that every method has a unit κ-to-$\rho_{\text{persistent}}$ slope. It does not show that natural κ alone predicts $\rho_{\text{persistent}}$ independently of $\widehat{\lambda}$. It does not turn finite endpoint completion into a population-retention statement.

3.9 Scope and limitations of the causal analysis

The controlled intervention and its follow-up studies define a specific empirical scope. Five points are important for interpreting the results.

1. **The requested-zero condition is a minimum-compatibility boundary in the vision systems.** Mean realized compatibility at the requested-zero condition is 0.0107 for the CNN and 0.0413 for the ViT. We therefore describe this condition as minimum compatibility rather than exact zero.

2. **The matched controlled intervention is deliberately local.** The weakest of the six unrestricted directions determines the common positive learning target, making the matched causal probe small relative to ordinary vision updates. This is useful for causal identification but does not imply that the same normalized slope must hold at arbitrary finite step size. The multiscale study and fixed-norm bridge quantify this boundary directly.

3. **Near-zero theorem-aligned margins require the theorem conditions to be distinguished from the diagnostic quantity.** All 65 accepted requested-zero observations with negative theorem-aligned margins lacked certification of the relevant finite curvature or step condition. Numerical verification agreed with the reported compatibility values to within 1.5883×10^{-9} and with the persistent-ratio values exactly. These rows are therefore not counted as certified theorem violations. Nonzero intervention levels have uniformly nonnegative theorem-aligned empirical margins.

4. **Natural-state ratios can be unstable when unrestricted progress is small.** In the CNN natural-state data, small values of Δ_0 can produce large ratios. Medians and absolute persistent progress are therefore reported alongside ratio summaries where this matters.

5. **Natural-state validation is observational.** It demonstrates that compatibility, accepted path length and retention-feasible AFM updates arise during ordinary stream learning, but the controlled matched-state intervention supplies the causal identification. Complete architecture, optimizer, replay, projector and analysis configurations are available with the released code identified in the paper.

3.10 Controlled causal experiment synthesis

The controlled experiment succeeds at the question it was designed to answer. It changes functional compatibility while holding the parent continual-learning state fixed, uses realized rather than requested compatibility in analysis, matches the genuine unrestricted finite decrease across compatibility conditions, and evaluates all methods under the same state-specific retention budget. The resulting projected compatibility slopes are large, reproducible across seeds, and close to one across CNN, ViT, and text-transformer systems once the retention budget permits the compatible direction to be expressed.

The natural-state validation then removes the engineered compatibility targets and demonstrates the same framework on ordinary stream states. Native AFM satisfies the stored retention criterion on all 750 natural states, while the empirical $\rho_{\text{persistent}} - \widehat{\lambda}\kappa/3$ margin is positive on all 750 observations. Together, the causal and natural experiments support a general retention-plasticity interpretation centered on functional compatibility,

but they also show that compatibility must be considered jointly with the accepted path fraction and the update mechanism.

The most important scientific distinction is therefore between a broad geometric result and a method-specific performance claim. The evidence supports the former: functional compatibility causally changes the persistent-learning frontier under matched retention constraints. AFM supplies a theorem-aligned mechanism for measuring and exploiting that geometry.

3.11 Complete cross-system slope summary

For completeness, Table 11 reports the aggregate cross-system point slopes and 95% intervals. The primary inferential analysis in the paper uses the independently reconstructed seed-level effects, which provide the explicit inferential unit and interval construction used for the main claims.

Table 11: Aggregate cross-system slopes.

Method	β	CNN	ViT	Text	Pooled
Projection / AFM-base	0.00	0.000	0.000	-0.000	-0.000
Projection / AFM-base	0.01	0.072	0.682	0.012	0.255
Projection / AFM-base	0.05	0.999	0.996	0.131	0.708
Projection / AFM-base	0.10	1.014	1.001	0.287	0.767
Projection / AFM-base	0.25	1.014	1.001	0.728	0.914
Projection / AFM-base	0.50	1.014	1.001	0.965	0.993
Projection / AFM-base	1.00	1.014	1.001	1.000	1.005
Unrestricted	0.00	0.000	0.000	0.000	0.000
Unrestricted	0.01	0.055	0.537	0.010	0.201
Unrestricted	0.05	0.728	0.769	0.105	0.534
Unrestricted	0.10	0.723	0.743	0.195	0.554
Unrestricted	0.25	0.668	0.679	0.442	0.597
Unrestricted	0.50	0.606	0.586	0.509	0.567
Unrestricted	1.00	0.000	0.000	0.000	0.000
Linearized distillation	0.00	0.000	0.000	0.000	0.000
Linearized distillation	0.01	0.060	0.552	0.010	0.208
Linearized distillation	0.05	0.838	0.826	0.101	0.589
Linearized distillation	0.10	0.910	0.833	0.212	0.652
Linearized distillation	0.25	0.801	0.689	0.483	0.658
Linearized distillation	0.50	0.790	0.606	0.558	0.651
Linearized distillation	1.00	0.789	0.597	0.511	0.632
EWC-prox	0.00	0.000	0.000	0.000	0.000
EWC-prox	0.01	0.057	0.535	0.010	0.200
EWC-prox	0.05	0.731	0.768	0.105	0.535
EWC-prox	0.10	0.728	0.741	0.195	0.555
EWC-prox	0.25	0.678	0.683	0.443	0.601
EWC-prox	0.50	0.570	0.584	0.506	0.554
EWC-prox	1.00	0.010	0.007	0.006	0.008
Replay	0.00	0.000	0.000	0.000	0.000
Replay	0.01	0.000	0.000	0.000	0.000
Replay	0.05	0.000	0.000	0.000	0.000
Replay	0.10	0.001	0.000	0.000	0.001
Replay	0.25	0.007	0.002	0.001	0.003
Replay	0.50	0.013	0.005	0.002	0.007
Replay	1.00	0.022	0.009	0.001	0.011

Method	β	CNN	ViT	Text	Pooled
DER++	0.00	0.000	0.000	0.000	0.000
DER++	0.01	0.000	0.000	0.000	0.000
DER++	0.05	0.001	0.000	0.000	0.001
DER++	0.10	0.003	0.003	0.001	0.002
DER++	0.25	0.006	0.002	0.003	0.004
DER++	0.50	0.019	0.010	0.002	0.010
DER++	1.00	0.049	0.020	0.000	0.023

3.12 Natural-state descriptive ratios

Table 12 reports both means and medians to expose ratio skew. The unrestricted branch is exactly one by construction because it defines Δ_0. Large positive or negative means in other CNN branches arise from a small number of states with small Δ_0; the corresponding medians are substantially more stable.

Table 12: Natural-state persistent-progress ratios and retention pass rate.

System	Method	Mean $\rho_{\text{persistent}}$	Median $\rho_{\text{persistent}}$	Retention pass
CIFAR-10 CNN	AFM	0.937	0.348	100.0%
CIFAR-10 CNN	Projection	1.787	0.631	12.0%
CIFAR-10 CNN	Linearized distillation	1.789	0.632	11.6%
CIFAR-10 CNN	EWC-prox	1.075	1.008	0.0%
CIFAR-10 CNN	Replay	-13.393	-3.674	0.0%
CIFAR-10 CNN	DER++	-2.477	-0.223	0.0%
CIFAR-10 CNN	Unrestricted	1.000	1.000	0.0%
CIFAR-10 ViT	AFM	0.341	0.277	100.0%
CIFAR-10 ViT	Projection	0.503	0.503	34.8%
CIFAR-10 ViT	Linearized distillation	0.505	0.505	32.0%
CIFAR-10 ViT	EWC-prox	0.999	0.999	0.0%
CIFAR-10 ViT	Replay	0.651	0.748	0.0%
CIFAR-10 ViT	DER++	0.929	0.941	0.0%
CIFAR-10 ViT	Unrestricted	1.000	1.000	0.0%
Text transformer	AFM	0.102	0.119	100.0%
Text transformer	Projection	0.950	0.952	0.0%
Text transformer	Linearized distillation	0.952	0.953	0.0%
Text transformer	EWC-prox	0.999	1.000	0.0%
Text transformer	Replay	1.003	0.999	0.0%
Text transformer	DER++	1.005	1.004	0.0%
Text transformer	Unrestricted	1.000	1.000	0.0%

4 Generality, natural-state behaviour, and finite-scale boundary

4.1 Robust across directions and models

A causal variable should not depend on one specially selected direction. We therefore generated four independent learning directions at each nonzero compatibility level. At $\beta = 0.5$, the within-κ variation across these directions was only 3.8%, 2.1% and 3.0% of the between-κ variation in the convolutional network, vision transformer and text transformer, respectively, while the corresponding compatibility slopes were 1.046, 1.029 and 0.986 (Fig. 2a,b). This substantially weakens the alternative explanation that the original dose response arose from a particular generalized-eigenmode construction.

Table 13: Independent-direction robustness across budgets

β	CNN R/slope	ViT R/slope	Text R/slope
0.01	0.3718 / 0.0725	0.2467 / 0.2740	0.4353 / −0.0066
0.05	0.0922 / 0.6876	0.0409 / 0.8535	0.5321 / 0.0927
0.10	0.0577 / 0.8917	0.0220 / 0.9566	0.3445 / 0.3367
0.25	0.0400 / 1.0130	0.0254 / 1.0133	0.1151 / 0.8285
0.50	0.0379 / 1.0465	0.0206 / 1.0294	0.0302 / 0.9861
1.00	0.0347 / 1.0556	0.0205 / 1.0295	0.0178 / 1.0033

$R = R_{\mathrm{dir}/\kappa}$ is within-compatibility standard deviation across four independently constructed directions divided by between-compatibility variation.

Table 14: Ten-seed pooled compatibility response

β	Equal-system pooled slope	95% CI
0.01	0.36331	[0.35075,0.37667]
0.05	0.77171	[0.76572,0.77759]
0.10	0.82114	[0.81433,0.82756]
0.25	0.93121	[0.92425,0.93841]
0.50	0.99360	[0.99123,0.99600]
1.00	1.00379	[1.00326,1.00442]

Equal-system pooled AFM/projection slopes over ten common seed trajectories in the CNN, original ViT, stronger ViT and text transformer. Intervals use 100,000 deterministic seed-bootstrap resamples.

We next doubled the number of seed trajectories and added a stronger vision transformer. Across ten seeds, the $\beta = 0.5$ slopes were 1.015 for the convolutional network, 1.001 for the original vision transformer, 0.999 for the stronger vision transformer and 0.960 for the text transformer; the equal-system pooled slope was 0.994 (Fig. 2c). A separate multiscale experiment enlarged the designed intervention while preserving the matched causal construction. The vision slopes remained close to one over this expanded local regime (Fig. 2d). Importantly, calibration against ordinary training showed that even the largest such interventions remained far smaller than natural updates in vision. The multiscale result therefore establishes robustness of the local causal principle, not invariance at arbitrary step size.

4.2 Compatibility in ordinary learning

Compatibility also arose spontaneously during unmanipulated continual learning. Across 750 ordinary states, mean κ was 0.228 in the convolutional network, 0.422 in the vision transformer and 0.950 in the text transformer (Fig. 3a). Native AFM satisfied the common recorded retention criterion on all 750 states. Yet the text model, despite its high compatibility, accepted only a mean persistent path fraction of 0.107 and had a median persistent-progress ratio of 0.119, compared with path fractions of 0.552 and 0.670 in the two vision systems (Fig. 3b). Natural-state behaviour therefore reinforces the causal result while showing why compatibility cannot be interpreted as a complete scalar law: a compatible direction must still fit inside the permitted retention path.

AFM provides a constructive realization of this geometry rather than only a measurement procedure (Fig. 3d). In the controlled causal suite, every nonzero native-AFM request accepted a persistent update and every such row had a nonnegative theorem-aligned empirical margin; finite completion succeeded in 3,745 of 3,750 nonzero cases. The chronological programme evaluated AFM on CORe50(Lomonaco & Maltoni, 2017), CLEAR-10(Lin et al., 2021) and CLAD-C(Verwimp et al., 2023) over five frozen seeds per protocol. AFM produced retention-plasticity hypervolumes of 0.248, 0.381 and 0.513, respectively, and exceeded the strongest tested classical non-oracle family on every paired seed in all three protocols. Against six recent challengers

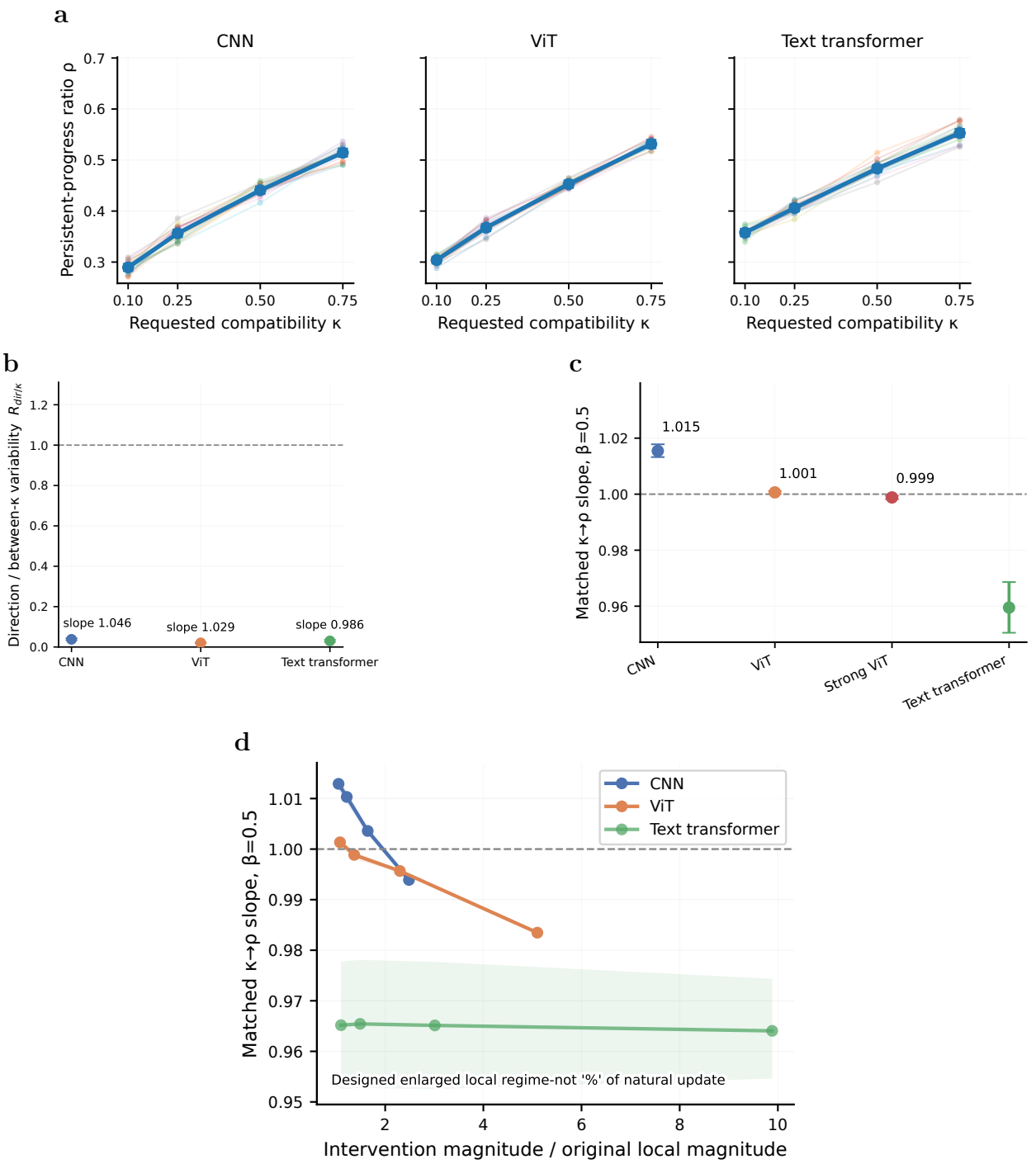

Figure 2: **The compatibility effect generalizes across directions, models and local scales.** **a**, Four independently constructed directions at each nonzero compatibility level; light traces and points show individual directions and the dark line shows the mean. **b**, Ratio of within-compatibility directional variation to between-compatibility variation at $\beta = 0.5$. For **a,b**, $n = 5$ seed trajectories per system, 50 states per seed, and four directions at each of four nonzero compatibility levels. **c**, Ten-seed compatibility slopes for the convolutional network, original vision transformer, stronger vision transformer and text transformer; $n = 10$ seed trajectories per system and intervals use 100,000 deterministic seed-bootstrap resamples. **d**, Multiscale compatibility slopes in the designed local expansion; $n = 5$ seed trajectories per system and 50 states per seed. The scale coordinate is an enlargement of the local intervention and is not a percentage of a natural training update.

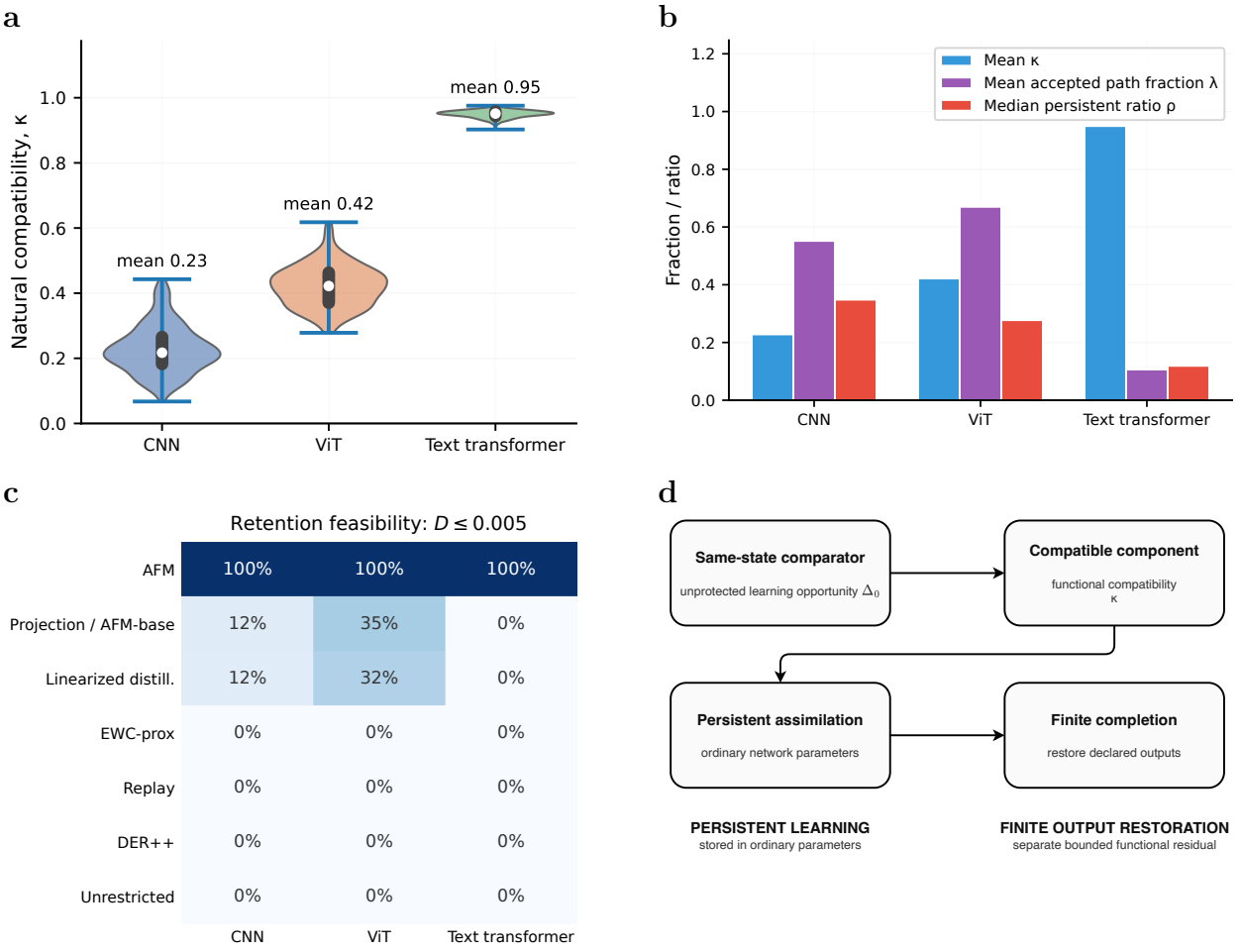

Figure 3: **Functional compatibility arises during ordinary learning and AFM converts it into a protected learning frontier. a**, Natural compatibility distributions across 750 unmanipulated states, comprising $n = 5$ seed trajectories and 50 states per seed in each of three systems. **b**, Natural compatibility, accepted AFM path fraction and persistent-progress ratio. **c**, Retention-pass rates under the common recorded $D \leq 0.005$ criterion; rates describe feasibility under this criterion rather than universal method quality. **d**, Schematic of the AFM transaction separating persistent learning in ordinary network parameters from finite output restoration.

under a common predictive-base and learner-visible-information interface, AFM had the highest dataset-mean primary score on CLEAR-10 and CLAD-C; FGH(Michel et al., 2026) was the only dataset-level reversal, exceeding AFM on CORe50 by 2.8% (Table 15). On the equal-weighted three-dataset primary summary, AFM exceeded CCL-DC(Wang et al., 2024), MKD(Michel et al., 2024), LPR(Yoo et al., 2024), FGH, aL-SAR(Seo et al., 2025) and OCAR(Urettini & Carta, 2025) by 7.0%, 11.4%, 16.3%, 21.7%, 31.7% and 36.8%, respectively. These are not equal-compute or equal-storage claims, and the benchmark representation prefix was frozen after initial learning. They establish that the certified retention-plasticity frontier is executable and empirically competitive, while the causal experiments address the broader scientific question of why persistent learning is possible in some directions and not others.

4.3 Nonlinear geometry sets the boundary

The strongest challenge to a local compatibility law is finite update scale. The largest controlled multiscale update was about 12,170-fold smaller than the median natural update for the convolutional network and 1,118-fold smaller for the vision transformer, whereas the gap was 13.5-fold in text (Fig. 4a). We therefore

Table 15: AFM chronological retention-plasticity benchmarks

Dataset	AFM	CCL-DC	MKD	FGH	aL-SAR	LPR	OCAR
CORe50	0.2478	0.2016	0.1799	**0.2546**	0.1801	0.1640	0.1436
CLEAR-10	**0.3807**	0.3731	0.3513	0.3773	0.3572	0.3485	0.3164
CLAD-C	**0.5129**	0.4921	0.4931	0.3058	0.3296	0.4694	0.3747

Dataset	AFM	Strongest classical non-oracle	Mean difference	95% paired CI
CORe50	0.2478	No protection 0.2388	+0.0090	[0.0059,0.0126]
CLEAR-10	0.3807	No protection 0.3740	+0.0068	[0.0053,0.0082]
CLAD-C	0.5129	A-GEM 0.5019	+0.0110	[0.0054,0.0166]

Primary score is the protocol-specific retention-plasticity hypervolume. Modern challengers use frozen method-native configurations; this comparison controls predictive base and learner-visible information but is not an equal-compute/storage study.

Table 16: Near-zero theorem-aligned boundary assessment

System	Negative requested-zero empirical margins
CNN	52
ViT	1
Text transformer	12
Total	65

Every row lacked certification of the relevant finite curvature or step condition. Numerical verification agreed with the reported κ values to within 1.5883×10^{-9} and with the persistent-ratio values exactly.

constructed a fixed-norm bridge in the two vision systems using ten seeds and 500 preselected states per architecture. At each state we attempted four compatibility levels at exactly 1%, 10%, 50% or 100% of the median natural unrestricted update norm. The admissibility criterion required all four same-norm directions to produce positive unrestricted finite progress; Δ_0 was measured, not matched.

At 1% natural norm, 231 of 500 vision-transformer states across all ten seeds supported the four-level comparison, whereas only one convolutional-network state did. At 10% and above, neither system supported a jointly feasible four-level continuum (Fig. 4b). This is not a zero or negative compatibility effect: the matched intervention itself ceases to exist because finite nonlinear geometry makes one or more same-norm directions non-descending.

Where the bridge remained feasible, compatibility still increased the *absolute* amount of persistent learning. In the 231 vision-transformer states, the slope of $\Delta_{\text{persistent}}$ on κ was 3.064×10^{-4} with a 95% interval of $[3.031, 3.093] \times 10^{-4}$ (Fig. 4c). The normalized ratio, however, had slope -0.572 with an interval crossing zero because Δ_0 itself changed with compatibility. Stratifying states by Δ_0 heterogeneity made the distinction explicit: the absolute persistent slope was essentially unchanged across low, medium and high heterogeneity, whereas the ratio slope shifted from $+2.25$ and $+3.15$ to -7.54 (Fig. 4d). The original matched-Δ_0 experiment therefore identifies the fraction of an equal learning opportunity that can persist; at fixed finite norm, absolute persistent progress is the cleaner causal outcome.

The following sections report the prespecified robustness tests, the near-zero boundary assessment, the fixed-norm bridge, and the decomposition that distinguishes absolute persistent progress from denominator-sensitive normalized ratios.

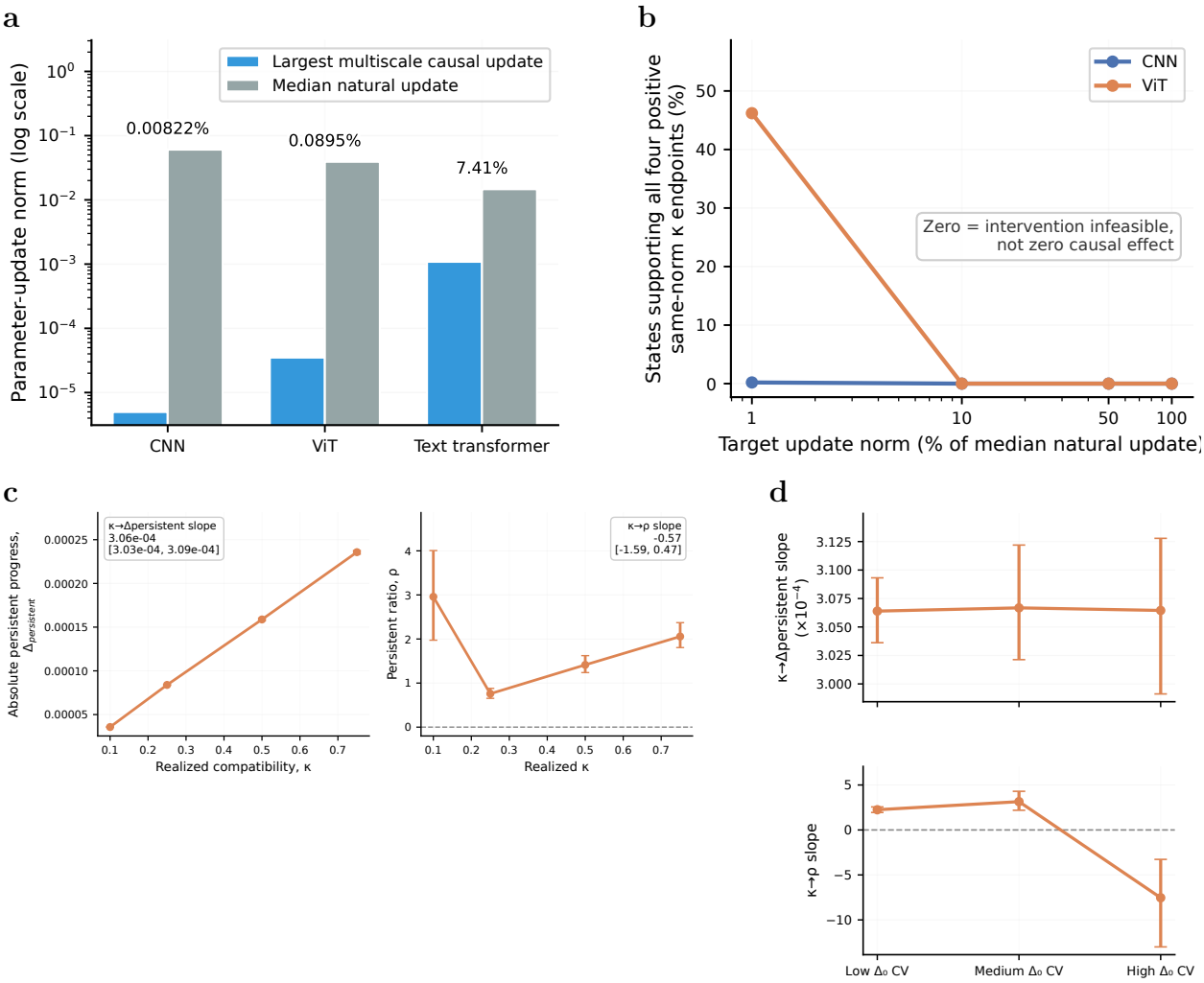

Figure 4: **Nonlinear geometry sets the finite-scale boundary of compatibility-controlled learning.** **a**, Largest designed causal update norms compared with median natural unrestricted update norms; natural calibration uses 250 ordinary states per original system. **b**, Fraction of 500 preselected states per architecture, 10 seeds by 50 states, supporting all four positive same-norm compatibility endpoints at 1% to 100% of median natural update norm. Zero feasibility denotes absence of the matched four-level intervention, not a zero causal effect. **c**, In 231 feasible vision-transformer states at 1% natural norm, distributed across all 10 seeds, compatibility increases absolute persistent progress whereas the normalized ratio does not show a positive slope. **d**, Stratification of the same 231 states by unrestricted-progress heterogeneity leaves the absolute persistent slope essentially invariant while the ratio slope reverses sign. Bridge intervals use $n = 10$ seed-level effects and 100,000 deterministic seed-bootstrap resamples.

Table 17: Fixed-norm bridge feasibility and decomposition

System	Natural-norm fraction	Feasible / attempted	Seeds with feasibility	Target norm
CNN	0.01	1/500 (0.2%)	1	0.00060570
CNN	0.10-1.00	0/1500	0	0.00605702-0.06057018
ViT	0.01	231/500 (46.2%)	10	0.00039144
ViT	0.10-1.00	0/1500	0	0.00391441-0.03914410

Method, ViT 1%, $\beta = 0.5$	Absolute persistent-progress slope [95% CI]	Normalized-ratio slope [95% CI]
EWC-prox	4.3693×10^{-5} [3.5585,5.1115]$\times 10^{-5}$	-7.895 [-10.701,-5.197]
Linearized distillation	2.59767×10^{-4} [2.54328,2.65080]$\times 10^{-4}$	-3.494 [-5.588,-1.502]
Projection / AFM base	3.06428×10^{-4} [3.03053,3.09322]$\times 10^{-4}$	-0.572 [-1.591,0.455]
Unrestricted	4.38287×10^{-5} [3.55718,5.12805]$\times 10^{-5}$	-7.891 [-10.717,-5.190]

Bridge admission requires positive unrestricted finite progress for all four same-norm compatibility directions. Across 4,000 attempted state-scale conditions, 232 were feasible. Maximum absolute compatibility error was 9.656×10^{-6} and maximum relative update-norm error was 3.674×10^{-5}.

4.4 Purpose and scope

These follow-up experiments test direction specificity, intervention scale, seed and architecture generality, the near-zero diagnostic boundary, and finite-norm extension toward ordinary update magnitudes. Positive, null and infeasible outcomes are reported because each defines part of the empirical scope of the compatibility claim.

The experimental claim under investigation is deliberately narrower than a universal statement that compatibility alone determines learning. The causal proposition is:

> At a fixed neural state and under a fixed retention criterion, changing the functionally compatible component of the incoming learning signal changes how much new learning can be stored persistently. The magnitude of that effect depends on the retention budget and learning rule, and finite nonlinear geometry can limit how far the local compatibility intervention can be extended.

The experiments below distinguish three questions:

1. Is the observed effect specific to one constructed direction or does it survive independently generated directions with the same compatibility?

2. Does the local effect persist as the intervention is enlarged, across more seeds and more architectures?

3. What happens when the intervention is brought toward ordinary update magnitude, where finite nonlinear loss geometry is no longer negligible?

4.5 Notation and common estimands

Let q denote the incoming functional learning signal and let P denote the projection onto the locally protected-compatible subspace. The realized compatibility used throughout the causal interventions is

$$\kappa = \frac{\|Pq\|^2}{\|q\|^2}. \tag{116}$$

The unrestricted finite learning opportunity from the same pre-update state is denoted

$$\Delta_0 > 0, \tag{117}$$

and the amount of current learning that remains stored in ordinary parameters after enforcing the retention criterion is denoted $\Delta_{\text{persistent}}$. When the unrestricted denominator is appropriate for comparison, the normalized persistent-progress ratio is

$$\rho = \frac{\Delta_{\text{persistent}}}{\Delta_0}. \tag{118}$$

The retention budget is indexed by β. In the original causal suites, the grid was

$$\beta \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 1\}. \tag{119}$$

Unless stated otherwise, causal slopes are estimated within a fixed neural state across the compatibility intervention levels, state-level slopes are averaged within seed, and inference is performed at the seed level rather than treating states as independent replicates.

A theorem-aligned diagnostic used in the native AFM analyses is

$$M = \rho - \frac{\widehat{\lambda}\kappa}{3}. \tag{120}$$

This quantity is an empirical alignment diagnostic. It is not interpreted as a newly certified finite nonlinear theorem bound unless the required finite curvature and step conditions have themselves been certified.

4.6 Reference: original controlled causal experiment

The follow-up studies were motivated by the original controlled causal experiment. That experiment used a fully unfrozen CIFAR-10 CNN, a fully unfrozen CIFAR-10 ViT, and a fully unfrozen character-level text transformer. There were five seed trajectories (11, 29, 47, 71, 101), 50 causal states per seed, and six requested compatibility levels

$$\kappa^\star \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}. \tag{121}$$

The seven method branches were AFM/projection, unrestricted, replay, linearized distillation, EWC-prox, DER++, and the native AFM mechanism. The controlled analysis contained 31,500 wide method outcomes, while the nonlinear frontier analysis contained 220,500 rows; validation reported complete expected coverage.

In this experiment, the unrestricted finite decrease Δ_0 was deliberately matched across the compatibility intervention levels within a state. Therefore ρ directly compared the fraction of approximately equal unrestricted learning opportunities that survived the retention requirement. At $\beta = 0.5$, the matched $\kappa \mapsto \rho$ slope for the projection/AFM-base branch was near one in all three systems (Table 18).

These reference results established the phenomenon in a deliberately local, Δ_0-matched regime. The four experiments below were run specifically to challenge alternative explanations and test generality.

4.7 Experiment 1: multiscale causal compatibility

4.7.1 Question

The first stress test asked whether the matched compatibility effect survives when the intervention is expanded beyond the original very local operating scale. The design retained the same three systems, five seeds, seven

Table 18: Reference causal slopes at retention budget $\beta = 0.5$.

System	Matched slope	95% CI
CIFAR-10 CNN	1.01377	[1.01171, 1.01611]
CIFAR-10 ViT	1.00082	[1.00029, 1.00135]
Text transformer	0.96497	[0.95283, 0.97710]
Equal-system pooled	0.99319	[0.98831, 0.99781]

methods, six requested compatibility levels, and seven retention budgets. Four intervention scale indices were used:

$$s \in \{0.05, 0.2, 0.5, 0.9\}. \tag{122}$$

The multiscale design defines these values as a *logarithmic expansion from the original local intervention to a common peak*. They are therefore intervention-scale coordinates and must not be interpreted as 5%, 20%, 50%, or 90% of a natural training update.

Validation reported 882,000 expected and observed frontier rows, 15/15 completed runs, and no validation failures.

4.7.2 Key result

For AFM/projection at a moderate retention budget ($\beta = 0.5$), the matched compatibility slope remained close to one throughout the designed multiscale range in both vision systems (Table 19).

Table 19: AFM/projection matched κ slopes across the designed multiscale range at $\beta = 0.5$.

System	$s = 0.05$	$s = 0.2$	$s = 0.5$	$s = 0.9$
CIFAR-10 CNN	1.01290	1.01032	1.00359	0.99390
CIFAR-10 ViT	1.00135	0.99886	0.99567	0.98347

The CNN also illustrates the interaction between compatibility and the retention budget. At the tight budget $\beta = 0.01$, the AFM/projection matched slope rose from 0.10272 at $s = 0.05$ to 0.16003 at $s = 0.2$, 0.48762 at $s = 0.5$, and 0.96935 at $s = 0.9$. At $\beta = 0.05$, the corresponding CNN slopes were 1.01171, 1.01032, 1.00359, and 0.99390. Thus compatibility is not the sole determinant of attainable persistent progress: the path fraction permitted by the retention budget also matters.

4.7.3 Scale calibration against natural updates

A calibration compares the multiscale intervention norms with ordinary unrestricted update norms measured in the natural-state validation. Update norm is the meaningful cross-experiment scale quantity because the causal and natural experiments use different current objectives; their raw Δ_0 values are not directly commensurate.

At the largest multiscale index $s = 0.9$, the causal updates were still much smaller than ordinary natural updates, especially in vision (Table 20).

Table 20: Largest designed multiscale intervention compared with the median natural unrestricted update norm.

System	Mean causal norm	Median natural norm	Causal/natural ratio
CIFAR-10 CNN	4.9766×10^{-6}	0.0605702	8.2163×10^{-5}
CIFAR-10 ViT	3.5026×10^{-5}	0.0391441	8.9479×10^{-4}
Text transformer	1.0863×10^{-3}	0.0146534	0.0741306

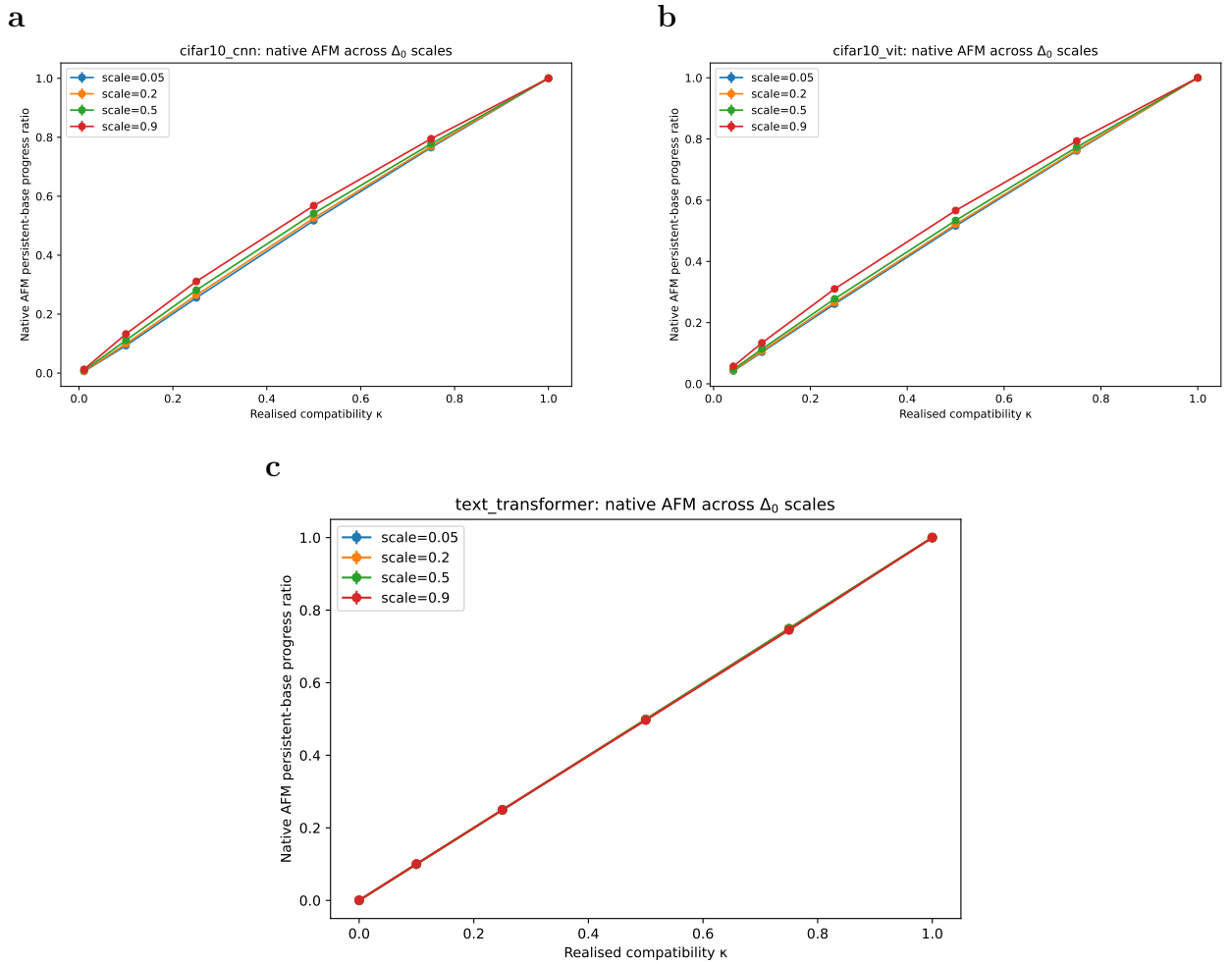

Figure 5: **Native AFM multiscale dose responses. a–c**, Native-AFM responses across the designed local multiscale suite for the **a**, CIFAR-10 convolutional network, **b**, CIFAR-10 vision transformer, and **c**, text transformer. Internal scale indices are logarithmic expansion coordinates and must not be interpreted as percentages of a natural update.

These correspond approximately to gaps of 12,170-fold for the CNN, 1,118-fold for the ViT, and 13.5-fold for the text transformer. The resulting scale gap motivates the fixed-norm natural-scale bridge in Section 4.11.

4.7.4 Interpretation

The multiscale experiment supports robustness over a deliberately enlarged *local* causal regime. It does not establish natural-update-scale invariance. Its principal value is that the compatibility effect is not confined to one single infinitesimal numerical setting, while the calibration makes clear where the local experiment still sits relative to ordinary training updates.

4.8 Experiment 2: independently constructed directions at fixed compatibility

4.8.1 Question and design

The second stress test addressed a construction-specific confound: perhaps the original dose response was a property of the particular generalized-eigenmode direction used to realize each κ, rather than of compatibility itself. Four independently constructed directions were therefore generated for each nonzero requested

compatibility level

$$\kappa^\star \in \{0.1, 0.25, 0.5, 0.75\}. \tag{123}$$

The experiment used the CNN, ViT, and text transformer; five seeds; all seven retention budgets; and seven method branches. Validation reported four directions per κ, 147,000 fixed groups, and 588,000 expected and observed frontier rows, with all 15 runs complete.

The principal diagnostic compares variability across independently generated directions at a fixed κ with variability across κ. Let

$$R_{\mathrm{dir}/\kappa} = \frac{\text{within-}\kappa\text{ SD across directions}}{\text{between-}\kappa\text{ SD}}. \tag{124}$$

Small values indicate that changing the direction while holding compatibility fixed produces much less variation than changing compatibility itself.

4.8.2 Results

Table 21 records the AFM/projection results across retention budgets. At nontrivial budgets, the matched compatibility slope approaches one and the within-κ directional variation is generally far smaller than the between-κ variation. The text transformer is the most budget-limited system at tight budgets, but it approaches the same near-unit regime as the retention budget is relaxed.

Table 21: Independent-direction robustness for AFM/projection. Each cell gives $R_{\mathrm{dir}/\kappa}$ / matched κ slope.

β	CIFAR-10 CNN	CIFAR-10 ViT	Text transformer
0.01	0.3718 / 0.0725	0.2467 / 0.2740	0.4353 / −0.0066
0.05	0.0922 / 0.6876	0.0409 / 0.8535	0.5321 / 0.0927
0.10	0.0577 / 0.8917	0.0220 / 0.9566	0.3445 / 0.3367
0.25	0.0400 / 1.0130	0.0254 / 1.0133	0.1151 / 0.8285
0.50	0.0379 / 1.0465	0.0206 / 1.0294	0.0302 / 0.9861
1.00	0.0347 / 1.0556	0.0205 / 1.0295	0.0178 / 1.0033

At $\beta = 0.5$, the 95% seed-level confidence intervals for the matched slope were [1.03046, 1.06024] in the CNN, [1.02516, 1.03442] in the ViT, and [0.97849, 0.99454] in the text transformer. The mean fraction of states whose within-κ direction SD was below the between-κ variation was 1.0 for all three systems at this budget.

4.8.3 Interpretation and reporting limitation

This experiment substantially weakens the explanation that the effect is tied to a single specially chosen generalized-eigenmode recipe. It does not imply that all directions sharing the same κ are equivalent; rather, at the tested local scale and ordinary retention budgets, direction-to-direction variation is much smaller than the systematic variation induced by changing compatibility.

The independent-direction analysis does not assume exact equality of realized κ, update norm or Δ_0 across the four constructed directions. The reported inference is based on the observed within-compatibility directional variation relative to the systematic between-compatibility effect.

4.9 Experiment 3: 10-seed generality and a stronger ViT

4.9.1 Question and design

The original controlled causal experiment used five seeds. The generality experiment doubled the number of seed trajectories to ten and added a stronger CIFAR-10 ViT, yielding four systems:

1. CIFAR-10 CNN,
2. original CIFAR-10 ViT,

3. stronger CIFAR-10 ViT,

4. text transformer.

Uncertainty was summarized with a deterministic Monte Carlo seed bootstrap using 100,000 resamples.

4.9.2 System-level results

At $\beta = 0.5$, all four AFM/projection matched slopes were strongly positive and near one (Table 22).

Table 22: Ten-seed AFM/projection matched slopes at $\beta = 0.5$.

System	Mean matched slope	95% CI
CIFAR-10 CNN	1.01544	[1.01324, 1.01778]
CIFAR-10 ViT	1.00064	[1.00010, 1.00114]
Stronger CIFAR-10 ViT	0.99883	[0.99824, 0.99942]
Text transformer	0.95950	[0.95049, 0.96863]

4.9.3 Equal-system pooled dose response

The four-system equal-weight pooled AFM slope increased with the retention budget (Table 23). This pattern is consistent with a compatibility signal whose exploitable persistent fraction is itself retention-budget dependent.

Table 23: Equal-system pooled AFM/projection slope across ten common seeds.

β	Pooled slope	95% CI
0.01	0.36331	[0.35075, 0.37667]
0.05	0.77171	[0.76572, 0.77759]
0.10	0.82114	[0.81433, 0.82756]
0.25	0.93121	[0.92425, 0.93841]
0.50	0.99360	[0.99123, 0.99600]
1.00	1.00379	[1.00326, 1.00442]

4.9.4 Interpretation

The generality experiment addresses the small-seed concern and shows that the dose response is not specific to a single CNN, a single ViT implementation, or a single modality. The stronger ViT result is especially useful because it tests a distinct vision-transformer configuration without changing the scientific intervention. It is not a foundation-model-scale demonstration and should not be described as one.

4.10 Experiment 4: near-zero-compatibility boundary assessment

4.10.1 Boundary cases

Among accepted native-AFM rows near requested $\kappa = 0$, 65 have a negative empirical margin relative to the coarse theorem-aligned reference $\widehat{\lambda}\kappa/3$. The minimum recorded margin is approximately -0.003857. Because the empirical reference is not itself a certified bound unless the relevant finite assumptions hold, these cases were assessed explicitly for certification status.

4.10.2 Assessment result

All 65 rows were classified as

lacking certification of the relevant finite curvature or step condition.

The system breakdown is given in Table 24.

Table 24: Classification of all negative theorem-aligned-margin cases near requested $\kappa = 0$.

System	Count
CIFAR-10 CNN	52
CIFAR-10 ViT	1
Text transformer	12
Total	65

Numerical verification agreed with the reported κ values to within 1.5883×10^{-9} and with the reported ρ values exactly across all 65 cases.

4.10.3 Interpretation

This assessment does *not* prove the nonlinear theorem assumptions for these rows. Its conclusion is narrower: the negative empirical margins occurred outside the set for which the finite curvature/step assumptions had been certified. Therefore these rows cannot legitimately be counted as certified theorem violations. The quantity $\widehat{\lambda}\kappa/3$ is accordingly described as a theorem-aligned empirical reference unless the relevant assumptions are checked for that row.

4.11 Natural-update-scale bridge

4.11.1 Motivation

The calibration in Section 4.7 showed that the largest designed causal interventions remained far below ordinary unrestricted update norm in the two vision systems. The fixed-norm bridge therefore targets update norms defined directly as fractions of the median natural unrestricted update norm. Only the CIFAR-10 CNN and existing CIFAR-10 ViT are used, because these are the systems with the largest scale gap. Ten seed trajectories and 50 preselected states per seed are used; the states are fixed independently of the resulting feasibility outcomes.

The requested compatibility levels were

$$\kappa^\star \in \{0.1, 0.25, 0.5, 0.75\}, \tag{125}$$

and the target update norms were

$$R^\star \in \{0.01, 0.1, 0.5, 1.0\} \times R_{\text{natural,median}}. \tag{126}$$

The primary intervention therefore controls two quantities:

$$\kappa \approx \kappa^\star, \qquad \|d\| = R^\star. \tag{127}$$

For a state-scale condition to support a four-level within-state comparison, all four same-norm compatibility interventions must yield positive unrestricted finite progress,

$$\Delta_0(\kappa) > 0 \qquad \text{for all four tested } \kappa. \tag{128}$$

4.11.2 Admission criterion

Because the bridge fixes both update direction and update norm, finite unrestricted progress Δ_0 is an outcome of the nonlinear loss surface. A state-scale condition supports the four-level within-state comparison when all required κ interventions have positive unrestricted finite progress at the prescribed common norm.

Within-state variation in Δ_0 is reported descriptively because it can affect normalized ratios, but it is not an admission variable.

Across 4,000 attempted state-scale conditions, 232 satisfied the positive-endpoint criterion. Validation reported:

- 232 feasible state-scale conditions among 4,000 attempted conditions;
- maximum absolute compatibility error 9.656×10^{-6};
- maximum relative update-norm error 3.674×10^{-5};
- 14,848 expected and observed merged frontier rows;
- 928 expected and observed native-AFM rows;
- no validation failures.

4.11.3 Feasibility

The feasibility result is shown in Table 25. The ViT supports a substantial four-level comparison at 1% of natural update norm: 231/500 states across all ten seeds. The CNN supports only one such state. At 10%, 50%, and 100% of natural update norm, neither architecture has a state for which all four target-compatibility directions remain positive same-norm finite endpoints.

Table 25: Fixed-norm bridge feasibility.

System	Natural-norm fraction	Target norm	Feasible/attempted	Seeds with any feasible
CNN	0.01	0.00060570	1/500 (0.2%)	1
CNN	0.10	0.00605702	0/500	0
CNN	0.50	0.03028509	0/500	0
CNN	1.00	0.06057018	0/500	0
ViT	0.01	0.00039144	231/500 (46.2%)	10
ViT	0.10	0.00391441	0/500	0
ViT	0.50	0.01957205	0/500	0
ViT	1.00	0.03914410	0/500	0

Among the 231 feasible ViT states at 1%, the mean within-state Δ_0 coefficient of variation is 0.22586 and the maximum is 0.58485. This is direct evidence that finite unrestricted progress varies appreciably across same-norm compatibility directions once the intervention reaches this regime.

4.11.4 Why the raw ratio appears to fail

In the 1% ViT bridge, the raw fixed-norm $\kappa \mapsto \rho$ slope for projection/native AFM is

$$-0.57177, \qquad \mathrm{CI}_{95\%} = [-1.59, 0.46], \tag{129}$$

so the ratio does not show a positive dose response. If the analysis stopped at ρ, the bridge would appear to undermine a naive universal claim that compatibility must always increase the normalized ratio.

However, this fixed-norm intervention does *not* hold Δ_0 constant. Because

$$\rho = \frac{\Delta_{\mathrm{persistent}}}{\Delta_0}, \tag{130}$$

its derivative with respect to compatibility is

$$\frac{d\rho}{d\kappa} = \frac{\Delta_0 \frac{d\Delta_{\mathrm{persistent}}}{d\kappa} - \Delta_{\mathrm{persistent}} \frac{d\Delta_0}{d\kappa}}{\Delta_0^2}. \tag{131}$$

A flat or negative ratio slope can therefore coexist with strongly increasing absolute persistent learning if the unrestricted denominator also increases with compatibility.

4.12 Decomposition of the fixed-norm bridge

4.12.1 Primary decomposition

The bridge was decomposed into three within-state effects across the four compatibility interventions:

1. $\kappa \mapsto \Delta_0$,
2. $\kappa \mapsto \Delta_{\text{persistent}}$,
3. $\kappa \mapsto \rho$.

The state-level slopes were averaged within seed and summarized across the ten ViT seeds.

For native AFM/projection at 1% natural update norm, the decomposition is:

$$\frac{d\Delta_0}{d\kappa} = 7.78981 \times 10^{-5}, \tag{132}$$

$$\frac{d\Delta_{\text{persistent}}}{d\kappa} = 3.06428 \times 10^{-4}, \qquad \text{CI}_{95\%} = [3.03053, 3.09315] \times 10^{-4}, \tag{133}$$

$$\frac{d\rho}{d\kappa} = -0.57177, \qquad \text{CI}_{95\%} = [-1.59145, 0.45542]. \tag{134}$$

The mean persistent-update acceptance fraction is 1.0 and the mean finite-completion fraction is 0.994118.

Table 26 shows that the positive absolute-persistent effect is not restricted to one method, although its magnitude is method dependent.

Table 26: ViT at 1% natural update norm, $\beta = 0.5$: absolute persistent-progress and normalized-ratio slopes.

Method	$\kappa \mapsto \Delta_{\text{persistent}}$ slope [95% CI]	$\kappa \mapsto \rho$ slope [95% CI]
EWC-prox	4.3693×10^{-5} [3.5585, 5.1115]$\times 10^{-5}$	−7.895 [−10.701, −5.197]
Linearized distillation	2.59767×10^{-4} [2.54328, 2.65080]$\times 10^{-4}$	−3.494 [−5.588, −1.502]
Projection / AFM base	3.06428×10^{-4} [3.03053, 3.09322]$\times 10^{-4}$	−0.572 [−1.591, 0.455]
Unrestricted branch	4.38287×10^{-5} [3.55718, 5.12805]$\times 10^{-5}$	−7.891 [−10.717, −5.190]

A separate regression that adjusts for realized finite Δ_0 yields a positive compatibility coefficient for the projection branch (about 2.694 with a 95% interval of roughly [2.36, 3.08]). This is useful as a decomposition/sensitivity analysis but should not be treated as the primary causal estimand because Δ_0 is a post-intervention quantity in the fixed-norm experiment.

4.12.2 Δ_0-heterogeneity stratification

The 231 feasible ViT states were divided into low-, medium-, and high-Δ_0-CV strata using the first and second tertile boundaries, 0.07620 and 0.33483. The key projection result is striking: the absolute persistent-progress slope is essentially unchanged across all three strata, while the normalized ratio slope changes dramatically and reverses sign in the high-heterogeneity stratum (Table 27).

For example, at $\beta = 0.5$, the 95% intervals for the absolute persistent-progress slope are approximately [3.03618, 3.09315]$\times 10^{-4}$ in the low band, [3.02131, 3.12192]$\times 10^{-4}$ in the medium band, and [2.99120, 3.12784]$\times 10^{-4}$ in the high band. By contrast, the ratio-slope intervals are [1.945, 2.561], [2.192, 4.298], and [−13.001, −3.271], respectively.

4.12.3 Interpretation

The stratification provides a direct explanation for the apparent ratio failure. The absolute amount of persistently stored learning responds to compatibility with nearly the same slope regardless of how heterogeneous

Table 27: Projection branch at ViT 1% natural update norm, stratified by Δ_0 heterogeneity. Values are identical across the tested projection retention budgets because the projection endpoint saturates to the same branch in this setting.

Δ_0-CV band	States	$\kappa \mapsto \Delta_{\text{persistent}}$ slope	$\kappa \mapsto \rho$ slope
Low	78	3.06394×10^{-4}	2.2463
Medium	77	3.06677×10^{-4}	3.1477
High	76	3.06452×10^{-4}	−7.5366

the unrestricted finite denominator is. The normalized ratio, however, is highly sensitive to denominator heterogeneity and can reverse sign even when absolute persistent progress continues to increase strongly.

This does not make ρ an invalid quantity in general. In the original Δ_0-matched experiment it answers a clean question: what fraction of an approximately equal unrestricted learning opportunity survives the retention constraint? In the fixed-norm bridge, by contrast, Δ_0 is itself an outcome of changing direction, so ρ is no longer a stable standalone causal summary.

4.13 Finite-step feasibility boundary

The natural-scale bridge also exposes a second phenomenon that is distinct from the ratio decomposition. Increasing the fixed update norm eventually destroys the ability to construct a common four-level continuum of positive-descent directions at the target compatibilities. This happens before 10% of the median natural update norm in both tested vision systems under the present direction construction and current objective.

The correct interpretation is therefore conditional:

1. **Where the four-level intervention is jointly feasible**, compatibility can be causally varied at fixed state and fixed update norm, and the ViT 1% data show a strong positive effect on absolute persistent progress.

2. **Where joint feasibility disappears**, there is no matched four-level causal slope to estimate. Those cases are not negative slopes; they are failures of the local compatibility continuum to remain a positive finite descent construction at that step magnitude.

This boundary is itself informative. It separates the locally controlled functional geometry from a larger-step regime in which finite curvature and nonlinear loss-surface structure determine whether the nominally compatible direction remains a useful descent endpoint.

4.14 Integrated interpretation of the follow-up experiments

Taken together, the follow-up studies support the following evidence chain.

1. **The effect is not tied to one intervention direction.** Four independently generated directions at each nonzero compatibility level produce much less within-κ variation than the systematic between-κ effect at ordinary retention budgets.

2. **The effect is robust over a designed local multiscale range.** In the CNN and ViT, AFM/projection slopes remain near one as the intervention is enlarged within the multiscale construction.

3. **The effect generalizes across more independent seed trajectories and architectures.** Ten-seed results in the CNN, original ViT, stronger ViT, and text transformer reproduce a positive dose response, with a pooled slope of 0.9936 at $\beta = 0.5$.

4. **Near-zero theorem-aligned margins are separated from certified theorem violations.** All 65 negative theorem-aligned margins occur where the relevant finite curvature/step condition is not certified; numerical verification agrees with the reported quantities to the stated precision.

5. **At a fixed finite update norm, absolute persistent progress remains compatibility-sensitive even when the normalized ratio does not.** In 231 feasible ViT states at 1% natural update norm, the absolute AFM/projection slope is 3.06428×10^{-4} with a narrow positive confidence interval, whereas the raw ratio slope crosses zero.

6. **Denominator heterogeneity explains the ratio instability.** Across low-, medium-, and high-Δ_0 heterogeneity strata, the absolute persistent slope is nearly invariant while the ratio slope changes from positive to strongly negative.

7. **Finite nonlinear geometry imposes a scale boundary.** Above the 1% target, the present four-level same-norm positive-endpoint construction becomes infeasible in the CNN and ViT; no causal slope should be claimed there.

The combined evidence supports the following synthesis:

> Functional compatibility is a causal determinant of retention-constrained persistent learning in the locally feasible regime. The amount of compatibility that can be exploited depends on the retention budget and learning rule. At finite fixed norm, compatibility continues to increase absolute persistent learning where the intervention remains feasible, but unrestricted finite progress also becomes compatibility-dependent, making the normalized ratio an unstable summary. At still larger step magnitudes, nonlinear geometry can eliminate the common positive-descent compatibility continuum itself.

This synthesis is intentionally different from two stronger statements that the data do not establish: (i) that ρ must increase with κ at every finite scale, or (ii) that a four-level compatibility intervention remains realizable at ordinary full natural update magnitude.

4.15 Additional statistical considerations

1. **Seed as inferential unit.** The original, multiscale, and independent-direction analyses aggregate state-level effects within seed and use seed-level uncertainty. The generality and bridge analyses use ten seeds with 100,000 deterministic Monte Carlo bootstrap resamples where reported.

2. **No state selection on large-scale success.** Bridge states are fixed before attempting large-scale interventions. Feasibility rates are reported over all attempted states rather than replacing failed states until a balanced successful sample is obtained.

3. **Positive-endpoint conditioning.** A bridge slope is conditional on joint feasibility: all tested target-κ directions must produce positive same-norm unrestricted progress. The unconditional feasibility rate is therefore a separate result and must accompany conditional slopes.

4. **Δ_0 is post-intervention in the fixed-norm bridge.** Analyses that adjust for realized Δ_0 are decompositional/sensitivity analyses, not the primary causal estimand.

5. **Fixed-norm inclusion criterion.** Δ_0 similarity is not an admission variable because finite unrestricted progress is an outcome at fixed direction and norm. Feasibility is defined by positive unrestricted finite progress for all tested same-norm compatibility directions and is reported over all 4,000 attempted state-scale conditions.

5 Experimental methods and reproducibility

5.1 Study design and inferential unit

The study contains three linked experimental layers. The first is the original AFM chronological evaluation on CORe50, CLEAR-10 and CLAD-C, which establishes the framework as an executable retention-plasticity mechanism. The second is a controlled causal experiment in fully unfrozen neural networks that intervenes directly on functional compatibility. The third consists of follow-up stress tests addressing direction specificity, seed and architecture generality, local intervention scale, the near-zero diagnostic boundary and fixed-norm extension toward ordinary update magnitudes. Causal rows are repeated measurements within neural states and seed trajectories; they are never treated as independent observations merely because the row count is large. Unless otherwise stated, a slope is estimated within state across compatibility levels, state slopes are averaged within seed, and uncertainty is computed across seed-level effects.

5.2 Causal systems and reconstruction

The controlled experiment used a CIFAR-10 convolutional network, a CIFAR-10 vision transformer and a character-level text transformer trained on WikiText-2. All trainable representations remained unfrozen. The original causal suite used seeds 11, 29, 47, 71 and 101, with 50 causal states per seed and system, giving 750 states. A saved pre-probe parent checkpoint contained the complete model, optimizer, stream position, replay reservoir and random-number-generator state. Every method and compatibility intervention at one causal state was reconstructed from that identical parent checkpoint. The generality experiment used ten seed trajectories and added a stronger CIFAR-10 vision transformer.

5.3 Functional compatibility intervention

At a fixed state, the current-learning signal was modified in function space rather than by inserting an arbitrary parameter-space vector. Residual modes were constructed from the generalized eigenproblem

$$J_c P J_c^{\mathsf{T}} r = \lambda J_c J_c^{\mathsf{T}} r, \tag{135}$$

and low- and high-compatibility modes were mixed to target six requested levels from 0 to 1 (0, 0.1, 0.25, 0.5, 0.75 and 1). Realized compatibility was recomputed as $\kappa = \|Pq\|^2/\|q\|^2$ and used in every analysis. The requested-zero condition is a minimum-compatibility boundary condition rather than exact zero in the two vision systems: mean realized values were 0.010681 for the convolutional network, 0.041266 for the vision transformer and 8.3×10^{-5} for text. For nonzero requests, mean absolute compatibility error was at most 1.1×10^{-5} across systems.

5.4 Matching unrestricted learning

Changing direction can change the finite improvement available without protection. In the original causal experiment, the weakest attainable unrestricted endpoint at each state defined a common positive target decrease. A one-dimensional bisection along each original direction matched this target without rotating the direction. The resulting denominator was

$$\Delta_0 = \mathcal{L}_{\text{current}}(\theta_{\text{pre}}) - \mathcal{L}_{\text{current}}(\theta_{\text{unrestricted}}). \tag{136}$$

The mean within-state relative spread of Δ_0 was 0.283%, 0.282% and 0.268% in the convolutional network, vision transformer and text transformer, with maxima below 0.40%. This matched-Δ_0 design makes $\rho = \Delta_{\text{persistent}}/\Delta_0$ interpretable as the fraction of an approximately equal unrestricted learning opportunity that remains persistent.

5.5 Retention budgets and methods

At each causal state, the reference protected drift was the maximum unrestricted drift across compatibility interventions. Every method and compatibility level was evaluated under the same absolute budget $D \leq$

$\max(10^{-8}, \beta D_{\mathrm{ref}})$ for $\beta \in \{0, 0.01, 0.05, 0.1, 0.25, 0.5, 1\}$. Because nonlinear protected drift need not be monotone in step scale, generic branches evaluated 33 candidate endpoint scales from zero to the full proposal and selected the best feasible persistent endpoint. The seven evaluated branches were unrestricted learning, replay, projection, linearized distillation, EWC-proximal updating, DER++ and an AFM-compatible projected proposal. The generic AFM proposal is identical to the projection family in this controlled suite and is collapsed with it for cross-system slope summaries. Native AFM was evaluated separately with its accepted backtracking fraction, finite functional completion and endpoint checks.

5.6 AFM framework and protected transaction

AFM operates on a nonanticipating supervised stream without learner-visible task, session, episode, segment, intervention or boundary identifiers. It separates persistent model parameters from bounded structural state. A protected transaction begins with a genuine same-state no-protection comparator: the declared unrestricted learner is run from the identical predictive and optimizer state, with the same current minibatch, active coordinates, learning-rate and backtracking rules, mutable buffers and declared private randomness, but without retention constraints. The accepted endpoint is used only as a counterfactual reference and is never installed as the persistent base model.

Protected behaviour is represented functionally. Committed records retain bounded whole-behaviour sensitivity sketches, formed by streaming Jacobian rows through Frequent Directions,(Ghashami et al., 2016) rather than a permanent scalar importance value for every parameter. A finite family of timescale and spectral-rank policies predicts which protected sensitivities remain relevant. The controller is charged both for residual protected leakage and for current gradient energy blocked by protection. Task-free routing uses only declared observable signatures; outcome evidence may trigger reopening, but creation of a new observable route requires separately controlled signature evidence. When observables do not distinguish semantic states, AFM reports the resulting routing obstruction rather than using evaluator identities.

For a protected update, AFM forms the compatible projection $v_t = \Pi_t d_t^0$ of the same-state comparator. A predeclared assimilation coordinate η_t allocates a fraction of the certified retention charge along this reference path. The safe-base operator accepts only a persistent endpoint satisfying the retention, descent, trust-region, collinearity and normalized-assimilation checks. Under the assumptions stated and proved in the methods below, $\lambda_t \geq \eta_t$ and the scalar-gradient comparator yields Eq. 2. The protected projector also has an exact first-order rank-plasticity characterization: for protected Jacobian J with singular values $\sigma_1 \geq \cdots \geq \sigma_d$, the smallest worst-case first-order leakage over any $(d-r)$-dimensional plastic subspace is $\sigma_{r+1}(J)$.

The persistent safe-base move generally does not reproduce the unrestricted finite endpoint. AFM therefore performs a second, explicitly separate function-space operation. A bounded compact-cardinal residual restores active finite protected outputs, safeguards unselected certified candidates, serves a selected finite transfer target when requested, and completes the current minibatch to the same logits as the no-protection comparator when finite consistency, support, capacity and numerical checks pass. The residual is zero outside its finite support union. Failed checks reject the transaction atomically; they do not authorize installation of the unrestricted base endpoint or reduction of the declared persistent-assimilation requirement.

AFM additionally supports function-preserving structural renewal through a fixed pool of dormant zero-gated modules. Resetting internal parameters while the functional gate is exactly zero leaves the deployed predictor and active protected behaviours unchanged. After reset, AFM recomputes the gradient, protected projector and certificates, and a module is activated only through an accepted protected update. The integrated theory combines these deterministic mechanisms with time-uniform consolidation, evidence-based reopening, route-refinement conditions, bounded resources, a non-convex projected-stationarity identity and a separate convex dynamic-regret specialization. Complete definitions, proofs, obstruction results, algorithm pseudocode, benchmark protocols, ablations, runtime analysis and controlled mechanism tests are provided throughout this paper.

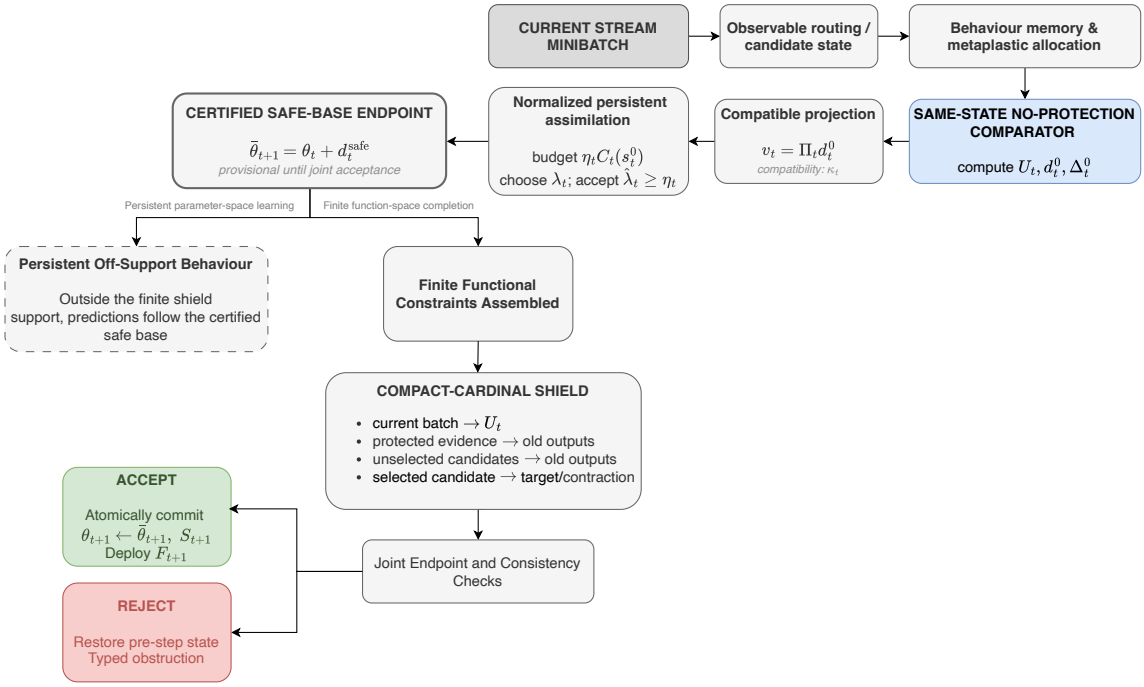

Figure 6: **Full AFM protected-update transaction.** Detailed original AFM transaction showing same-state comparator construction, compatibility projection, normalized persistent assimilation, finite functional constraints, compact-cardinal shield, endpoint checks and accept/reject logic.

5.7 Natural-state validation

The natural validation removed compatibility targeting. Starting from the same parent trajectories, 50 pre-update states were sampled at a fixed every-tenth-step schedule over the next 500 ordinary supervised updates for each system and original seed. No state was selected using its measured compatibility. Validation branches were discarded after measurement; only the ordinary parent update was committed, so the measurement branch did not alter subsequent trajectory state. The 750 natural states recorded the supervised gradient, naturally occurring κ, unrestricted decrease, method-specific persistent decrease, protected drift and retention pass. The common stored tolerance was $D \leq 0.005$.

5.8 Independent-direction experiment

For each requested nonzero compatibility $\{0.1, 0.25, 0.5, 0.75\}$, four independently constructed directions were generated in each of the three original systems and five seeds. All seven retention budgets and seven proposal branches were retained. Validation reported 147,000 fixed groups and 588,000 expected and observed frontier rows. Direction sensitivity was summarized by

$$R_{\text{dir}/\kappa} = \frac{\text{SD across directions at fixed } \kappa}{\text{SD across compatibility levels}}. \tag{137}$$

At $\beta = 0.5$, the mean fraction of states whose within-compatibility directional standard deviation was below the between-compatibility variation was 1.0 in all three systems.

5.9 Multiscale and generality tests

The multiscale experiment retained the original three systems, five seeds, seven methods, six requested compatibility levels and seven retention budgets. Internal scale coordinates $s \in \{0.05, 0.2, 0.5, 0.9\}$ form a

logarithmic expansion from the original local intervention to a common peak and are not fractions of a natural update. Validation reported 882,000 expected and observed frontier rows with all 15 runs complete. A separate calibration compared update norms with the median unrestricted update norm from natural states. The ten-seed generality experiment doubled seed trajectories and added a stronger vision transformer; reported confidence intervals use 100,000 deterministic Monte Carlo bootstrap resamples over seeds.

5.10 Near-zero boundary assessment

Among accepted requested-zero native-AFM rows, 65 had a negative empirical margin $M = \rho - \widehat{\lambda}\kappa/3$: 52 convolutional-network, one vision-transformer and 12 text rows. None of these rows had certification of the relevant finite curvature or step condition. Numerical verification agreed with the reported compatibility values to within 1.5883×10^{-9} and with the persistent-ratio values exactly. These cases are therefore not counted as certified theorem violations; $\widehat{\lambda}\kappa/3$ is called a theorem-aligned empirical reference when finite assumptions were not independently certified.

5.11 Fixed-norm bridge

The bridge used the convolutional network and CIFAR-10 vision transformer, ten seeds and 50 preselected states per seed. Target compatibilities were $\{0.1, 0.25, 0.5, 0.75\}$ and target update norms were 1%, 10%, 50% and 100% of the architecture-specific median natural unrestricted update norm. A state-scale condition was feasible only when all four target-compatibility directions produced positive unrestricted finite progress at the same target norm. States were preselected independently of the resulting feasibility outcomes.

At fixed norm and direction, finite unrestricted progress is an outcome of the nonlinear loss surface and is therefore measured rather than constrained to match across compatibility levels. The bridge admission criterion is that all four target-compatibility directions produce positive unrestricted finite progress at the prescribed common norm. Under this criterion, 232 state-scale conditions were feasible among 4,000 attempted conditions. Validation reported maximum compatibility error 9.656×10^{-6}, maximum relative norm error 3.674×10^{-5}, 14,848 expected and observed frontier rows, 928 native-AFM rows and no validation failures.

For the 231 feasible vision-transformer states at 1% natural norm, primary bridge slopes were estimated within state and summarized across ten seeds. Because Δ_0 is post-intervention in the fixed-norm experiment, absolute persistent progress $\Delta_{\text{persistent}}$ is the primary causal outcome and ρ is a secondary decomposition. The 231 states were additionally divided into tertiles of within-state Δ_0 coefficient of variation with boundaries 0.07620 and 0.33483. Regression adjustment for realized Δ_0 is treated only as a sensitivity/decomposition analysis, not a primary causal estimand.

5.12 Chronological AFM benchmarks

The original AFM evaluation used CORe50, CLEAR-10 and CLAD-C, the chronological object-classification component of CLAD built from labelled SODA10M imagery.(Han et al., 2021) All compared methods used the same compact convolutional predictive base and seed-specific initialization. A common representation prefix was learned and then frozen; thereafter methods operated on the same available head and adapter coordinates without learner-visible task or boundary identifiers. AFM was evaluated at $\eta \in \{0.10, 0.50, 1.00\}$. Classical controls included no protection, matched SGD, replay, A-GEM(Chaudhry et al., 2019) and online EWC. The modern comparison used OCAR, LPR, aL-SAR, FGH, CCL-DC and MKD. The common predictive base, stream, representation-freeze boundary, checkpoints and five seeds were held fixed, but method-native auxiliary state was retained; the comparison is therefore not an equal-storage or equal-compute study.

Retention R was one minus mean peak-to-final forgetting, clipped to $[0, 1]$. Plasticity P was the mean prequential online accuracy over the first and last quarters of predeclared adaptation episodes. The primary family score was origin-referenced hypervolume of non-dominated (R, P) operating points. Raw hypervolume is protocol-specific and is not interpreted as directly comparable across datasets. Paired five-seed primary benchmark comparisons used percentile bootstrap intervals; additional five-seed analyses used exact enumeration of all 5^5 resamples. Full protocol definitions and benchmark result tables are provided in the detailed evaluation sections below.

5.13 Statistics

For the original causal analysis, a least-squares slope of ρ on realized κ was computed within each causal state across the six interventions. Fifty state slopes were averaged within each seed; the five seed means were the inferential replicates for each system, with two-sided 95% Student-t intervals (four degrees of freedom). Follow-up five-seed suites use the same state-to-seed aggregation. Ten-seed generality and bridge intervals use a deterministic 100,000-resample seed bootstrap. The bridge feasibility rate is reported over all attempted states and is not replaced by conditional resampling of successful states. No method-level row count is used as an independent-sample size.

6 Chronological AFM evaluation and benchmark evidence

6.1 Experimental evaluation

The experiments address three questions: whether the AFM transaction yields a useful retention-plasticity frontier on real chronological streams, whether that frontier remains competitive against recent continual-learning methods under a common predictive architecture, and whether the internal execution behaves consistently with the theorem-aligned quantities. The nonlinear experiments use complete empirical endpoint checks; they are not presented as outward-rounded pre-step numerical certificates of the full nonlinear theorem.

6.1.1 Datasets, architecture, and protocol

We evaluate on CORe50 (Lomonaco & Maltoni, 2017), CLEAR-10 (Lin et al., 2021), and CLAD-C (Verwimp et al., 2023), the chronological object-classification component of CLAD built from labeled SODA10M imagery (Han et al., 2021). Table 28 summarizes the frozen protocols. All methods use the same compact convolutional predictive base and seed-specific initialization. A common representation prefix is learned and then frozen; after that boundary, all methods operate on the same available head and adapter coordinates. No method receives task or boundary identifiers, and no challenger receives an external pretrained representation in the controlled comparison.

AFM is evaluated at the predeclared coordinates $\eta \in \{0.10, 0.50, 1.00\}$. The classical comparison matrix includes the genuine no-protection counterpart, matched SGD, experience replay (Rolnick et al., 2019), A-GEM (Chaudhry et al., 2019), online EWC, and a task-aware oracle EWC reference (Kirkpatrick et al., 2017). Five paired seeds are used throughout. Benchmark protocol specification and the separation between exploratory and confirmatory runs are documented in Section 6.2.

6.1.2 Metrics and primary analysis

For semantic regime r, let $A_{j,r}$ denote checkpoint-j accuracy, with $i(r)$ its introduction checkpoint and $e(r)$ its final valid checkpoint. Forgetting and backward transfer are

$$F_r = \max_{i(r)\le j\le e(r)} A_{j,r} - A_{e(r),r}, \qquad \mathrm{BWT}_r = A_{e(r),r} - A_{i(r),r}. \tag{138}$$

The retention score is

$$R = \mathrm{clip}\left(1 - \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} F_r, 0, 1\right). \tag{139}$$

Plasticity P is the mean online accuracy over the first and last quarters of the predeclared adaptation episodes. Online accuracy is prequential: for each incoming minibatch, predictions and minibatch accuracy are recorded before that minibatch is used for the learning update. The same prediction-before-update ordering is used during the initial representation-learning prefix. The primary score for a method family is the origin-referenced area dominated by its non-dominated (R, P) operating points,

$$\mathrm{HV}(\mathcal{Q}) = \mu\left(\bigcup_{(R,P)\in \mathrm{ND}(\mathcal{Q})} [0,R]\times[0,P]\right). \tag{140}$$

Table 28: Experimental protocols. All learners receive labels for supervised updates but no task, session, episode, bucket, segment, intervention, or boundary identifiers.

	CORe50	CLEAR-10	CLAD-C
Dataset structure	Official 128×128 CORe50 stream Lomonaco & Maltoni (2017); 10 classes and 12 predeclared episodes covering new contexts, recurrence, long-dormancy return, an explicit $0 \leftrightarrow 1$ target conflict, gradual visual drift, and capacity pressure.	Official CLEAR-10 imagery Lin et al. (2021); 11 labels including background, 10 chronological supervised buckets, and natural temporal drift.	CLAD-C chronological object classification Verwimp et al. (2023) on labeled SODA10M Han et al. (2021); 6 classes and 6 official chronological training segments. Only labeled object crops are used: no SODA10M unlabeled pretraining, no CLAD-D, and no external pretrained backbone.
Learner stream	Batch size 32; task/session/episode metadata withheld; evaluator-only semantic regimes and validity intervals.	32960 learner examples, 3296 per bucket, batch size 32, 1030 updates; bucket and period metadata withheld.	22249 official object crops in six preserved segments with item counts 5157, 1154, 6742, 2560, 4517 and 2119; maximum batch size 10 with partial boundary batches retained. Segment metadata and original labels remain evaluator-only.
Representation prefix	20 batches / 640 images, then backbone freeze.	Complete first supervised episode: 103 batches / 3296 images, then backbone freeze. Optional unlabeled bucket-0 pretraining is not used.	Complete first official training segment: 516 batches / 5157 crops, then backbone freeze.
Candidate fitting	256 routed examples, 20 fixed epochs, validation horizon 2048, risk threshold 0.7.	256 routed examples, 100 fixed epochs, validation horizon 4096, risk threshold 0.7.	256 routed examples, 100 fixed epochs, validation horizon 4096, risk threshold 0.7.
Evaluation	Checkpoint matrix over active semantic regimes; held-out test and complete original-semantics evaluator.	Checkpoint at every bucket boundary; 1137 validation and 4363 held-out test examples from supervised buckets 1-10; next-bucket evaluation for near-future accuracy.	Checkpoint after every official training segment; the official data loader yields 32967 validation and 46915 held-out test object crops.
Matrix	17 variants per seed and 5 seeds: 85 jobs per dataset, 255 jobs total across the three frozen matrices.		

Raw hypervolume is protocol-specific and is not interpreted as directly comparable across datasets. The paired primary statistic is the within-dataset difference between AFM and the strongest eligible non-oracle family. For the original five-seed primary comparisons, 95% confidence intervals are paired percentile-bootstrap intervals over the five seedwise AFM-minus-comparator differences. We use 20,000 resamples with fixed RNG seed 20260729; each resample draws five paired differences with replacement and recomputes their mean, and

Table 29: Primary paired retention-plasticity hypervolume results across five seeds. Confidence intervals are paired across seeds; relative improvement is normalized by the comparator mean.

Dataset	Comparator	AFM HV	Comparator HV	Mean Δ	95% CI	Wins	Relative
CORe50	No protection	0.2478	0.2388	+0.0090	[0.0059, 0.0126]	5/5	3.76%
CLEAR-10	No protection	0.3807	0.3740	+0.0068	[0.0053, 0.0082]	5/5	1.81%
CLAD-C	A-GEM	0.5129	0.5019	+0.0110	[0.0054, 0.0166]	5/5	2.20%

Table 30: Five-seed mean family hypervolume for AFM and the classical comparison families. Oracle EWC receives task information and is reported only as a task-aware reference.

Family	CORe50	CLEAR-10	CLAD-C
AFM	**0.2478**	**0.3807**	**0.5129**
No protection	0.2388	0.3740	0.4872
Matched SGD	0.2031	0.3596	0.4714
Replay	0.1839	0.3503	0.4797
A-GEM	0.1845	0.3521	0.5019
Online EWC	0.1829	0.3511	0.4834
Oracle EWC	0.1823	0.3506	0.4478

the interval endpoints are the 2.5% and 97.5% order-statistic positions. For subsequent five-seed ablation and modern-comparison analyses, the paired bootstrap is evaluated exactly by enumerating all $5^5 = 3125$ possible resamples and taking the 2.5th and 97.5th percentiles with linear interpolation.

6.1.3 Classical comparison

Table 29 reports the confirmatory primary comparison. AFM exceeds the strongest eligible non-oracle family on every paired seed in all three protocols. The five-seed mean hypervolumes are 0.248 on CORe50, 0.381 on CLEAR-10, and 0.513 on CLAD-C. The corresponding paired mean advantages are approximately 0.0090, 0.0068, and 0.0110, and each paired confidence interval remains above zero.

Table 30 places the result in the complete classical comparison matrix. No protection is the strongest classical non-oracle family on CORe50 and CLEAR-10, while A-GEM is the strongest on CLAD-C. The task-aware oracle EWC reference is not eligible for the primary non-oracle comparison.

AFM is a frontier rather than a single fixed operating point. Table 31 shows the expected movement from stronger preservation at $\eta = 0.10$ toward greater acquisition at $\eta = 1.00$. On CLEAR-10, all three coordinates reduce forgetting relative to no protection while maintaining similar plasticity. On CLAD-C, the conservative coordinate substantially reduces forgetting and improves final-test accuracy, whereas the full-assimilation coordinate approaches unrestricted acquisition and exhibits more peak-to-final forgetting. That reversal is examined at class level in Section 6.2.

6.1.4 Comparison with modern continual-learning methods

The modern comparison includes OCAR (Urettini & Carta, 2025), LPR (Yoo et al., 2024), aL-SAR (Seo et al., 2025), FGH (Michel et al., 2026), CCL-DC (Wang et al., 2024), and MKD (Michel et al., 2024). These methods span curvature-aware replay, layerwise proximal replay, adaptive freezing and retrieval, learned hypergradient reweighting, collaborative distillation, and momentum-teacher distillation. Each challenger uses one frozen operating configuration. The challengers are not themselves defined by a common predeclared three-point retention-plasticity coordinate analogous to AFM's η; accordingly, the comparison uses their frozen method-native operating configurations rather than constructing method-specific sweeps over unrelated tuning axes solely to equalize frontier cardinality. The common predictive base, learner-visible chronological stream, representation-freeze rule, checkpoints, and seeds are held fixed. Method-native auxiliary state is

Table 31: Five-seed mean operating points. Fgt. denotes mean forgetting, BWT backward transfer, Test held-out final-test accuracy, Online chronological stream accuracy, and NF CLEAR-10 next-bucket accuracy.

Dataset	Point	R	P	Fgt.↓	BWT↑	Test↑	Online↑	NF↑
CORe50	AFM 0.10	0.9648	0.2088	0.0352	-0.0186	0.2291	0.2125	-
CORe50	AFM 0.50	0.9185	0.2458	0.0815	-0.0658	0.2362	0.2594	-
CORe50	AFM 1.00	0.9020	0.2595	0.0980	-0.0842	**0.2398**	0.2764	-
CORe50	No protection	0.8866	**0.2693**	0.1134	-0.0991	0.2341	**0.2866**	-
CLEAR-10	AFM 0.10	**0.9933**	0.3750	**0.0067**	**0.0056**	0.3412	0.3753	0.3263
CLEAR-10	AFM 0.50	0.9828	0.3793	0.0172	-0.0013	**0.3417**	0.3799	0.3309
CLEAR-10	AFM 1.00	0.9768	**0.3834**	0.0232	-0.0065	0.3386	**0.3835**	0.3329
CLEAR-10	No protection	0.9756	0.3834	0.0244	-0.0073	0.3376	0.3834	**0.3338**
CLAD-C	AFM 0.10	**0.8534**	0.5706	**0.1466**	-0.0081	**0.5732**	0.5713	-
CLAD-C	AFM 0.50	0.7915	0.5874	0.2085	0.0293	0.5440	0.6025	-
CLAD-C	AFM 1.00	0.7433	0.6040	0.2567	**0.0382**	0.5457	0.6155	-
CLAD-C	No protection	0.7685	**0.6341**	0.2315	0.0370	0.5497	**0.6409**	-

Table 32: Five-seed mean retention-plasticity hypervolume for AFM and six modern challengers. AFM contributes its predeclared three-point frontier; each challenger contributes the RP area of its fixed operating configuration.

Dataset	AFM HV	CCL-DC HV	MKD HV	FGH HV	aL-SAR HV	LPR HV	OCAR HV
CORe50	0.2478	0.2016	0.1799	**0.2546**	0.1801	0.1640	0.1436
CLEAR-10	**0.3807**	0.3731	0.3513	0.3773	0.3572	0.3485	0.3164
CLAD-C	**0.5129**	0.4921	0.4931	0.3058	0.3296	0.4694	0.3747

retained rather than removed: for example, CCL-DC keeps its second learner and replay state, and MKD keeps its EMA teacher and replay state. The comparison is therefore controlled for the predictive base and information interface, but it is not described as equal-storage or equal-compute. Detailed compatibility decisions are reported in Section 6.2.

Table 32 gives the dataset-level primary scores. These values compare AFM's declared operating family with the challengers' frozen method-native configurations; they are not estimates of the challengers' complete achievable hypervolumes under arbitrary hyperparameter sweeps. AFM has the highest mean primary score on CLEAR-10 and CLAD-C. FGH is the sole dataset-level mean reversal, exceeding AFM on CORe50 by about 2.8% while operating at a more acquisitive point with lower retention and greater forgetting than every AFM coordinate. AFM exceeds FGH slightly on CLEAR-10 and substantially on CLAD-C.

For a compact description of conventional metrics, one AFM coordinate is selected per dataset using AFM's own five-seed mean final-test accuracy, independently of challenger identity: $\eta = 1.00$ on CORe50, $\eta = 0.50$ on CLEAR-10, and $\eta = 0.10$ on CLAD-C. Table 33 shows the equal-weighted descriptive summary. AFM has the highest aggregate plasticity, online accuracy, and primary score; CCL-DC has the highest final-test accuracy, OCAR the highest retention and lowest forgetting, and aL-SAR the highest mean BWT. The comparison therefore does not support a claim that AFM maximizes every scalar metric. Its advantage is a stronger joint acquisition-retention profile.

Table 34 quantifies that trade-off against each challenger. The closest aggregate primary challenger is CCL-DC: AFM is about 7.0% higher in the descriptive primary score, with about 4.7% greater plasticity and 3.5% higher online accuracy, while CCL-DC retains higher mean retention and final-test accuracy. Against the six-challenger mean, AFM has a 19.9% higher primary score, together with higher plasticity, online accuracy, and final-test accuracy, but lower mean retention.

Table 33: Equal-weighted descriptive summary of conventional metrics across the three datasets. AFM uses the coordinate with highest AFM five-seed mean final-test accuracy on each dataset; the Primary column remains the mean of the declared benchmark-level primary scores.

Method	$R\uparrow$	$P\uparrow$	Fgt.↓	BWT↑	Online↑	Test↑	Primary↑
AFM	0.9127	**0.4031**	0.0873	-0.0312	**0.4092**	0.3849	**0.3805**
CCL-DC	0.9424	0.3850	0.0576	0.0081	0.3955	**0.3950**	0.3556
MKD	0.9634	0.3604	0.0366	0.0175	0.3714	0.3916	0.3414
LPR	0.9538	0.3512	0.0462	0.0202	0.3634	0.3799	0.3273
FGH	0.8716	0.3592	0.1284	-0.0215	0.3816	0.3591	0.3126
aL-SAR	0.9625	0.3018	0.0375	**0.0343**	0.3277	0.2934	0.2890
OCAR	**0.9952**	0.2799	**0.0048**	0.0050	0.3041	0.2791	0.2782

Table 34: AFM gain or loss relative to each modern challenger in the equal-weighted descriptive summary. R, P, Online, Test, and Primary are relative changes; Fgt. and BWT are AFM-minus-challenger percentage-point differences.

Comparison	ΔR	ΔP	ΔFgt.	ΔBWT	ΔOnline	ΔTest	ΔPrimary
AFM vs CCL-DC	-3.14%	+4.72%	+2.96 pp	-3.93 pp	+3.46%	-2.55%	**+7.01%**
AFM vs MKD	-5.26%	+11.86%	+5.07 pp	-4.87 pp	+10.19%	-1.70%	**+11.44%**
AFM vs LPR	-4.31%	+14.80%	+4.11 pp	-5.14 pp	+12.62%	+1.32%	**+16.26%**
AFM vs FGH	+4.72%	+12.24%	-4.12 pp	-0.97 pp	+7.22%	+7.19%	**+21.73%**
AFM vs aL-SAR	-5.17%	+33.58%	+4.98 pp	-6.55 pp	+24.88%	+31.20%	**+31.67%**
AFM vs OCAR	-8.29%	+44.00%	+8.25 pp	-3.62 pp	+34.54%	+37.93%	**+36.76%**
AFM vs mean of six challengers	-3.74%	+18.72%	+3.54 pp	-4.18 pp	+14.53%	+10.08%	**+19.90%**

Contemporary methods outside the common-architecture comparison. The primary comparison is designed to isolate differences in continual-learning strategy while holding the predictive architecture, representation source, and learner-visible stream fixed. We therefore distinguish between methods that can be instantiated on the common compact network without removing their defining mechanism and methods whose contribution is intrinsically coupled to a different architecture, pretrained representation, or method-specific structural capacity. The latter remain important scientific comparators, but including them in the same numerical table would change more than the continual-learning rule itself.

DEMD is a task-free method that represents past experience through a dynamic memory distribution whose structure is expanded, augmented, and reduced as streaming novelty changes (Ye & Bors, 2025). AdaLin addresses a different aspect of the stability-plasticity problem by adaptively controlling neuronal linearity to maintain useful gradient propagation during continual learning (Rohani et al., 2026). Both methods were considered for direct empirical comparison. However, when the experimental protocol was frozen, a reproducible authors' implementation suitable for integration into the controlled same-network evaluation was not available for either method. Their published experiments also use datasets and evaluation protocols that are not directly commensurate with the CORe50, CLEAR-10, and CLAD-C streams used here. Importing their reported accuracies would therefore mix results obtained under different architectures, data streams, tuning procedures, and evaluation rules, while an independent reimplementation would introduce an additional implementation variable into an otherwise controlled comparison. We consequently discuss DEMD and AdaLin as relevant contemporary approaches but do not assign them numerical entries in the primary table.

SinglePrompt is excluded for a different reason. Its task-free adaptation mechanism places learned prompts inside Transformer self-attention blocks (Park et al., 2026). The common predictor used in our controlled comparison is a compact convolutional network with residual adapter capacity and contains no self-attention blocks in which the SinglePrompt mechanism can be instantiated. Adding a Transformer solely for this comparator would replace the shared predictive architecture and alter both parameterization and representation geometry. Conversely, replacing SinglePrompt's prompt mechanism by an adapter compatible with the

common network would no longer constitute an evaluation of the published method. We therefore regard SinglePrompt as a relevant task-free architectural alternative rather than a valid member of the same-base numerical comparison.

Online-LoRA likewise relies on assumptions that conflict directly with the frozen experimental interface. The method performs online low-rank adaptation of an externally pretrained Vision Transformer (Wei et al., 2025). External pretrained representations are deliberately excluded from the primary protocol: every evaluated method receives the same representation learned from the declared learner-visible prefix before the backbone is frozen. Supplying a pretrained Vision Transformer only to Online-LoRA would therefore change both the initial representation and the model architecture, whereas removing the pretrained Transformer would remove a defining component of the published method. Its reported results are consequently not numerically combined with the common-network comparison.

The remaining structural methods also change the learner in ways that cannot be separated cleanly from their continual-learning mechanisms. SERENA creates and freezes specialized concept cells within an over-parameterized architecture (Yildirim et al., 2026); hence its stability mechanism is tied to method-specific structural capacity rather than to an update rule that can be applied unchanged to the common predictor. EG-CNN makes continual evolution of its feature extraction, refinement, and classification components part of the learning process itself (Leite, 2026). Evaluating EG-CNN on the fixed common architecture would therefore suppress the structural evolution that defines the method, while permitting that evolution would break the architecture-matched control.

S6MOD augments the learner with an additional state-space-model branch and class-conditional routing after the backbone (Liu et al., 2025). These components introduce both additional adaptive capacity and a different routing mechanism, so a comparison against the unchanged common network would conflate the continual-learning strategy with an architectural expansion. Dual-Arch makes this distinction still more explicit by assigning stability and plasticity to separate lightweight networks with different functional roles (Lu et al., 2025). Collapsing the method into the single common predictor would remove its defining dual-network construction; retaining both networks would instead compare different model systems and different adaptive capacities.

For these reasons, SinglePrompt, Online-LoRA, SERENA, EG-CNN, S6MOD, and Dual-Arch are not omitted because they are considered less relevant or less competitive. They answer the stability-plasticity problem partly through architectural or representation choices that lie outside the controlled question addressed by the primary experiment: how different continual-learning rules behave when the underlying predictive network, representation source, learner-visible information, and chronological stream are held fixed. Their proper empirical assessment would require a separate architecture- and resource-aware comparison in which model capacity, pretraining, computational cost, and auxiliary state are treated as experimental variables rather than held constant.

6.1.5 Ablation and theorem-aligned execution evidence

The main ablation asks whether the joint AFM transaction improves on a persistent protected base update alone. Table 35 reports the paired CORe50 result at $\eta = 0.50$. Full AFM improves the single-point $R \times P$ score by about 18.6% relative to both base-only variants, with positive paired intervals. In contrast, collapsing the multiscale trace bank to one timescale produces essentially the same aggregate score at this operating point. The latter result does not invalidate the multiscale theorem; it shows that this particular CORe50 coordinate does not empirically require the full timescale bank. Complete conventional metrics for the ablation are given in Section 6.3.

Table 36 summarizes the execution-level checks over all protected AFM coordinates. Across the three datasets, accepted finite restorations keep the maximum endpoint error below 10^{-6}. The minimum deployed progress ratio remains extremely close to one, and the realized persistent-base progress ratio exceeds the analytic lower bound on every accepted protected update examined. The near-one deployed ratio verifies execution of the finite endpoint-completion mechanism; it is not interpreted as evidence of population retention away from the finite shield support. The high precision in this table is retained because the numerical margins themselves are the object being audited.

Table 35: Paired five-seed CORe50 $R \times P$ ablation results. Differences are full AFM at $\eta = 0.50$ minus the indicated ablation.

Ablation	Full AFM	Ablation	Mean Δ	95% CI	Wins	Relative gain
Base-only, finite normalized budget	0.2257	0.1903	**+0.0354**	**[+0.0276, +0.0445]**	**5/5**	**+18.61%**
Base-only, fixed-absolute budget	0.2257	0.1903	**+0.0354**	**[+0.0266, +0.0446]**	**5/5**	**+18.59%**
Single-timescale AFM	0.2257	0.2257	+0.0000	[-0.0008, +0.0005]	4/5	+0.01%

Table 36: Execution-level mechanism audit over the three protected AFM coordinates and five seeds on each real-world benchmark. The analytic margin is the minimum realized persistent-base progress ratio minus the theorem lower bound.

Dataset	Runs	Certs	Commits	Protected nonzero	Exact restore accepted/attempts	Max endpoint error	Min deployed ratio	Min analytic margin
CORe50	15	90	90	8621	8636/8829	4.768×10^{-7}	0.99999494	$+2.107 \times 10^{-3}$
CLEAR-10	15	99	99	10416	10431/10494	9.537×10^{-7}	0.99999718	$+2.940 \times 10^{-2}$
CLAD-C	15	488	488	22917	22932/24980	9.537×10^{-7}	0.99993644	$+1.378 \times 10^{-4}$

Component-level computational cost. The present AFM implementation has substantial wall-clock overhead: median runtime is approximately 35× matched SGD on CORe50, 15× on CLEAR-10, and 184× on CLAD-C. To identify the source of this cost, we instrumented the existing implementation without changing the learning rule, frozen hyperparameters, data streams, or operating-point selection. Profiling used the independently selected conventional-metric operating points, $\eta = 1.0$ on CORe50, $\eta = 0.5$ on CLEAR-10, and $\eta = 0.1$ on CLAD-C, over the same five frozen seeds used elsewhere in the study. For each profiled operation we recorded its accumulated host-observed elapsed time, invocation count, and per-invocation duration, while recording CUDA-event elapsed time separately. The default profiling mode deliberately avoids per-region CUDA synchronization to limit instrumentation-induced perturbation.

Table 37 shows that the host-observed execution cost is highly concentrated. The component-wise median shares of protection geometry, finite counterfactual completion, and endpoint verification sum to 94.82% on CORe50, 91.70% on CLEAR-10, and 95.78% on CLAD-C. Ordinary learning accounts for only 0.36%, 0.44%, and 0.24%, respectively, while construction and evaluation of the same-state no-protection comparator accounts for 0.52%, 0.27%, and 0.25%. The host-timing decomposition is therefore concentrated in constructing and verifying protected executable updates rather than in the ordinary forward/backward regions or the counterfactual-comparator region. Because CUDA execution is asynchronous in this profiling mode, these component shares characterize host-observed execution time and are not interpreted as a synchronized decomposition of device execution time.

The dominant category varies with the stream. Protection geometry is largest on CORe50 at 37.33% and on CLAD-C at 42.06%, whereas endpoint verification dominates CLEAR-10 at 52.68%. Finite counterfactual completion remains substantial on all three datasets, accounting for 25.95%, 22.90%, and 28.55%, respectively. The operation-level CLAD-C decomposition in Section 6.3 separates these costs into invocation frequency and per-call duration and identifies the principal implementation-level bottlenecks.

6.1.6 Empirical scope and limitations

The empirical evidence supports a restricted claim. AFM yields a positive paired frontier advantage over the strongest tested classical non-oracle family on all five seeds of each frozen protocol, and its primary performance remains competitive or superior across six recent challengers under the common predictive-base interface. The evaluated neural protocols learn a common representation prefix and then freeze that representation, so the empirical results establish continual adaptation over the shared learned representation rather than unrestricted end-to-end continual representation learning. The experiments do not establish universal empirical dominance or strict outward numerical certification of the nonlinear theorem.

Table 37: Component-level host-timing decomposition of AFM at the independently selected operating points. Entries are median percentages across the five frozen seeds. Profiling uses low-perturbation host timers without per-region CUDA synchronization; CUDA-event timing is recorded separately. Because each component is summarized independently, columns need not sum exactly to 100%.

Component	CORe50	CLEAR-10	CLAD-C
Ordinary learning	0.36%	0.44%	0.24%
Same-state no-protection comparator	0.52%	0.27%	0.25%
Protection geometry	**37.33%**	16.12%	**42.06%**
Candidate machinery	1.94%	2.17%	2.00%
Finite counterfactual completion	**25.95%**	**22.90%**	**28.55%**
Endpoint verification	**31.54%**	**52.68%**	**25.17%**
Routing/metaplastic/control	1.05%	1.77%	0.88%
Other/unattributed	1.87%	3.48%	1.39%

Two limitations are particularly informative. First, all protected CLAD-C runs report insufficient signature calibration for a positive route-identification conclusion, so CLAD-C supports the assimilation, restoration, transfer, and finite-protection mechanisms but not a positive routing claim. Second, the CLAD-C full-assimilation forgetting increase is concentrated in Pedestrian and reflects a population-level collapse that also occurs in the no-protection learner after a Pedestrian-free chronological interval. Exact finite Pedestrian evidence can remain protected while held-out population accuracy collapses. The classwise analysis and an inverse-frequency diagnostic are reported in Section 6.2.

6.2 Complete experimental results and protocol details

6.2.1 Benchmark protocol specification and analysis separation

All reported five-seed benchmark results use protocols fixed before analysis of the corresponding benchmark outcomes. For CLEAR-10, protocol-development diagnostics conducted before the five-seed evaluation determined the shared representation prefix and candidate-fitting budget using candidate-side fitting and activation behaviour only; held-out test accuracy and hypervolume were not used. The resulting protocol uses the complete first supervised chronological episode as the shared representation prefix and 100 candidate-fitting epochs. Protocol-development runs are excluded from all reported five-seed benchmark summaries. The risk threshold, counterfactual comparator, normalized assimilation rule, and finite restoration construction follow the specification given in this paper.

For CLAD-C, the study uses the chronological object-classification stream defined by CLAD (Verwimp et al., 2023) over labeled SODA10M data (Han et al., 2021). Only labeled bounding-box crops are used; unlabeled pretraining and the detection task are excluded. The official loader version used here yields 22,249 training crops, 32,967 validation crops, and 46,915 test crops; these are the exact cardinalities used in the reported experiments. The official ordering is preserved, and examples are not moved or discarded to match cardinalities from earlier benchmark releases. Segment identity and original-label side information remain evaluator-only. Loader and frontier integrity checks use a separate development seed that is excluded from all reported five-seed results; no CLAD-C test score or exploratory classwise diagnostic is used to tune AFM.

Finite-shield support and guards. The finite shield uses the predeclared compact-address support construction of Section 2.5. The replay envelope ε_a was fixed at 10^{-8}, the support multiplier was fixed at $\kappa = 4$, and the label-free guard bank was bounded by $Q_{\max} = 640$ addresses. These quantities are shield support parameters rather than task-routing thresholds. Conditional on an accepted transaction, the requested value at each constraint center is reproduced exactly; changing the support parameters does not define an approximate center-matching tolerance. Instead, they determine the spatial extent of the finite residual and whether the required center-center and center-guard separation conditions are satisfied, thereby affecting acceptance or obstruction of a shield transaction.

Table 38: Per-seed hypervolume for AFM and the strongest non-oracle comparator within each protocol. Hypervolume magnitudes are not compared across datasets.

Seed	CORe50 AFM	CORe50 no-prot.	Δ	CLEAR AFM	CLEAR no-prot.	Δ	CLAD AFM	CLAD A-GEM	Δ
11	0.2481	0.2403	+0.0078	0.3852	0.3763	+0.0089	0.5065	0.4961	+0.0104
29	0.2281	0.2205	+0.0076	0.3693	0.3652	+0.0041	0.5001	0.4951	+0.0051
47	0.2520	0.2371	+0.0149	0.3836	0.3778	+0.0058	0.5231	0.5198	+0.0033
71	0.2621	0.2586	+0.0036	0.3752	0.3684	+0.0069	0.5216	0.5010	+0.0207
101	0.2489	0.2377	+0.0111	0.3903	0.3822	+0.0081	0.5133	0.4977	+0.0156

Table 39: AFM operating-point trade-offs relative to no protection. Test differences are percentage points; positive values favor AFM.

Dataset	Coordinate	Forgetting reduction	Relative reduction	Test difference
CORe50	0.10	0.0783	69.00%	-0.499
CORe50	0.50	0.0319	28.12%	+0.205
CORe50	1.00	0.0154	13.59%	+0.565
CLEAR-10	0.10	0.0177	72.66%	+0.358
CLEAR-10	0.50	0.0072	29.42%	+0.408
CLEAR-10	1.00	0.0012	5.06%	+0.101
CLAD-C	0.10	0.0850	36.70%	+2.356
CLAD-C	0.50	0.0231	9.96%	-0.569
CLAD-C	1.00	-0.0252	-10.87%	-0.394

6.2.2 Classical seedwise results and operating-point trade-offs

Table 38 gives the paired seedwise comparison underlying the classical primary analysis. The sign of the difference is positive for every seed in all three protocols.

Table 39 expresses the operating-point trade-offs relative to no protection. The conservative AFM coordinate strongly reduces forgetting on all three datasets. Higher η shifts the frontier toward acquisition, as intended.

For CLEAR-10 at $\eta = 0.50$, the paired 95% interval for the final-test difference relative to no protection is approximately $[+0.11, +0.77]$ percentage points. This interval is computed by exact paired bootstrap over the five seedwise AFM-minus-no-protection final-test differences: all $5^5 = 3125$ resamples with replacement are enumerated, each replicate is the mean of five resampled paired differences, and the 2.5% and 97.5% order-statistic endpoints are used without percentile interpolation. Next-bucket accuracy is about 0.29 percentage points lower. Thus the balanced coordinate improves retained/final performance without dominating the unrestricted trajectory on every temporal generalization metric. On CLAD-C, A-GEM with memory 32 attains lower mean forgetting than AFM $\eta = 0.10$ but also lower plasticity; the family hypervolume therefore evaluates the complete retention-plasticity trade-off rather than selecting the single lowest-forgetting operating point.

6.2.3 Modern challenger compatibility

All six modern challengers are run on the same chronological observations, common predictive base, representation-freeze boundary, checkpoint schedule, and five seeds as AFM. The method-native state required by each algorithm is retained. OCAR and LPR use replay reservoirs of 128 learner-visible examples. aL-SAR retains its larger 4000-example memory; CCL-DC and MKD retain 1000-example replay memories together with, respectively, a second collaborating learner and a complete EMA teacher. Consequently, this experiment controls architecture and learner-visible information but not total storage or compute.

Only interface adaptations required by the common base are made. OCAR's curvature preconditioning and LPR's proximal preconditioning act on the available affine adapter and classifier coordinates. FGH's architecture-agnostic gradient-reweighting and prototype mechanisms are applied to the same trainable path

Table 40: Frozen configurations used for the six modern challengers. Settings are method-native unless a shared benchmark value is stated.

Method	Persistent auxiliary state	Optimization	Principal frozen settings
OCAR	Replay memory, 128 examples	Shared learning rate 10^{-3}	Curvature/Fisher, damping, and replay rules retained
LPR	Replay memory, 128 examples	Learning rate 0.01	$\omega_0 = 1$, $\beta = 1$, preconditioner refresh every 100 iterations
aL-SAR	Replay memory, 4000 examples	Adam, 3×10^{-4}	Update batch 16, online factor 0.015625, unfreeze rate 0.5, temperature 0.125, $k = 4$, warm-up 50
FGH	Prototype state	Adam, 5×10^{-3}	Hypergradient rate 1, gradient-weight clamp 1000, one online epoch, prototype coefficient 1
CCL-DC	Second learner, replay memory 1000	AdamW, 5×10^{-4}, weight decay 10^{-4}	Replay batch 64, one memory iteration, distillation weight 2, temperature 4
MKD	EMA teacher, replay memory 1000	Adam, 5×10^{-4}	Replay batch 64, distillation weight 5.5, temperature 4, EMA 0.01, correction interval 10

because its complete published prompt configurations require pretrained Transformer architectures that would violate the common-base condition (Michel et al., 2026). CCL-DC retains both collaborating peers and MKD retains the teacher; removing these components would change the methods (Wang et al., 2024; Michel et al., 2024).

Two compatibility consequences are reported because they affect interpretation. First, aL-SAR's public coordinate-sampling rule selects 0.01% of the weights of a layer after integer truncation. On the 4096-weight adapter matrices used here this gives zero sampled similarity coordinates, so retrieval on those blocks reduces to the method's frequency-balancing component. The rule is not altered to improve aL-SAR under the compact architecture (Seo et al., 2025). Second, FGH's prototype state is indexed by numeric class label. The CORe50 protocol contains an explicit $0 \leftrightarrow 1$ target conflict, but FGH receives neither the conflict boundary nor evaluator-side semantic remapping. This preserves the same learner-visible information restriction as AFM.

All 90 modern-challenger runs completed with valid evaluation records. The frozen configurations are summarized in Table 40.

Table 41 gives the complete seedwise primary scores. Across the 90 matched seed-dataset-challenger comparisons, AFM has the higher primary score in 84. At the dataset-mean level it is higher in 17 of 18 pairwise comparisons. Relative to CCL-DC, AFM's dataset-level mean primary scores are higher by approximately 22.9%, 2.1%, and 4.2% on CORe50, CLEAR-10, and CLAD-C, respectively. The corresponding gains over MKD are 37.8%, 8.4%, and 4.0%; over OCAR, 72.6%, 20.3%, and 36.9%; over LPR, 51.1%, 9.3%, and 9.3%; and over aL-SAR, 37.6%, 6.6%, and 55.6%. FGH is the only dataset-level reversal: it is about 2.8% higher on CORe50, whereas AFM is about 0.9% higher on CLEAR-10 and 67.8% higher on CLAD-C.

Table 42 reports the equal-weighted mean of the three benchmark-level primary scores. It is included only as a cross-protocol descriptive summary because the raw hypervolume scale is dataset specific.

Table 43 gives the complete conventional metrics for every AFM coordinate and modern challenger. It shows explicitly that several challengers exceed AFM on individual stability or final-accuracy measures even when AFM has the stronger joint primary profile.

Table 41: Per-seed retention-plasticity hypervolume for AFM and the six modern challengers. AFM contributes its three-point family hypervolume and each challenger its fixed-configuration RP area.

CORe50

Seed	AFM	CCL-DC	MKD	FGH	aL-SAR	LPR	OCAR
11	0.2481	0.2004	0.1854	0.2632	0.1875	0.1702	0.1483
29	0.2281	0.1904	0.1739	0.2452	0.1763	0.1618	0.1428
47	0.2520	0.2030	0.1827	0.2520	0.1878	0.1592	0.1345
71	0.2621	0.2030	0.1817	0.2620	0.1729	0.1672	0.1497
101	0.2489	0.2111	0.1758	0.2509	0.1761	0.1616	0.1427

CLEAR-10

Seed	AFM	CCL-DC	MKD	FGH	aL-SAR	LPR	OCAR
11	0.3852	0.3698	0.3413	0.3832	0.3573	0.3499	0.3101
29	0.3693	0.3747	0.3492	0.3736	0.3455	0.3379	0.3181
47	0.3836	0.3763	0.3646	0.3847	0.3641	0.3584	0.3303
71	0.3752	0.3639	0.3431	0.3658	0.3558	0.3432	0.3111
101	0.3903	0.3807	0.3582	0.3794	0.3635	0.3531	0.3124

CLAD-C

Seed	AFM	CCL-DC	MKD	FGH	aL-SAR	LPR	OCAR
11	0.5065	0.4970	0.4914	0.2995	0.2989	0.4611	0.3499
29	0.5001	0.4960	0.4765	0.3413	0.3136	0.4599	0.3490
47	0.5231	0.5035	0.4974	0.3062	0.3526	0.4826	0.4221
71	0.5216	0.4818	0.4867	0.2905	0.3300	0.4628	0.3419
101	0.5133	0.4820	0.5136	0.2913	0.3526	0.4804	0.4105

Table 42: Equal-weighted mean of the three benchmark-level primary scores. This descriptive summary complements, but does not replace, the dataset-specific primary analysis.

Method	Equal-weighted three-benchmark primary score
AFM	**0.3805**
CCL-DC	0.3556
MKD	0.3414
LPR	0.3273
FGH	0.3126
aL-SAR	0.2890
OCAR	0.2782

6.2.4 CLAD-C Pedestrian analysis

The full-assimilation coordinate has more peak-to-final forgetting than no protection on CLAD-C. Classwise checkpoint analysis localizes most of the difference to Pedestrian. As shown in Table 44, AFM $\eta = 1.00$ reaches substantially higher Pedestrian accuracy before the third chronological segment, after which both AFM and no protection collapse to zero held-out Pedestrian accuracy at the same boundary. The corresponding margin shift is negative for both learners and predictions are taken over by Car or Truck. The larger AFM forgetting value therefore reflects a higher acquired peak before a collapse shared with the unrestricted learner, not a collapse unique to AFM.

Finite protection and population performance coexist in a way consistent with the theorem. In individual seeds, active protected Pedestrian evidence remains present while held-out Pedestrian accuracy falls to zero. The theorem protects the declared finite evidence, not the complete unseen class distribution.

Table 43: Five-seed mean conventional continual-learning metrics for AFM and the six modern challengers. Fgt. denotes forgetting, BWT backward transfer, Test held-out final-test accuracy, and Online chronological stream accuracy.

Dataset	Method	$R\uparrow$	$P\uparrow$	Fgt.↓	BWT↑	Test↑	Online↑
CORe50	AFM 0.10	0.9648	0.2088	0.0352	-0.0186	0.2291	0.2125
	AFM 0.50	0.9185	0.2458	0.0815	-0.0658	0.2362	0.2594
	AFM 1.00	0.9020	0.2595	0.0980	-0.0842	0.2398	0.2764
	CCL-DC	0.9756	0.2066	0.0244	0.0339	0.2648	0.2240
	MKD	0.9952	0.1808	0.0048	0.0374	0.2415	0.1929
	FGH	0.8755	0.2910	0.1245	-0.1088	0.2493	0.3087
	aL-SAR	0.9862	0.1827	0.0138	0.0392	0.2338	0.1914
	LPR	0.9934	0.1651	0.0066	0.0273	0.2091	0.1698
	OCAR	0.9991	0.1437	0.0009	0.0023	0.1469	0.1474
CLEAR-10	AFM 0.10	0.9933	0.3750	0.0067	0.0056	0.3412	0.3753
	AFM 0.50	0.9828	0.3793	0.0172	-0.0013	0.3417	0.3799
	AFM 1.00	0.9768	0.3834	0.0232	-0.0065	0.3386	0.3835
	CCL-DC	0.9906	0.3766	0.0094	0.0102	0.3515	0.3769
	MKD	0.9931	0.3537	0.0069	0.0106	0.3337	0.3509
	FGH	0.9698	0.3891	0.0302	-0.0095	0.3407	0.3911
	aL-SAR	0.9918	0.3602	0.0082	0.0109	0.3350	0.3586
	LPR	0.9969	0.3496	0.0031	0.0167	0.3414	0.3486
	OCAR	0.9989	0.3168	0.0011	0.0020	0.3045	0.3173
CLAD-C	AFM 0.10	0.8534	0.5706	0.1466	-0.0081	0.5732	0.5713
	AFM 0.50	0.7915	0.5874	0.2085	0.0293	0.5440	0.6025
	AFM 1.00	0.7433	0.6040	0.2567	0.0382	0.5457	0.6155
	CCL-DC	0.8608	0.5716	0.1392	-0.0198	0.5685	0.5856
	MKD	0.9020	0.5467	0.0980	0.0044	0.5995	0.5703
	FGH	0.7694	0.3974	0.2306	0.0538	0.4872	0.4451
	aL-SAR	0.9095	0.3625	0.0905	0.0529	0.3113	0.4330
	LPR	0.8710	0.5388	0.1290	0.0165	0.5892	0.5717
	OCAR	0.9878	0.3793	0.0122	0.0106	0.3857	0.4477

Table 44: Exploratory CLAD-C Pedestrian diagnostic from saved checkpoints, reported as five-seed means unless stated otherwise. This diagnostic is not part of the primary benchmark analysis, and no retraining is involved.

Quantity	AFM $\eta = 1.00$	No protection
Pedestrian peak accuracy at boundary 2	77.24%	60.34%
Pedestrian accuracy at boundary 3	0.00%	0.00%
Final Pedestrian accuracy	1.58%	1.76%
Pedestrian peak-to-final forgetting	75.66%	58.59%
Mean Pedestrian margin at boundary 2	+0.482	+0.053
Mean Pedestrian margin at boundary 3	-4.202	-5.587
Mean boundary-2-to-3 margin change	-4.684	-5.640
Dominant wrong class at boundary 3	Car 3/5; Truck 2/5	Car 5/5

A separate exploratory inverse-frequency diagnostic tested whether aggregate post-prefix class frequency alone explained the full-assimilation penalty. Four paired seeds completed. The mean excess forgetting of AFM $\eta = 1.00$ over no protection increased from about 0.0269 in the primary benchmark runs to about 0.0837 under inverse-frequency weighting, giving a mean attenuation of approximately -0.0568 with paired

Table 45: Five-seed mean CORe50 metrics for full AFM at $\eta = 0.50$ and the targeted ablations. The final column is the single-operating-point $R \times P$ product, not the AFM family hypervolume.

Method	$R\uparrow$	$P\uparrow$	Fgt.↓	BWT↑	Online↑	Test↑	$R \times P\uparrow$
Full AFM, $\eta = 0.50$	0.9185	**0.2458**	0.0815	-0.0658	**0.2594**	**0.2362**	**0.2257**
Base-only, finite normalized budget	**0.9984**	0.1906	**0.0016**	-0.0005	0.1885	0.2007	0.1903
Base-only, fixed-absolute budget	0.9977	0.1908	0.0023	-0.0004	0.1886	0.2016	0.1903
Single-timescale AFM	0.9205	0.2452	0.0795	-0.0638	0.2589	0.2354	0.2257

bootstrap interval $[-0.1113, -0.0023]$. The weighting was itself extreme and degraded both learners. This result rejects the simple explanation based only on global class counts. It does not establish temporal class absence as the unique cause; proving that stronger causal claim would require a dedicated reordering or replay intervention.

6.2.5 Resource accounting

The modern comparison preserves method-native auxiliary state. In particular, CCL-DC carries two complete learners and replay, while MKD carries the student, a complete EMA teacher, and replay. AFM carries candidate, projector, metaplastic, event, and finite-shield state. Runtime and memory are therefore interpreted together with predictive performance rather than normalized away. The present AFM implementation has substantial runtime overhead, especially on CLAD-C, as reported in the main text.

6.3 Ablations, execution audit, and controlled mechanism tests

6.3.1 Transaction ablation

The component ablation is performed at the predeclared CORe50 coordinate $\eta = 0.50$. Because the normalized persistent update and finite residual are coupled inside the transaction, the experiment does not claim to isolate either component individually. Instead, full AFM is compared with two base-only protected transactions, one using the normalized finite budget and one using an absolute budget, together with a single-timescale AFM variant. Table 45 gives the conventional metrics.

The base-only variants obtain near-perfect retention by suppressing acquisition. Relative to the normalized base-only transaction, full AFM increases plasticity by approximately 0.0551, online accuracy by 0.0708, and held-out final-test accuracy by 0.0354; the corresponding paired intervals remain positive. The two base-only budget parameterizations are almost indistinguishable in $R \times P$, so the ablation supports the importance of completing the deployed AFM transaction beyond the persistent base step, not a claim that one budget parameterization alone explains the gain. The single-timescale result is nearly identical to full AFM at this coordinate, indicating that the observed CORe50 aggregate performance is not sensitive to the multiscale bank in this particular setting. This empirical equivalence does not conflict with the multiscale results in Section 2.9. Theorems 2.32 and 2.33 establish a worst-case representational separation when temporal relevance must be represented over a long age range with a broad, approximately scale-free profile; they do not imply that every finite stream induces such a profile or that the resulting approximation gap must alter downstream accuracy. A practical multiscale advantage is therefore expected when protected relevance spans sufficiently separated ages and the resulting allocation decisions are sensitive to those temporal weights. The present CORe50 coordinate does not provide empirical evidence that this regime is active.

6.3.2 Execution-level audit

Across the 45 protected real-world AFM runs, there are 677 certifications and commits and 41,954 nonzero protected base updates. Finite restoration is accepted 41,999 times out of 44,303 attempts; failed attempts are rejected rather than committed. The maximum accepted finite endpoint error remains below 10^{-6} on every dataset. The minimum realized-minus-requested path-fraction margins are at floating-point roundoff

Table 46: Invocation frequency and host-observed per-call cost of the principal AFM operations on CLAD-C at $\eta = 0.1$. Values are medians across the five frozen seeds. Host timing uses the low-perturbation profiling mode without per-region CUDA synchronization; CUDA-event timing is recorded separately. The final column reports the median share of profiled host-observed elapsed time.

Operation	Median calls	Median host-observed time/call	Host-timing share
Protected-projector construction	1,711	2861.552 ms	**37.89%**
Compact-cardinal shield construction	1,526	1611.751 ms	**20.58%**
Safe-base backtracking	1,665	613.933 ms	8.60%
Finite-address preparation	1,665	601.503 ms	8.08%
Protected prestate evaluation	1,711	599.454 ms	8.06%
Protected poststate evaluation	1,664	599.695 ms	8.05%
Frontier evaluation	25,665	16.670 ms	3.46%
Candidate fitting	37	4846.976 ms	1.86%
Sketch update	142	356.499 ms	0.47%
Post-restoration re-evaluation	1,526	28.977 ms	0.39%
Comparator endpoint/backtracking	1,665	15.096 ms	0.24%
Ordinary backward pass	1,711	4.122 ms	0.07%
Ordinary forward pass	1,711	3.930 ms	0.07%

scale, while the persistent-base progress ratio remains above the analytic theorem floor on every accepted protected update examined.

A tighter finite-smoothness diagnostic is deliberately more aggressive than the analytic theorem bound. On CLAD-C it exceeds the realized decrease in 11 of 22,932 accepted restoration events, with maximum ratio discrepancy 3.8×10^{-4} and maximum absolute loss-decrease discrepancy 9.1×10^{-6}. These events do not violate the analytic lower bound, whose margin stays positive. No corresponding negative tight-diagnostic events occur on CORe50 or CLEAR-10.

The routing audit also matches the declared information conditions. No protected CORe50 or CLEAR-10 run reports the calibration obstruction, whereas all protected CLAD-C runs do. The implementation therefore withholds the positive route-identification conclusion on CLAD-C. No evaluator-side segment information is introduced.

6.3.3 Runtime bottleneck analysis

The component-level decomposition in Table 37 identifies protection geometry, finite counterfactual completion, and endpoint verification as the principal runtime categories. CLAD-C provides the clearest operation-level decomposition because it also has the largest overall runtime overhead. Table 46 therefore separates invocation frequency from per-call cost for this protocol.

Protected-projector construction is the largest individual component of the CLAD-C host-timing profile, accounting for 37.89% of measured host-observed time. It combines a median of 1,711 invocations with an approximately 2.862 s host-observed duration per invocation. Compact-cardinal shield construction is the second largest component, accounting for 20.58% across 1,526 median invocations at approximately 1.612 s per call. Safe-base backtracking, finite-address preparation, and protected prestate and poststate evaluations each account for approximately 8-9% of the host-timing profile.

The invocation analysis distinguishes operations that are individually expensive in the host-timing profile from operations that dominate aggregate execution. Candidate fitting has the largest median host-observed duration per invocation among the listed operations, approximately 4.847 s, but occurs only 37 times and therefore contributes 1.86%. Frontier evaluation occurs 25,665 times but has a median host-observed duration of only 16.67 ms per call and contributes 3.46%. Protected-projector construction is therefore the largest host-visible implementation bottleneck because substantial per-call duration is combined with repeated invocation.

Table 47: Controlled AFM mechanism tests on three seeds. These experiments test recurrence, conflict, route refinement, capacity pressure, and observational indistinguishability rather than external predictive performance.

Scenario	Seeds passed	Structural events	Observed behavior
Favourable recurrence	3/3	commits 6/6/6; protected steps 20/13/10	recurrent protected learning completed
Semantic conflict	3/3	reopenings 2/3/1; splits 0/0/0	conflict handled without an unsupported route split
Observable shift	3/3	commits 3/3/3; protected steps 41/27/27	observable-shift control completed
Within-route observable shift	3/3	splits 1/1/1	positive route splitting exercised on every seed
Capacity pressure	3/3	commits 11/11/11; protected steps 30/18/21	bounded-capacity control completed
Identical-observation impossibility	3/3	commits 3/3/3; protected steps 12/8/11	indistinguishable observations did not induce a false semantic resolution

These host-timing measurements identify protected-projector construction, compact-cardinal shield construction, finite-address preparation, and protected-state evaluation as the principal host-visible targets for implementation optimization. Such optimization can focus on reducing repeated construction and reevaluation costs without changing the same-state comparator, persistent-assimilation rule, or finite endpoint constraints that define the AFM transaction. The separately recorded CUDA-event timings provide the corresponding device-side diagnostic.

Profiling validity. The profiling instrumentation does not modify the AFM learning rule, frozen hyperparameters, prepared streams, or operating-point selection. The default profiler uses low-perturbation host timing and deliberately avoids per-region CUDA synchronization; CUDA-event elapsed times are recorded separately. This design reduces synchronization-induced perturbation while requiring the host-region percentages to be interpreted as a decomposition of host-observed execution rather than synchronized device execution. CORe50 and CLEAR-10 reproduce the corresponding unprofiled metrics exactly. CLAD-C uses the same AFM configuration and byte-identical prepared training and evaluation streams, but its frozen execution mode is nondeterministic. Independent unprofiled CLAD-C executions with the same nominal seed and AFM settings likewise produce distinct numerical trajectories. The profiling results therefore characterize execution of the frozen AFM algorithm rather than reproduction of one particular numerical realization.

6.3.4 Controlled mechanism tests

The controlled experiments exercise recurrence, semantic conflict, observable route change, capacity pressure, and observational indistinguishability. Table 47 summarizes the results. The within-route observable-shift condition produces a route split on every seed. Semantic conflict produces reopening but no unsupported split, and the identical-observation condition remains unresolved, consistent with the absence of learner-observable distinguishing information.

The transfer-conflict diagnostic did not instantiate the required theorem precondition: no transfer attempt occurred on any of its three seeds, so the designed zero-dimensional feasible subspace was not exercised. It is therefore non-informative for the transfer theorem and is reported only to delimit the empirical coverage of the mechanism tests.

Table 48: AFM transaction ablation and execution validation

Retention-plasticity summary

Method	R	P	$R \times P$
Full AFM, $\eta = 0.50$	0.9185	**0.2458**	**0.2257**
Base-only, normalized budget	**0.9984**	0.1906	0.1903
Base-only, fixed absolute budget	0.9977	0.1908	0.1903
Single-timescale AFM	0.9205	0.2452	0.2257

Forgetting and predictive metrics

Method	Forgetting	BWT	Online	Test
Full AFM, $\eta = 0.50$	0.0815	-0.0658	**0.2594**	**0.2362**
Base-only, normalized budget	**0.0016**	-0.0005	0.1885	0.2007
Base-only, fixed absolute budget	0.0023	-0.0004	0.1886	0.2016
Single-timescale AFM	0.0795	-0.0638	0.2589	0.2354

Execution-level quantity across 45 protected benchmark runs	Recorded value
Certifications and commits	677
Nonzero protected base updates	41,954
Finite restorations accepted / attempted	41,999 / 44,303
Maximum accepted finite endpoint error	$< 10^{-6}$ on every dataset
Accepted protected updates below analytic theorem floor	0
CLAD-C tighter finite-smoothness diagnostic exceedances	11 / 22,932 accepted restoration events

The ablation uses $n = 5$ frozen CORe50 seed trajectories at the predeclared $\eta = 0.50$ coordinate. Base-only variants attain near-perfect retention by suppressing acquisition; full AFM increases plasticity, online accuracy and held-out final-test accuracy. The single-timescale variant is nearly identical on this particular finite stream, which does not contradict the worst-case multiscale representation theorem. Failed finite-restoration attempts are rejected rather than committed. Runtime decomposition and controlled mechanism tests are reported earlier in this section.

7 Discussion

AFM changes the unit at which the stability-plasticity trade-off is quantified. The protected learner is not compared with a nominal unconstrained direction but with the actual same-state endpoint that the declared no-protection learner would accept. This creates a meaningful denominator for persistent plasticity. The resulting guarantee separates the requested assimilation coordinate η_t from the compatibility fraction κ_t: a large η_t does not imply that an incompatible gradient can be stored persistently, and the theorem exposes rather than conceals that loss of compatibility.

The second conceptual distinction is between persistent learning and finite functional completion. Exact restoration of protected outputs and exact reproduction of the current comparator endpoint are useful deployed properties, but they do not imply that the unrestricted parameter trajectory has been stored. Outside the finite support of the residual, future behavior is governed by the protected base. This separation is also what makes the empirical ablation interpretable: a base-only learner can preserve almost everything by suppressing acquisition, whereas the complete transaction recovers substantially more plasticity and predictive performance while the finite residual remains distinct from persistent base learning.

The finite/population distinction is equally important. AFM gives exact finite protection for declared evidence and bounded certificates for broader behavior maps when the corresponding assumptions are available. The CLAD-C Pedestrian result shows why these claims must remain separate. A finite set of protected Pedestrian

observations can be preserved exactly while the unseen Pedestrian population is overtaken by classifier interference. Treating finite interpolation as population retention would therefore overstate the theorem.

Task-free operation also has an information-theoretic boundary. AFM can route exactly under observable separation and can control false route refinement with separately allocated sequential evidence, but no learner can identify a semantic distinction that induces the same observation law. AFM therefore withholds the positive routing claim when observable separation is insufficient and does not use evaluator boundaries as learner-visible information.

The main practical limitation is computational overhead. The profiling analysis in Table 37 localizes the host-observed execution cost primarily to protection geometry, finite counterfactual completion, and endpoint verification rather than ordinary learning or construction of the same-state comparator. The detailed CLAD-C analysis in Section 6.3 further identifies repeated protected-projector construction and compact-cardinal shield construction as the two largest individual components of the host-timing profile. These results provide concrete optimization targets, including reducing redundant projector construction, finite-address preparation, shield construction, and protected-state reevaluation. Such optimizations must preserve the same-state comparator and the distinction between persistent learning and finite functional completion. Reducing these implementation costs while preserving the certified AFM transaction is therefore an important direction for future work. The bounded-resource theorem concerns structural state and explicit precision, not low wall-clock cost.

Finally, the convex dynamic-regret specialization and the nonconvex neural experiments have different status. The former supplies a global tracking statement under a supplied exact convex projection oracle; the latter are empirical executions of the complete endpoint transaction with numerical verification, not a claim that the real neural network satisfies all outward-certified convex or interval assumptions. Keeping these levels separate is necessary for the theorem and experiments to reinforce rather than overstate one another.

These experiments identify functional compatibility as an experimentally controllable property of incoming learning, not merely a retrospective description of forgetting. The finding reframes the stability-plasticity problem. A neural state does not offer a single undifferentiated capacity for learning: some incoming functional changes are more compatible with protected behaviour than others, and learning rules differ in how effectively they capture that opportunity. This perspective complements replay, regularization, gradient projection and plasticity-preservation approaches rather than replacing them.(Schwarz et al., 2018; Dohare et al., 2024; Störk, 2026; Abbes et al., 2026)

The result also defines its own boundary. Compatibility does not determine persistent learning independently of retention allowance, method or finite curvature; ρ is not a universal function of κ; exact finite output restoration is not population-level retention; and the present fixed-norm construction does not establish a four-level causal continuum at ordinary full update magnitude. The near-zero compatibility assessment similarly showed that 65 negative theorem-aligned empirical margins occurred where the required finite curvature or step condition had not been certified, rather than constituting certified theorem violations (Table 16). These qualifications are part of the result: they distinguish a causal local learning geometry from the nonlinear regime in which that geometry can no longer be realized by the same matched intervention.

Together, the theory, intervention and constructive mechanism support a concise principle: functional compatibility controls the local persistent-learning frontier, retention constraints determine how much of that frontier can be used, and nonlinear geometry determines how far it extends. This offers a measurable way to ask not only how an artificial learner should protect the past, but which components of new learning can safely become part of its persistent future.

Acknowledgements

The author thanks the ADAPT Centre for access to high-performance computing resources used for the experimental runs. The ADAPT Centre had no role in conceptualization, mathematical or methodological development, software development, study design, data analysis, interpretation or manuscript preparation.

Funding

The author received no specific funding for this work.

Author contributions

H.J. conceived the study, developed the theory and methodology, implemented the software, designed and ran the experiments, performed the formal and statistical analyses, validated the results, and wrote and revised the manuscript.

Competing interests

The author declares no competing interests.

Data availability

All datasets used in this study are publicly available from the original sources cited in this paper. No new primary dataset was generated. The public benchmark datasets are not redistributed with this submission. Numerical summaries underlying the reported figures and tables are provided throughout this paper. Saved checkpoints and row-level derived experimental outputs required for additional verification are available from the corresponding author on reasonable request. The released analysis and experiment code described under Code availability provides the procedures and configurations used to reproduce the derived results.

Code availability

The AFM implementation, causal-compatibility and follow-up experiment code, configuration files, analysis scripts and reproduction instructions are publicly available at `https://github.com/hosseinjavidnia/afm/`.

References

Istabrak Abbes, Gopeshh Subbaraj, Matthew Riemer, Nizar Islah, Tsuguchika Tabaru, Hiroaki Kingetsu, Sarath Chandar, and Irina Rish. Revisiting replay and gradient alignment for continual pre-training of large language models. In *Proceedings of The 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 465–486. PMLR, 2026.

Ari Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019.

Marcus K. Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19:1697–1706, 2016. doi: 10.1038/nn.4401.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016. doi: 10.1137/15M1009718.

Tieliang Gong, Zhongbo Zhang, Wen Wen, and Yong-Jin Liu. Drift and dependence: Layer-wise information-theoretic bounds for replay-based continual learning. *arXiv preprint*, 2026.

Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. SODA10M: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint*, 2021.

Sooyong Jang, Seongjun Park, Insup Lee, and Osbert Bastani. Sequential covariate shift detection using classifier two-sample tests. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9845–9880, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Chayanon Kitkana and Shivam Arora. Sustained gradient alignment mediates subliminal learning in a multi-step setting: Evidence from mnist auxiliary logit distillation experiment. *arXiv preprint arXiv:2604.25779*, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Daniel F. Leite. Task-free continual learning with expansion-based granular cnn: Gradual partitioning of manifolds in image stream classification. *Neurocomputing*, 671:132665, 2026. doi: 10.1016/j.neucom.2026.132665.

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The CLEAR benchmark: Continual learning on real-world imagery. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Sihao Liu, Yibo Yang, Xiaojie Li, David A. Clifton, and Bernard Ghanem. Enhancing online continual learning with plug-and-play state space model and class-conditional mixture of discretization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20502–20511, 2025.

Vincenzo Lomonaco and Davide Maltoni. CORe50: A new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 17–26, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aojun Lu, Hangjie Yuan, Tao Feng, and Yanan Sun. Rethinking the stability-plasticity trade-off in continual learning from an architectural perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40888–40902, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Nicolas Michel, Maorong Wang, Ling Xiao, and Toshihiko Yamasaki. Rethinking momentum knowledge distillation in online continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35607–35622, 2024.

Nicolas Michel, Maorong Wang, Jiangpeng He, and Toshihiko Yamasaki. From offline to online memory-free and task-free continual learning via fine-grained hypergradients. *Transactions on Machine Learning Research*, 2026. URL `https://openreview.net/forum?id=Tu9rHiczLm`.

Seoyoung Park, Haemin Lee, and Hankook Lee. Is prompt selection necessary for task-free online continual learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7883–7892, 2026.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

Seyed Roozbeh Razavi Rohani, Khashayar Khajavi, Wesley Chung, Mo Chen, and Sharan Vaswani. Preserving plasticity in continual learning with adaptive linearity injection. In *Proceedings of the 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 418–444, 2026.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, 2018.

Minhyuk Seo, Hyunseo Koh, and Jonghyun Choi. Budgeted online continual learning by adaptive layer freezing and frequency-based sampling. In *International Conference on Learning Representations*, 2025.

Yao Shu, Jian Mu, and Zhongxiang Dai. Why zeroth-order adaptation may forget less: A randomized shaping theory. *arXiv preprint*, 2026.

Julius Störk. Interference and retention in continual learning. *arXiv preprint*, 2026.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Edoardo Urettini and Antonio Carta. Online curvature-aware replay: Leveraging second-order information for online continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60590–60609, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Eli Verwimp, Kai Yang, Sarah Parisot, Lanqing Hong, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. CLAD: A realistic continual learning benchmark for autonomous driving. *Neural Networks*, 161:659–669, 2023. doi: 10.1016/j.neunet.2023.02.001.

Maorong Wang, Nicolas Michel, Ling Xiao, and Toshihiko Yamasaki. Improving plasticity in online continual learning via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23460–23469, 2024.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Xiwen Wei, Guihong Li, and Radu Marculescu. Online-lora: Task-free online continual learning via low rank adaptation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 6634–6645, 2025.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 2016.

Fei Ye and Adrian G. Bors. Online task-free continual learning via dynamic expansionable memory distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20512–20522, 2025.

Murat Onur Yildirim, Elif Ceren Gok Yildirim, Decebal Constantin Mocanu, and Joaquin Vanschoren. Self-regulated neurogenesis for online data-incremental learning. In *Proceedings of the 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 657–671, 2026.

Jinsoo Yoo, Yunpeng Liu, Frank Wood, and Geoff Pleiss. Layerwise proximal replay: A proximal point method for online continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57199–57216, 2024.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.