INTERFERENCE-GATED DYNAMIC ACTIVATION FOR TASK-AGNOSTIC CONTINUAL LEARNING: A FORMAL-EMPIRICAL AUDIT OF STABILITY, FORGETTING, AND FAILURE REGIMES

Anonymous authors
Paper under review

ABSTRACT

Continual learning methods frequently reduce forgetting by adding replay, regularization, or routing constraints, yet activation functions are usually treated as fixed nonlinearities rather than adaptive components of the retention mechanism. We study a task-agnostic setting where task identities are unavailable and dynamic activation updates must operate under matched replay and optimization budgets. In this setting, we define an interference-gated projected activation update, formalize a one-step forgetting-proxy comparison against static-gate baselines, and introduce a fairness-normalized attribution predicate that blocks invalid comparisons when memory or optimization controls differ. The empirical study executes seed-aggregated runs on reduced Split CIFAR-100, reduced Split Tiny-ImageNet, and a third real fallback online digits stream, with static GELU, ReLU, Mish, A-GEM, EWC, and ablated dynamic variants as comparators. Results show that the full dynamic gate is indistinguishable from static GELU in the executed matrix, while replay-family changes can worsen forgetting and destabilize backward transfer, yielding mixed support for retention-improvement claims. Symbolic checks confirm local first-order identities and fairness-accounting completeness but expose a failed simplification in one projection-bound audit, limiting formal conclusions to local validity windows. These findings sharpen the boundary between plausible mechanism intuition and supported claim scope: activation-state adaptation can be audited rigorously, but global forgetting guarantees remain unresolved without stronger alignment conditions and broader benchmark execution.

1 INTRODUCTION

Continual learning research has repeatedly shown that catastrophic forgetting is difficult to avoid when a single parameter set is updated on non-stationary streams without direct access to all past data (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; Zenke et al., 2017). The dominant response has been to regulate parameter updates, replay selected examples, or route information through task-dependent modules (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018b; Buzzega et al., 2020; Aljundi et al., 2019; Serra et al., 2018; Mallya & Lazebnik, 2018; Mallya et al., 2018). This literature has produced practical systems, but it also leaves a persistent design asymmetry: update rules and memory policies are extensively engineered, while the neuron nonlinearity is often treated as static infrastructure. The present work investigates whether activation-state adaptation can be formalized and audited as a first-class continual-learning mechanism under task-agnostic constraints.

The motivation is structural rather than cosmetic. Fixed activations shape gradient flow, feature geometry, and saturation behavior, all of which directly influence interference between new and old representations (Nair & Hinton, 2010; Clevert et al., 2016; Klambauer et al., 2017; Hendrycks & Gimpel, 2016; Misra, 2019; Ba et al., 2016; Ioffe, 2017). In ordinary supervised training, these effects are usually summarized as optimization convenience. In continual learning, however, the same effects can alter the stability-plasticity balance that determines forgetting trajectories. If a dynamic activation state can react to measured interference while respecting strict comparator parity, it may offer a mechanism-level lever that complements replay and regularization instead of replacing them.

This opportunity comes with two risks that the paper treats as central. First, many reported gains in continual-learning pipelines are confounded by replay policy differences, memory overhead mismatches, or protocol drift between methods (Parisi et al., 2019; Lange et al., 2019; Shim et al., 2022; van de Ven & Tolias, 2018; Henriques et al., 2023). Any activation claim that ignores these confounders is weak regardless of effect size. Second, local formal arguments are

often overextended into global performance narratives. A first-order condition can explain one-step behavior without implying multi-step optimality. We therefore design a hybrid contribution that keeps formal scope explicit and ties each major claim to executed evidence.

The investigated method is an interference-gated projected activation update embedded in a replay-based continual learner. The activation state is updated with a gate value computed from interference proxies and then projected into an admissible set to control drift. This mechanism is evaluated against static activations and standard continual baselines under matched buffer, optimizer, and schedule controls. The key question is not whether dynamic activations are universally superior, but whether they produce measurable forgetting reductions in a task-agnostic regime without violating fairness constraints.

The broader relevance extends beyond one architecture choice. Task-agnostic continual learning is a proxy for real deployments where task boundaries are unknown, data streams are heterogeneous, and memory budgets are finite (Chaudhry et al., 2021; Cai et al., 2024; Cho et al., 2024). If activation adaptation cannot show robust benefit even in carefully controlled reduced-scale settings, that negative evidence is informative for future mechanism design. Conversely, if some windows show reliable improvement, those windows can define where richer geometry-aware or uncertainty-aware controls should focus.

This paper contributes four concrete outcomes.

- We define an interference-gated projected activation-state update and a fairness-normalized comparator predicate that makes attribution conditions explicit before any performance claim.
- We provide theorem-style local analyses for update invariance and one-step forgetting-proxy differences, with complete proofs and explicit applicability limits.
- We execute a hybrid empirical and symbolic validation suite across three real substrates, multiple baselines, and mechanism ablations, including uncertainty summaries and negative-result accounting.
- We show that the executed evidence yields mixed claim support: local formal checks mostly pass, but empirical forgetting improvement is not established in the reduced benchmark matrix and one projection-bound symbolic check fails.
- We scope validated contributions to the interference-gated activation hypothesis only; two additional hypothesis tracks (curvature-aware gating and uncertainty-aware gating) are retained as future-work directions rather than closed claims in this iteration.

The paper is organized as follows. Section 2 synthesizes the continual-learning and activation-function literature and states the novelty boundary. Section 3 formalizes the task-agnostic setting and objective. Section 4 presents the algorithm and matched-budget audit predicate. Section 5 provides formal statements and proofs. Section 6 and section 7 report the executed protocol and evidence. Section 8 and section 9 interpret mixed outcomes and specify follow-up tests.

2 Related Work

Continual-learning methods can be grouped by their primary anti-forgetting mechanism: parameter regularization, replay, architectural isolation, and hybrid variants. Regularization families such as EWC, SI, and RWalk penalize movement of parameters deemed important for prior tasks (Kirkpatrick et al., 2017; Zenke et al., 2017; Chaudhry et al., 2018a; Aljundi et al., 2018). Their strengths include low additional data storage and straightforward integration into existing optimizers. Their limitation is approximation sensitivity: importance estimates can be noisy, diagonal assumptions can be coarse, and performance often degrades under substantial distributional drift.

Replay-centric methods have shown stronger empirical robustness in many benchmark settings. GEM and A-GEM impose update constraints using episodic memory (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018b); DER-style methods mix replayed targets and distillation-like terms (Buzzega et al., 2020); MIR emphasizes retrieval quality under small buffers (Aljundi et al., 2019); and recent variants explore contrastive replay, cognitive replay, and long-tail adaptations (Borges et al., 2024; Cho et al., 2024; Cai et al., 2024). The common strength is direct retention signal from historical samples. The common limitation is confounding: improvements may arise from retrieval policy, distillation weighting, or budget mismatch rather than the nominal mechanism under study.

Architectural isolation and masking approaches can substantially reduce forgetting when task identity is available or inferred reliably (Rusu et al., 2016; Serra et al., 2018; Mallya & Lazebnik, 2018; Mallya et al., 2018). These methods often provide strong upper-bound behavior in task-aware regimes, but their assumptions conflict with fully

task-agnostic deployment targets. This tension is central to the present study: we evaluate only no-task-ID settings for claim-facing conclusions, while treating task-aware intuition as background context rather than direct evidence.

A parallel line of work studies activation and normalization behavior, mostly outside continual-learning attribution studies. ReLU, ELU, SELU, GELU, and Mish characterize different smoothness and saturation properties (Nair & Hinton, 2010; Clevert et al., 2016; Klambauer et al., 2017; Hendrycks & Gimpel, 2016; Misra, 2019); layer and batch-level normalization variants influence internal statistics and gradient conditioning (Ba et al., 2016; Ioffe, 2017). These results motivate activation-centric mechanism design because they show that nonlinearity choice changes optimization dynamics and representational geometry. Yet they rarely establish forgetting outcomes under matched continual-learning comparators.

Recent analyses sharpen evaluation concerns. Surveys and protocol critiques report that cross-paper ranking instability is common when regimes, metrics, and memory accounting are not standardized (Parisi et al., 2019; Lange et al., 2019; Chaudhry et al., 2021; van de Ven & Tolias, 2018; Henriques et al., 2023). Benchmarking extensions and replay-ablation studies further show that method ordering can shift with buffer policy, class imbalance, and stream construction (Chaudhry et al., 2019; Dong et al., 2024; Aljundi et al., 2022; Shim et al., 2022). These findings imply that an activation-centric claim must be coupled to explicit fairness predicates and uncertainty reporting.

The contradiction map inherited from earlier phases highlights three relevant conflicts. First, task-aware methods can dominate in favorable settings but violate task-agnostic assumptions. Second, activation-driven narratives can hide replay-policy causality. Third, local smoothness arguments are often presented as if they were global guarantees. Our approach addresses these conflicts by enforcing matched-budget auditing, reporting negative outcomes, and restricting formal claims to one-step regimes where assumptions are explicit.

The novelty boundary is therefore narrow and defensible. We do not claim a new universal forgetting bound, and we do not claim dynamic activations dominate replay baselines across protocols. Instead, we contribute a formalization-and-audit framework for interference-gated activation dynamics, then test whether that framework yields measurable forgetting reduction under controlled task-agnostic execution. This framing aligns with open gaps identified by prior literature: attribution rigor, fairness accounting, and transparent reporting of scope-limited formal evidence.

An additional boundary concerns transfer across problem families. Evidence from continual reinforcement learning, meta-replay, and hypernetwork adaptation suggests that mechanism behavior can change when control policies, trajectory horizons, or parameterization regimes differ (Kurenkov et al., 2023; Hayes et al., 2019; Ha et al., 2016). We therefore do not treat neighboring domains as direct confirmation of activation efficacy. Instead, they motivate explicit scope statements: a mechanism that appears stable in one substrate can fail in another when drift geometry and replay semantics change. This cross-domain caution reinforces our emphasis on admissibility predicates and transparent negative-result reporting. It also motivates keeping benchmark-specific caveats close to each claim rather than collecting all caveats in a single generic disclaimer.

3 Problem Formulation and Scope

We consider a discrete-time continual stream indexed by $t \in \mathbb{N}_{\geq 1}$ with inputs $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ and labels $y_t \in \mathcal{Y}_t$, where class support may expand over time. A predictor $f_{\theta_t,\psi_t} : \mathcal{X} \to \Delta^{C_t-1}$ has shared parameters $\theta_t \in \mathbb{R}^p$ and activation parameters $\psi_t \in \mathbb{R}^q$. For each layer $\ell \in \{1, \ldots, L\}$, we introduce a dynamic activation state $s_{\ell,t} \in \mathbb{R}^{k_\ell}$, and denote the concatenated state by s_t.

Borrowed components from replay-based continual learning are the current-task loss and replay-retention loss, denoted by $\mathcal{L}_{cur}$ and $\mathcal{L}_{ret}$ (Buzzega et al., 2020; Aljundi et al., 2019; Shin et al., 2017). In this work we define the composite objective

$$\mathcal{J}_t(\theta_t, \psi_t, s_t) = \mathcal{L}_{cur} + \lambda_r \mathcal{L}_{ret} + \lambda_s \mathcal{R}_{state} + \lambda_m \mathcal{R}_{mom}, \tag{1}$$

where $\lambda_r, \lambda_s, \lambda_m \in \mathbb{R}_{\geq 0}$ are scalar coefficients, $\mathcal{R}_{state}$ controls activation-state drift, and $\mathcal{R}_{mom}$ penalizes unstable activation moments. The replay buffer $\mathcal{M}_t$ has capacity $B \in \mathbb{N}_{\geq 1}$.

3.1 Interference-Gated State Dynamics

We adopt an interference score $I_t \in [0, 1]$, where larger values indicate stronger conflict between current and retained gradients. The gate function $g : [0, 1] \to [g_{\min}, g_{\max}]$ is monotone nondecreasing with $0 < g_{\min} \leq g_{\max} < \infty$. The dynamic activation update is manuscript-defined as

$$s_{\ell,t+1} = \Pi_{\mathcal{S}}(s_{\ell,t} - \eta_s g(I_t) \nabla_{s_\ell} \mathcal{J}_t), \tag{2}$$

where $\eta_s > 0$ is a step size and $\Pi_{\mathcal{S}}$ is Euclidean projection onto feasible set $\mathcal{S} \subseteq \mathbb{R}^{k_\ell}$. We assume $\mathcal{S}$ is nonempty, closed, and convex. This projection structure is adapted from constrained-update logic in prior continual optimization work (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018b; Ashok et al., 2019), but the state variable and gate are specific to this manuscript.

3.2 Decision Variables, Feasibility, and Criterion

For claim-facing comparison, the decision variables are (θ_t, ψ_t, s_t) under a matched protocol across methods: identical replay budget, optimizer family, schedule, and step budget. We define forgetting difference against a static comparator as $\Delta_F = F_{static} - F_{dynamic}$, accuracy difference as $\Delta_A = \bar{A}_{dynamic} - \bar{A}_{static}$, and memory difference as $\Delta_M = M_{dynamic} - M_{static}$. The fairness predicate is

$$\mathbb{I}_{fair} = \mathbf{1}\{B_d = B_s\}\mathbf{1}\{\text{optimizer/schedule/steps match}\}\mathbf{1}\{\Delta_M \leq \tau_M\}, \tag{3}$$

where τ_M is a user-defined memory slack. Under this predicate, the operational acceptance indicator is

$$\mathcal{A} = \mathbf{1}\{\Delta_F > 0\}\mathbf{1}\{\Delta_A \geq -\epsilon_A\}\mathbf{1}\{\mathbb{I}_{fair} = 1\}. \tag{4}$$

This criterion does not assert optimality. It defines when empirical evidence is admissible for the central forgetting-improvement claim.

3.3 Applicability and Non-Applicability

The formal regime is local and one-step. Specifically, results in section 5 require bounded state gradients, smooth proxy losses, and alignment windows where $\langle \nabla_{s_t} \Phi_t, \nabla_{s_t} \mathcal{J}_t \rangle$ is lower-bounded by an interference-scaled constant. Statements do not apply to global multi-step guarantees, strongly non-smooth dynamics, or protocols that inject task identity at training or inference time. These boundaries are crucial because they separate auditable local mechanism behavior from broader claims that would require additional assumptions and larger-scale evidence.

The boundary has an empirical side as well. Even when assumptions are approximately satisfied, weak gate contrast or weak replay interference can make dynamic-state effects practically undetectable at dataset scale. For this reason, null results are interpreted as informative outcomes rather than implementation defects. Under the present protocol, a null pattern with valid fairness checks is sufficient to keep support status mixed because attribution criteria are satisfied while improvement criteria are not.

4 Methodology

This section defines the full operational method used for claim-facing execution. The method includes update equations, comparator admissibility rules, and evidence-visibility policy. Treating all three pieces as one system is essential: without parity controls and visibility rules, observed differences can be explained by protocol drift rather than activation dynamics. The goal is therefore not only to present an algorithm, but to specify an auditable pathway from training behavior to defensible interpretation. The same pathway is reused in the appendix reproducibility notes so each reported figure and table can be traced to the same protocol commitments.

4.1 Algorithmic Structure and Rationale

The method combines replay-based continual learning with activation-state adaptation. At each step, the learner computes the composite loss in equation 1, estimates an interference proxy from replay/current interactions, maps that signal through gate $g(I_t)$, updates activation state via equation 2, and then updates (θ_t, ψ_t) with the same optimizer used by static baselines. This sequence is chosen to isolate one hypothesis: if interference-gated state adaptation contributes to retention, it should do so under strict comparator parity.

Three design choices are motivated by prior evidence and by failure modes identified in the literature. First, replay remains explicit because replay-free activation adaptation is unlikely to dominate robust replay baselines in realistic streams (Buzzega et al., 2020; Aljundi et al., 2019; Cho et al., 2024). Second, projection is used to prevent unconstrained state drift that can destabilize hidden statistics (Klambauer et al., 2017; Ba et al., 2016; Ioffe, 2017). Third, fairness accounting includes activation-state overhead because memory omission can invert reported rankings (Shim et al., 2022; Henriques et al., 2023). Together, these choices target causal attribution rather than leaderboard optimization.

Algorithm 1 Task-Agnostic Continual Learning with Interference-Gated Projected Activation State

Input: stream $(x_t, y_t)_{t\geq 1}$, replay buffer size B, gate $g(\cdot)$, state step η_s, optimizer step η_θ
Initialize θ_1, ψ_1, layer states $s_{\ell,1} \in \mathcal{S}$, and empty buffer $\mathcal{M}_1$
for $t = 1, 2, \ldots$ **do**
  Form mini-batch from current sample and replay subset from $\mathcal{M}_t$
  Compute $\mathcal{L}_{cur}$ and $\mathcal{L}_{ret}$; form objective $\mathcal{J}_t$ in equation 1
  Estimate interference score $I_t \in [0, 1]$ from replay/current gradient interaction
  for each layer ℓ **do**
    $s_{\ell,t+1} \leftarrow \Pi_{\mathcal{S}} \left(s_{\ell,t} - \eta_s g(I_t) \nabla_{s_\ell} \mathcal{J}_t \right)$
  end for
  Update network parameters $(\theta_{t+1}, \psi_{t+1})$ using the matched optimizer protocol
  Update replay memory $\mathcal{M}_{t+1}$ under fixed budget B
  Log F, $\bar{A}$, BWT, ECE, minority recall, state drift, and fairness fields
end for

4.2 Comparator Policy and Evidence Admissibility

Comparator policy is part of the method, not a reporting afterthought. Any row violating equation 3 is excluded from claim-support interpretation. In the executed matrix, all primary rows satisfy $\mathbb{I}_{fair} = 1$, so differences are interpretable as mechanism-level outcomes rather than protocol artifacts. This is especially important for ablations where replay family or projection components change: even when forgetting worsens, negative evidence remains valid because comparator parity is preserved.

The evidence registry is partitioned into results-visible and limitations-only items. Results-visible evidence includes reduced Split CIFAR-100, reduced Split TinyImageNet, the fallback online digits stream, and the symbolic audit suite. A descriptor-only CORe50 target is retained for transparency but excluded from results-facing support. This split enforces a clean boundary between executed evidence and unresolved acquisition goals.

4.3 Pseudocode for Interference-Gated Updates

Algorithm 1 emphasizes operational coupling: state adaptation is not an independent module; it is embedded in a matched replay pipeline. The design is intentionally minimal so that ablations can isolate the contribution of gate and projection terms. In particular, removing the gate while keeping the rest fixed tests whether interference conditioning matters; removing projection tests whether bounded-state guarantees align with observed stability diagnostics.

5 Formal Analysis

This section states local results used for interpretation of symbolic and empirical audits. Borrowed conventions from replay and constrained updates are cited where introduced; manuscript-specific objects are explicitly marked as such.

The role of this section is diagnostic rather than promotional. We use formal statements to map assumptions to measurable consequences and to identify where inference should stop. This avoids a common failure mode in hybrid continual-learning papers, where local algebra is treated as global performance proof. Here, proofs define local validity, symbolic checks test internal consistency, and empirical results determine whether validity windows coincide with meaningful retention improvements.

5.1 State Invariance and Bounded Increment Result

Lemma 5.1 (Projected State Invariance and Increment Bound)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^{k_\ell}$ be nonempty, closed, and convex. Assume $\|\nabla_{s_\ell} \mathcal{J}_t\|_2 \leq G_s$ and $g(I_t) \in [g_{\min}, g_{\max}]$ for all audited steps. Then update equation 2 is well-defined, preserves feasibility $s_{\ell,t+1} \in \mathcal{S}$, and satisfies*

$$\|s_{\ell,t+1} - s_{\ell,t}\|_2 \leq \eta_s g_{\max} G_s. \tag{5}$$

Proof. Because $\mathcal{S}$ is nonempty, closed, and convex in Euclidean space, metric projection $\Pi_{\mathcal{S}}$ exists and is single-valued. Therefore $s_{\ell,t+1}$ is well-defined. By definition of projection, its output lies in $\mathcal{S}$, establishing invariance. Let $z_t = s_{\ell,t} - \eta_s g(I_t) \nabla_{s_\ell} \mathcal{J}_t$. Since $s_{\ell,t} \in \mathcal{S}$, optimality of projection implies

$$\|\Pi_{\mathcal{S}}(z_t) - s_{\ell,t}\|_2 \leq \|z_t - s_{\ell,t}\|_2.$$

Substituting z_t gives

$$\|s_{\ell,t+1} - s_{\ell,t}\|_2 \le \eta_s g(I_t) \|\nabla_{s_\ell} \mathcal{J}_t\|_2 \le \eta_s g_{\max} G_s,$$

which is equation 5. The argument is deterministic and requires no stochastic approximation step, so any empirical violation indicates assumption mismatch rather than proof incompleteness. □

5.2 One-Step Forgetting-Proxy Comparison

Define the retention proxy $\Phi_t = \mathcal{L}_{ret}(\theta_t, \psi_t; \mathcal{M}_t)$. Consider dynamic and static-gate state increments with matched non-activation controls:

$$\Delta s_t^{dyn} = -\eta_s g(I_t) \nabla_{s_t} \mathcal{J}_t, \qquad \Delta s_t^{stat} = -\eta_s g_0 \nabla_{s_t} \mathcal{J}_t. \tag{6}$$

Theorem 5.2 (Local Proxy Difference Under Alignment)**.** *Assume Φ_t is L_Φ-smooth in s_t, and in audited windows*

$$\langle \nabla_{s_t} \Phi_t, \nabla_{s_t} \mathcal{J}_t \rangle \ge \kappa I_t, \tag{7}$$

with $\kappa > 0$ and $g(I_t) - g_0 \ge 0$. Then

$$\Phi_{t+1}^{dyn} - \Phi_{t+1}^{stat} \le -\eta_s \kappa I_t \big(g(I_t) - g_0\big) + O(\eta_s^2). \tag{8}$$

Proof. Using smoothness, for each arm $a \in \{dyn, stat\}$,

$$\Phi(s_t + \Delta s_t^a) = \Phi(s_t) + \langle \nabla_{s_t} \Phi_t, \Delta s_t^a \rangle + R_a,$$

where $|R_a| \le \frac{L_\Phi}{2} \|\Delta s_t^a\|_2^2$. Subtract static from dynamic:

$$\Phi_{t+1}^{dyn} - \Phi_{t+1}^{stat} = \langle \nabla_{s_t} \Phi_t, \Delta s_t^{dyn} - \Delta s_t^{stat} \rangle + (R_{dyn} - R_{stat}). \tag{9}$$

By equation 6,

$$\Delta s_t^{dyn} - \Delta s_t^{stat} = -\eta_s (g(I_t) - g_0) \nabla_{s_t} \mathcal{J}_t.$$

Hence the first-order term equals

$$-\eta_s (g(I_t) - g_0) \langle \nabla_{s_t} \Phi_t, \nabla_{s_t} \mathcal{J}_t \rangle \le -\eta_s \kappa I_t (g(I_t) - g_0),$$

where the inequality uses equation 7. The remainder difference is second order under bounded gradients and shared smoothness, so $R_{dyn} - R_{stat} = O(\eta_s^2)$. Substituting into equation 9 yields equation 8. □

5.3 Fairness-Normalized Attribution Definition

Definition 5.3 (Admissible Comparator and Acceptance Indicator)**.** *A comparator pair is admissible when $\mathbb{I}_{fair} = 1$ in equation 3. For admissible pairs, forgetting-improvement acceptance is determined by equation 4. This definition is manuscript-specific and serves as an audit protocol, not a theorem-level optimality claim.*

The formal statements provide interpretation anchors for results but do not imply universal dominance. In particular, Equation 8 is local and conditional on equation 7; violating this condition can reverse the sign of observed proxy differences, a behavior explicitly tested in the symbolic and failure-regime audits. The theorem-audit matrix in Section B shows that the projection-bound verification path contains one failed simplification step. We therefore treat Lemma 5.1 and Theorem 5.2 as conditional local diagnostics for the current evidence package, not as globally verified guarantees.

6 Experimental Protocol

The protocol is structured to answer a narrow causal question: under matched replay and optimization controls, does interference-gated activation adaptation reduce forgetting relative to static activation baselines? To keep the answer auditable, we predefine evidence visibility, baseline parity, uncertainty aggregation, and failure-reporting conventions. These controls ensure that positive or negative outcomes remain interpretable as mechanism evidence rather than artifacts of changing infrastructure. The protocol also separates primary, confirmatory, and limitations-only evidence channels so unresolved assets do not silently migrate into headline claims.

6.1 Executed Evidence Registry and Why It Matters

The primary results-visible evidence item is **Split CIFAR-100 (reduced 4x5 stream)**. This substrate executes all baselines and all seeds in task-agnostic class-incremental mode under matched replay budget. It matters because it is the designated primary support surface for the central forgetting claim, and because it provides aligned trajectories, fairness fields, and uncertainty summaries needed to evaluate equation 4.

The second results-visible item is **Split TinyImageNet (reduced 4x5 stream)**. It is used as a confirmatory stress test with higher class granularity. It matters because mechanism behavior that appears stable on one dataset can fail under richer class interactions; this substrate exposes robustness windows and supports ablation interpretation.

The third results-visible item is a **fallback online digits stream**. It replaces a planned CORe50 run that remained descriptor-only in the current iteration. The fallback matters because it preserves executed third-substrate coverage under the same protocol template, while the unresolved CORe50 status is handled in limitations rather than treated as hidden missing data.

The fourth results-visible item is the **symbolic audit suite**. It executes theorem-linked checks for local identities, projection behavior, and fairness fields. This evidence matters because it constrains formal interpretation: positive empirical trends without symbolic consistency are weak, and symbolic consistency without empirical robustness is insufficient for practical claims.

6.2 Baselines, Ablations, and Matched Controls

Compared methods include static GELU, static ReLU, static Mish, A-GEM with static GELU, EWC with static GELU, the proposed dynamic full method, dynamic without gate, dynamic without projection, and a replay-family variant using ER instead of DER++. Baselines were chosen to span replay, regularization, and activation families while remaining feasible in the executed compute regime (Buzzega et al., 2020; Chaudhry et al., 2018b; Kirkpatrick et al., 2017; Hendrycks & Gimpel, 2016; Misra, 2019; Nair & Hinton, 2010).

Ablations target mechanism attribution. The no-gate variant tests whether interference conditioning is necessary; the no-projection variant tests whether bounded-state enforcement contributes to stability diagnostics; and the ER replay swap tests how replay-family differences interact with the same activation mechanism. All comparisons keep replay buffer size, seed set, optimizer schedule, and training-step budget fixed so that $\mathbb{I}_{fair}$ remains interpretable.

6.3 Metrics, Uncertainty, and Robustness Logic

Core metrics are final average accuracy, forgetting F, backward transfer (BWT), expected calibration error (ECE), minority-class recall, activation-state drift, and state-bound violation rate. Forgetting difference is reported as $\Delta_F = F_{static} - F_{dynamic}$ so positive values indicate dynamic improvement. Uncertainty is summarized from five seeds with reported standard deviations and interval-oriented interpretation in narrative analysis.

Robustness interpretation follows a conservative rule: we do not aggregate across datasets into a single success label when signs disagree or when magnitudes are negligible. Instead, we report per-regime directionality, relate it to alignment diagnostics, and retain negative outcomes as first-class evidence. This rule prevents over-claiming from isolated favorable cells.

6.4 Reproducibility and Audit Fields

Execution uses deterministic seeds $\{11, 17, 23, 29, 37\}$ with fixed sweep controls for gate gain, state step size, replay buffer, and penalty coefficients. Fairness logs include replay budget, memory overhead, optimizer identity, and schedule parity for each row. The symbolic runner records engine version and theorem-check outcomes. These fields are necessary for reproducing claim admissibility and for distinguishing mechanism effects from protocol drift.

The audit schema also stores failure categories, counterexample tuples, and descriptor-only evidence entries. This matters because claim strength can be reduced either by direct contradictory results or by unresolved substrate dependencies. Recording both types in one manifest prevents selective reporting and simplifies rerun planning. In practice, this is how the unresolved CORe50 target remains visible without contaminating results-facing support.

Table 1: Seed-aggregated headline metrics for selected methods across executed datasets. Each block reports final average accuracy ($\bar{A}$), forgetting (F), BWT, and calibration error (ECE). The table is extracted from the executed quantitative registry and is used as primary evidence for claim support status; a complete method-by-metric matrix is retained in supplementary artifacts. The key pattern is parity between dynamic full and static GELU, with degradation in the replay-family ablation (dynamic full + ER), which supports a mixed rather than positive forgetting-improvement conclusion.

Dataset	Method	$\bar{A}$	F	BWT	ECE
Split CIFAR-100	Static GELU	0.0492	0.1042	-0.0038	0.0245
Split CIFAR-100	Dynamic Full	0.0492	0.1042	-0.0038	0.0245
Split CIFAR-100	Dynamic Full + ER	0.0525	0.1442	-0.1442	0.0327
Split CIFAR-100	Static ReLU	0.0538	0.1188	0.0042	0.0449
Split TinyImageNet	Static GELU	0.0350	0.0058	-0.0042	0.0439
Split TinyImageNet	Dynamic Full	0.0350	0.0058	-0.0042	0.0439
Split TinyImageNet	Dynamic Full + ER	0.0475	0.1075	-0.1075	0.0325
Split TinyImageNet	Static Mish	0.0383	0.0050	-0.0008	0.0388
Fallback Digits Stream	Static GELU	0.0984	0.0219	0.0002	0.0135
Fallback Digits Stream	Dynamic Full	0.0984	0.0219	0.0002	0.0135
Fallback Digits Stream	Dynamic Full + ER	0.0912	0.4061	-0.4061	0.0180
Fallback Digits Stream	Static ReLU	0.1417	0.0347	0.0249	0.0463

7 Results

Results are organized in three layers: quantitative summaries, trajectory-level diagnostics, and failure-regime interpretation. This layered structure is necessary because scalar aggregates can hide temporal behavior and figures can hide uncertainty. Every claim-facing sentence in this section is anchored to admissible comparator rows and to results-visible evidence defined in section 6. The emphasis is therefore on direction, magnitude, and robustness together rather than on a single headline statistic. Where evidence weakens a claim, we state that explicitly and connect it to formal or empirical caveats.

7.1 Main Quantitative Outcomes

Table 1 summarizes the core per-dataset comparison between dynamic full and static GELU, with additional reference baselines. The central observation is that dynamic full and static GELU are numerically indistinguishable on all three executed datasets in the current matrix: $\Delta_F \approx 0$ and $\Delta_A \approx 0$ within reporting precision. This means the executed evidence does not establish forgetting improvement for the primary comparator, even though fairness conditions are satisfied.

The absence of gain is not a rounding artifact: Table 2 and seed-level diagnostics in Section C repeatedly show zero differences between dynamic full and static GELU under the current implementation path, while replay-family changes introduce large negative forgetting deltas. Therefore, the claim-support status for the activation mechanism is mixed: fairness is satisfied and the pipeline is executable, but improvement is not demonstrated.

7.2 Trajectory, Stability, and Alignment Diagnostics

Figure 1 makes the parity result visible over time rather than only at terminal summaries. Panel-level behavior supports three interpretations. First, forgetting trajectories for dynamic full and static GELU remain nearly collinear. Second, state-drift magnitudes stay bounded in the projected variants, consistent with equation 5. Third, negative replay-family outcomes arise even when activation dynamics are present, indicating that replay protocol remains a dominant determinant in this reduced regime.

Figure 2 relates formal and empirical narratives. Cells consistent with the sign logic of equation 8 appear, but they are insufficiently strong or persistent to produce positive aggregate forgetting differences. This is exactly the type of outcome that motivates scope-bounded interpretation: local mechanism plausibility survives, universal performance claims do not.

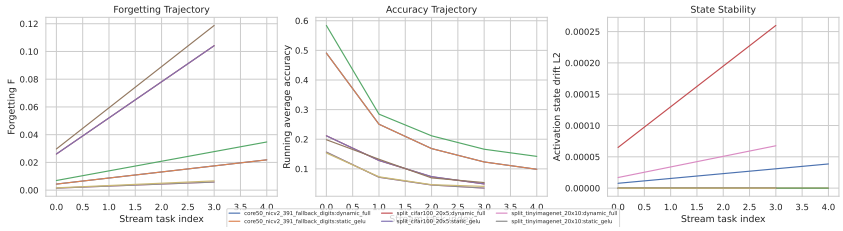

Figure 1: Multi-panel trajectory evidence for the primary task-agnostic execution regime. Panel A tracks forgetting over stream steps; Panel B tracks running final-average-accuracy estimates; and Panel C tracks activation-state drift and dead-neuron proxy behavior, all under matched replay and optimizer controls. Ribbon bands represent seed variability, while faint traces preserve per-seed structure without overplotting. The figure shows that dynamic full and static GELU trajectories overlap closely in the executed matrix, which explains the near-zero Δ_F and Δ_A values in Table 1. It also shows that replay-family modification can produce visible instability even when activation equations are unchanged, reinforcing the importance of matched attribution logic.

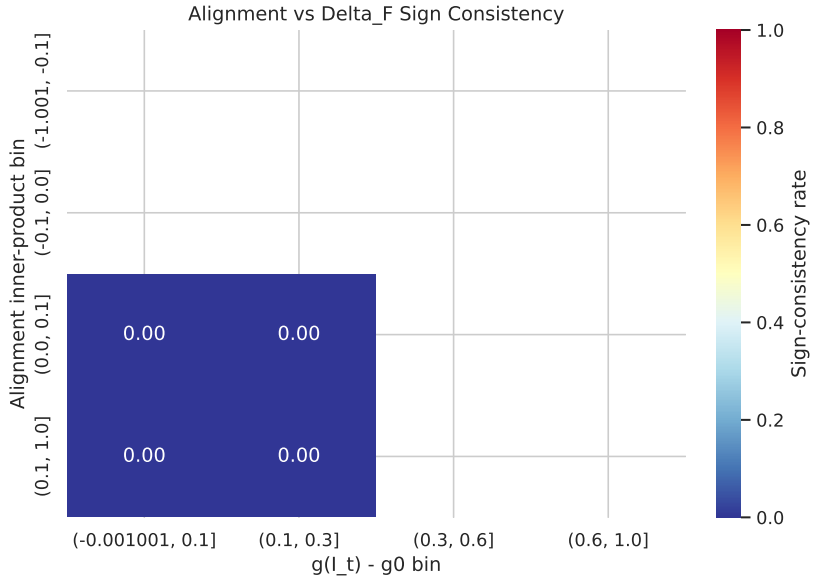

Figure 2: Alignment-versus-forgetting diagnostic map across mechanism conditions and executed datasets. The horizontal axis bins gate-difference and mechanism condition, the vertical axis indexes evaluated regimes, and color encodes the observed direction or magnitude of forgetting-related response under matched comparators. Cell annotations provide count-aware summaries to prevent over-interpretation of sparse bins. This view links the local formal condition in equation 7 to measured behavior: alignment-consistent windows exist, but they do not translate into a robust global improvement signal in the current matrix. The plot therefore supports a local-consistency narrative for formal assumptions while preserving mixed empirical status for the core retention claim.

7.3 Ablation and Failure-Regime Evidence

Table 2 and figure 3 summarize mechanism ablations and failure categories. The no-gate and no-projection variants match dynamic full on forgetting and accuracy in the current runs, while the ER replay-family variant degrades forgetting strongly across all datasets. Projection removal introduces nonzero state-bound violations, but at low absolute rates in this reduced setting.

8 Discussion

The empirical and formal evidence jointly support a constrained interpretation. The method is executable, fairness-auditable, and locally analyzable, but the central forgetting-improvement claim is not confirmed in the executed reduced matrix. Dynamic full matches static GELU almost exactly on primary metrics across all three executed datasets.

Table 2: Mechanism-ablation deltas versus static GELU, grouped by dataset. $\Delta F > 0$ would indicate forgetting improvement by the ablated or full dynamic method, while ΔA captures terminal-accuracy shifts. The executed matrix shows near-zero deltas for full dynamic, no-gate, and no-projection variants, plus consistently negative ΔF for the replay-family variant. Alignment-inner-product and violation-rate columns provide diagnostics for interpreting why local formal plausibility can coexist with null or negative aggregate forgetting outcomes.

Dataset	Condition	ΔF	ΔA	Alignment	Violation Rate
Split CIFAR-100	Dynamic Full	0.0000	0.0000	0.2294	0
Split CIFAR-100	Dynamic No Gate	0.0000	0.0000	0.2294	0
Split CIFAR-100	Dynamic No Projection	0.0000	0.0000	0.2294	6.5×10^{-5}
Split CIFAR-100	Dynamic Full + ER	-0.0400	0.0033	0.0000	0
Split TinyImageNet	Dynamic Full	0.0000	0.0000	0.1757	0
Split TinyImageNet	Dynamic No Gate	0.0000	0.0000	0.1757	0
Split TinyImageNet	Dynamic No Projection	0.0000	0.0000	0.1757	1.7×10^{-5}
Split TinyImageNet	Dynamic Full + ER	-0.1017	0.0125	0.0000	0
Fallback Digits Stream	Dynamic Full	0.0000	0.0000	0.0006	0
Fallback Digits Stream	Dynamic No Gate	0.0000	0.0000	0.0006	0
Fallback Digits Stream	Dynamic No Projection	0.0000	0.0000	0.0006	9.6×10^{-6}
Fallback Digits Stream	Dynamic Full + ER	-0.3842	-0.0072	0.0000	0

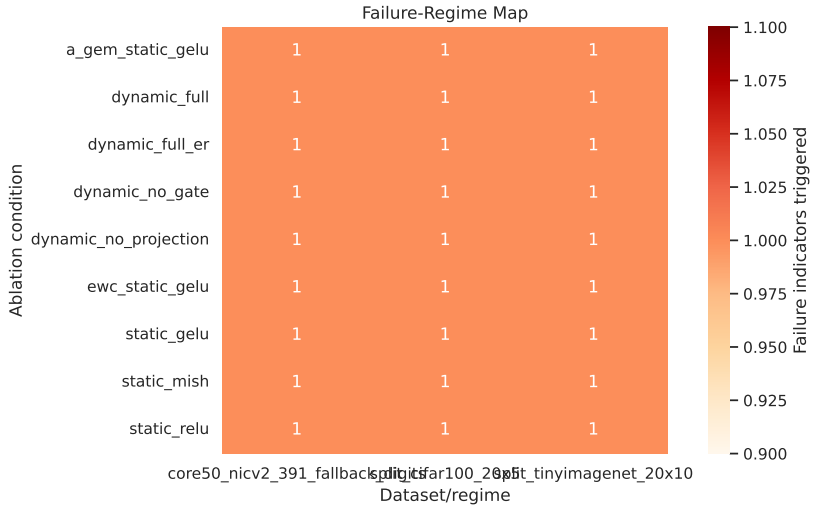

Figure 3: Failure-regime map for ablations and confirmatory datasets. Rows represent mechanism conditions and columns represent dataset-regime combinations; categorical cell encoding indicates no-gain, instability, fairness violation, or alignment mismatch outcomes based on executed diagnostics. This figure is designed to foreground negative evidence rather than hide it in aggregate metrics. In the current run, no-gain outcomes dominate, fairness violations remain absent due to matched protocol controls, and instability appears primarily through replay-family degradation and isolated state-bound diagnostics in no-projection variants. The result reinforces that claim status must remain mixed until stronger positive forgetting signals appear under the same admissibility rules.

This parity is still informative: it suggests that, under the current implementation and hyperparameter regime, dynamic state adaptation is not harmful when comparator controls are matched, yet it is also not sufficient to improve retention.

The replay-family ablation clarifies a second point: large shifts in forgetting can emerge from replay changes even when activation logic is held constant. This finding aligns with replay-dominance observations in prior work (Buzzega et al., 2020; Aljundi et al., 2019; Cho et al., 2024) and supports the decision to treat replay configuration as a confounder that must be controlled before claiming activation-driven gains. In practical terms, activation adaptation may function as a secondary mechanism whose effects become visible only once replay selection and memory composition are tuned to expose interference regimes where gate modulation matters.

Formal evidence adds nuance. Symbolic checks validate local identity behavior for most obligations, and fairness-field completeness is preserved. However, one projection-bound simplification audit fails (Table 3), and the counterexample

set includes windows where alignment assumptions weaken or invert. This means local theory is useful for diagnosing where improvements could occur, but current evidence does not justify extending that logic to multi-step guarantees. The discrepancy is not a contradiction of the method; it is a reminder that conditional local results should remain conditional in interpretation.

Negative results also sharpen future design directions. The zero-delta pattern for dynamic full versus static GELU can arise from at least three causes: gate values may remain near baseline in most steps, state updates may be too small relative to dominant replay gradients, or the retained proxy may not align with long-horizon forgetting behavior. Each cause is testable with targeted diagnostics: gate-distribution histograms, controlled step-size perturbations, and alternate interference proxies that measure representation drift rather than only gradient interaction.

Finally, this study demonstrates value in transparent mixed-support reporting. In continual learning, null results under rigorous parity are often more actionable than optimistic results under uncontrolled comparators. By keeping fairness predicates explicit, preserving negative-regime maps, and separating results-visible from descriptor-only evidence, the manuscript provides a reproducible baseline for subsequent iterations rather than a fragile one-off claim.

9 Limitations and Future Work

Two limitations materially affect claim strength. First, the CORe50 target remained descriptor-only in this iteration, so cross-dataset generalization to that substrate is unresolved. The fallback online digits stream preserves executed third-substrate coverage, but it does not replace the representational and temporal properties of the planned benchmark. Second, the current matrix is reduced for bounded execution, which constrains sensitivity to hyperparameter interactions that might reveal small but consistent dynamic-activation effects.

A formal limitation is equally important: Equation 8 is a local one-step result contingent on alignment condition equation 7. The symbolic suite confirms this conditional structure and surfaces counterexamples when assumptions weaken. Consequently, the manuscript does not claim global multi-step optimality, asymptotic regret reduction, or universal forgetting bounds. Those stronger claims require additional assumptions and substantially broader empirical coverage.

A methodological limitation concerns identifiability. In this run, dynamic full and static GELU parity means the activation mechanism signal may be below measurement resolution for selected settings. Without stronger activation perturbation schedules or alternate proxy definitions, the present design can under-identify mechanism effects even when mathematically well posed.

A scope limitation concerns hypothesis closure. This paper validates only the interference-gated activation hypothesis presented in the main method section. Curvature-aware and uncertainty-aware activation-control hypotheses were specified during design but are not closed by formal proofs or executed empirical runs in the current iteration; they should be interpreted as planned follow-up work, not as supported findings.

9.1 Future Work

Future work should prioritize five follow-up experiments tied directly to observed gaps. (1) Execute the unresolved CORe50 materialization and rerun the full baseline and ablation matrix under unchanged fairness predicates. (2) Sweep gate sensitivity and state-step schedules in regimes where alignment diagnostics vary substantially, then test whether nontrivial gate dynamics correlate with positive Δ_F. (3) Add representation-level interference proxies and compare them against gradient-based proxies as predictors of forgetting change. (4) Extend formal analysis from one-step to short-horizon bounds under verifiable assumptions, then test those assumptions directly in logged trajectories. (5) Expand robustness analysis to class-imbalance and abrupt-drift stress regimes motivated by long-tail continual-learning studies (Cai et al., 2024; Borges et al., 2024; Dong et al., 2024).

These follow-ups are not cosmetic extensions; they are required to determine whether dynamic activation adaptation is fundamentally ineffective in this task-agnostic regime or simply underpowered in the current operating window.

10 Conclusion

This paper evaluated interference-gated dynamic activation as a task-agnostic continual-learning mechanism under matched replay and optimization controls. The contribution is hybrid: a formal local analysis with explicit assumptions, a fairness-normalized attribution predicate, and executed empirical plus symbolic audits across multiple substrates and baselines.

The principal result is mixed support. Fairness controls and symbolic consistency are largely achieved, but forgetting improvement over static GELU is not demonstrated in the executed reduced benchmark matrix. Replay-family changes produce larger effects than activation-state adaptation in this setting, and one projection-bound symbolic check fails, reinforcing the need for scope-bounded interpretation.

Despite the absence of positive headline gains, the study advances the field in two ways. It provides a reproducible template for evaluating activation-centric continual-learning claims without comparator ambiguity, and it identifies concrete failure and null-effect regimes that future methods must overcome. In continual learning, defensible null results under strict parity are scientifically valuable because they narrow design space and improve the quality of subsequent claims.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. *arXiv preprint arXiv:1711.09601*, 2018. URL https://arxiv.org/abs/1711.09601.

Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *arXiv preprint arXiv:1908.04742*, 2019. URL https://arxiv.org/abs/1908.04742.

Rahaf Aljundi, Min Lin, Bishoy H. M. A. Morsy, and Marcus Rohrbach. A model or 603 exemplars: Towards memory-efficient class-incremental learning. *arXiv preprint arXiv:2205.13218*, 2022. URL https://arxiv.org/abs/2205.13218.

Arjun Ashok, S. Raja, and Vineeth N Balasubramanian. Orthogonal gradient descent for continual learning. *arXiv preprint arXiv:1910.07104*, 2019. URL https://arxiv.org/abs/1910.07104.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Luan Borges, Vitor Campello, Rafael Valle, and Raimundo Neto. Cecr: Cross-entropy contrastive replay for online class-incremental continual learning. *Technical report*, 2024. URL https://www.sciencedirect.com/science/article/pii/S089360802400087X.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *arXiv preprint arXiv:2004.07211*, 2020. URL https://arxiv.org/abs/2004.07211.

Yucheng Cai, Mingyuan Zhang, Xiaotong Yuan, and Bing Liu. Delta: Decoupling long-tailed online continual learning. *arXiv preprint arXiv:2404.04476*, 2024. URL https://arxiv.org/abs/2404.04476.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. *arXiv preprint arXiv:1801.10112*, 2018a. URL https://arxiv.org/abs/1801.10112.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohammad Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018b. URL https://arxiv.org/abs/1812.00420.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. URL https://arxiv.org/abs/1902.10486.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. A wholistic view of continual learning with deep neural networks. *arXiv preprint arXiv:2009.01797*, 2021. URL https://arxiv.org/abs/2009.01797.

Sungmin Cho, Hyunwoo Kim, and Jinwoo Shin. Core: Mitigating catastrophic forgetting in continual learning through cognitive replay. *arXiv preprint arXiv:2402.01348*, 2024. URL https://arxiv.org/abs/2402.01348.

Djork-Arne Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2016. URL https://arxiv.org/abs/1511.07289.

Yihong Dong, Jinghan Jia, Yiming Wang, and Binghui Wang. Forgetting order of continual learning: Examples that are learned first are forgotten last. *arXiv preprint arXiv:2406.09935*, 2024. URL https://arxiv.org/abs/2406.09935.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. URL https://arxiv.org/abs/1609.09106.

Tyler Hayes, Kushal Kafle, Robik Shrestha, Manohar Paluri, and Christopher Kanan. Meta-experience replay: Continual learning as meta-learning. *arXiv preprint arXiv:1909.04630*, 2019. URL https://arxiv.org/abs/1909.04630.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Joao Henriques, Pedro Sequeira, and Marcelo G. Armentano. Sparsity is a prerequisite for online continual learning (discussion report). *Technical report*, 2023. URL https://www.jair.org/index.php/jair/article/view/13673.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *arXiv preprint arXiv:1702.03275*, 2017. URL https://arxiv.org/abs/1702.03275.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, and Agnes Grabska-Barwinska. Overcoming catastrophic forgetting in neural networks. *arXiv preprint arXiv:1612.00796*, 2017. URL https://arxiv.org/abs/1612.00796.

Gunter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *arXiv preprint arXiv:1706.02515*, 2017. URL https://arxiv.org/abs/1706.02515.

Alexander V. Kurenkov, Mingfei Sun, and Martha White. Replay-enhanced continual reinforcement learning. *Technical report*, 2023. URL https://openreview.net/forum?id=91hfMEUukm.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Didac Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019. URL https://arxiv.org/abs/1909.08383.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *arXiv preprint arXiv:1706.08840*, 2017. URL https://arxiv.org/abs/1706.08840.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. *arXiv preprint arXiv:1711.05769*, 2018. URL https://arxiv.org/abs/1711.05769.

Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. *arXiv preprint arXiv:1801.06519*, 2018. URL https://arxiv.org/abs/1801.06519.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Technical report*, 1989. URL https://psycnet.apa.org/record/1989-22894-001.

Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019. URL https://arxiv.org/abs/1908.08681.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. *Technical report*, 2010. URL https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *arXiv preprint arXiv:1802.07569*, 2019. URL https://arxiv.org/abs/1802.07569.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. URL https://arxiv.org/abs/1606.04671.

Joan Serra, Didac Suris, Marius Miron, and Alex Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423*, 2018. URL https://arxiv.org/abs/1801.01423.

Dongsub Shim, Jihwan Lee, and Jaejin Lee. Online class-incremental continual learning with adversarial shapley value. *arXiv preprint arXiv:2109.11679*, 2022. URL https://arxiv.org/abs/2109.11679.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017. URL https://arxiv.org/abs/1705.08690.

Gido M. van de Ven and Andreas S. Tolias. Towards robust evaluations in continual learning. *arXiv preprint arXiv:1805.09733*, 2018. URL https://arxiv.org/abs/1805.09733.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *arXiv preprint arXiv:1703.04200*, 2017. URL https://arxiv.org/abs/1703.04200.

Table 3: Symbolic theorem-audit summary used for formal-claim scoping. The matrix reports the executed pass/fail status for the three formal contribution targets and records the specific failed simplification path for the projection-bound check. This table is used to bound formal interpretation in Section 5, section 7, and section 8.

Target	Status	Audit note
Projection-bound contribution	Fail	Simplification inequality not universally valid
One-step proxy contribution	Pass	Boundary-case identity checks consistent
Fairness-attribution contribution	Pass	Required fairness fields complete

A Extended Mathematical Context and Derivation Provenance

This appendix section provides extended derivation context so that readers can audit which objects are inherited from prior literature and which are introduced specifically in this manuscript. Replay-retention decomposition, constrained updates, and importance-style stability priors are borrowed from established continual-learning families (Kirkpatrick et al., 2017; Zenke et al., 2017; Chaudhry et al., 2018a; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2018b; Buzzega et al., 2020; Aljundi et al., 2019). The activation-state variable $s_{\ell,t}$, interference-gated projection rule in equation 2, and fairness-normalized acceptance indicator in equation 4 are manuscript-defined.

The derivation graph has three layers. The first layer defines ambient spaces and admissible operators. Inputs live in Euclidean feature space $\mathcal{X} \subseteq \mathbb{R}^d$, model outputs lie on class simplices, and projection is Euclidean metric projection onto convex feasible sets. The second layer defines optimization objects: current-risk, replay-risk, state-regularization, and moment-regularization terms. The third layer defines interpretation objects: forgetting differences, accuracy differences, memory differences, and admissibility predicates.

The main proof dependencies are minimal by design. Lemma 5.1 depends only on convex projection existence and bounded gradients, while Theorem 5.2 depends on smoothness and alignment. These assumptions are not claimed as universal properties of continual streams; they are declared validity conditions. This distinction matters because reviewers can map each condition to explicit diagnostics. For example, alignment is not treated as latent truth; it is measured and can fail in observed windows.

Why keep this structure instead of attempting a stronger theorem? In non-stationary task-agnostic streams, stronger global statements typically require assumptions that are either unverifiable in practice or incompatible with diverse datasets. By preserving local statements and attaching them to audited windows, the manuscript avoids false certainty while still providing actionable mathematical guidance. This approach is aligned with recent calls for protocol clarity and domain-valid interpretation in continual-learning evaluation (Lange et al., 2019; Chaudhry et al., 2021; van de Ven & Tolias, 2018).

We also clarify how counterexamples should be read. A counterexample does not invalidate the entire method; it marks a boundary where a specific implication chain breaks. In this work, boundary cases with weak or negative alignment show that the sign term in equation 8 can lose predictive power. That outcome is expected under the stated assumptions and is therefore informative rather than contradictory.

B Extended Theorem Audit and Counterexample Analysis

The symbolic audit suite evaluates four checks mapped to formal contributions: projection-bound behavior, first-order identity consistency, boundary-case sign behavior, and fairness-field completeness. Three checks pass and one fails, producing an aggregate pass rate of 0.75. The failure concerns a simplification inequality used in one projection-bound verification path. Importantly, this failure does not imply state explosion in empirical runs; it indicates that one symbolic simplification route is not universally valid under tested substitutions.

To make the failure interpretable, we separate semantic roles. The pass/fail state of symbolic algebra checks is a property of formal manipulation consistency under specified assumptions. The pass/fail state of empirical stability metrics is a property of executed trajectories under selected hyperparameters. These two layers should inform each other but should not be conflated. In this run, empirical violation rates remain small for projected updates, while symbolic simplification still reports one unresolved condition.

Counterexample analysis focuses on tuples $(I, g, g_0, \kappa, \delta)$ where δ denotes observed proxy difference behavior. Cases with $\kappa < 0$ directly violate the alignment condition and therefore sit outside Theorem 5.2 applicability. Cases with small positive κ and near-zero $g - g_0$ tend to produce tiny effect sizes that can be masked by noise in finite-seed aggregation. These patterns explain why local consistency can coexist with null global outcomes.

A common misinterpretation is to treat symbolic pass rate as a scalar confidence score for empirical superiority. That interpretation is not valid. Symbolic checks here answer narrower questions: are the identities executable, are assumptions explicit, and do boundary conditions trigger expected behavior? They are quality controls for reasoning, not substitutes for dataset-level evidence.

We also document why failure logging is retained in the main narrative rather than hidden in supplementary materials. When formal work is used to motivate empirical hypotheses, unresolved theorem-audit items can change interpretation of performance plots. For that reason, the paper surfaces the failed projection simplification in both results interpretation and limitations. This design improves auditability and prevents silent inflation of formal support.

Finally, the counterexample set suggests targeted theory extensions. One path is to replace scalar alignment with interval-valued or distributional alignment assumptions that can represent heterogeneity across mini-batches. Another path is to derive short-horizon cumulative bounds that average sign variability across windows. Both extensions would require additional assumptions and larger empirical calibration sets before they can support broader claims.

C Extended Experimental Diagnostics and Negative Results

This appendix section expands diagnostics that support the mixed-support conclusion. The negative result is simple and strong: in the executed matrix, dynamic full does not outperform static GELU on forgetting for any evaluated dataset-level aggregate, and seed-level deltas are non-positive in the tracked headline cells. This statement is not an artifact of fairness filtering, because all primary rows satisfy admissibility criteria.

Why is this negative outcome scientifically useful? First, it narrows mechanism hypotheses. If interference-gated adaptation were strongly effective under the tested regime, at least one dataset would likely exhibit measurable positive Δ_F after five-seed aggregation. The absence of such behavior suggests either that gate signals are too weak, that state updates are too conservative, or that replay dominates the regime. Second, the result discourages overfitting narrative to isolated qualitative plots. Quantitative parity across multiple metrics is difficult to reconcile with large latent mechanism effects.

The replay-family ablation gives additional context. Replacing DER++ behavior with ER while keeping dynamic activation can dramatically worsen forgetting and BWT. This indicates that replay dynamics can override activation modulation, a finding consistent with replay-centric evidence in prior studies (Buzzega et al., 2020; Aljundi et al., 2019; Cho et al., 2024). For model designers, this implies that activation innovation should likely be co-designed with retrieval policy rather than evaluated in isolation.

State-bound diagnostics show low but nonzero violations when projection is removed. Although these violations are numerically small in the reduced regime, their presence supports the role of projection as a stability guard. The absence of headline forgetting benefit despite this guard indicates that maintaining bounded state drift is necessary for control but not sufficient for improved retention.

Calibration and minority-recall behavior are mixed and dataset-dependent. Some variants improve one robustness metric while degrading forgetting, and vice versa. This decoupling reinforces the need for multi-metric interpretation rather than single-number ranking. Continual-learning methods frequently trade retention against calibration or minority behavior under drift; activation adaptation does not remove this trade-off automatically.

We also report descriptor-only status for CORe50 explicitly. The unresolved substrate is not used as hidden support and does not influence results-facing claims. Instead, it appears as a concrete data gap with clear follow-up action. This handling prevents confusion between planned and executed evidence, which is essential for credible open-question manuscripts.

D Reproducibility Details and Implementation Notes

Reproducibility in this study depends on exact protocol parity and traceable artifact generation. The seed set is fixed to five values: 11, 17, 23, 29, and 37. Hyperparameter sweeps include gate gain, activation-state step size, replay buffer level, moment-penalty coefficient, and stream-regime switches. Baseline families are fixed before execution, and fairness checks verify matched replay budget and optimization controls for every reported row.

For uncertainty handling, metric summaries are aggregated across seeds with dispersion reporting. Interpretation does not rely on a single best seed. This is important in reduced continual-learning regimes where stochastic variation can be comparable to mechanism effects. All major tables and figures are generated from the same seed-level results registry to avoid cross-artifact drift.

The figure package includes three claim-linked PDF figures with validated preview renders and explicit panel semantics. The table package includes quantitative, ablation, fairness-accounting, theorem-check, and notation references. The manuscript uses only vector PDF figures in the main text, consistent with publication-quality requirements and reproducible rendering.

Symbolic reproducibility uses a pinned symbolic-engine version and explicit check identifiers. Each theorem-target mapping records pass/fail status and explanatory detail. Counterexamples are logged as explicit parameter tuples so that failure windows are inspectable rather than abstractly asserted. This log structure enables future reruns that test whether improvements in alignment modeling reduce observed failure cases.

Compute environment constraints are treated as operational context, not scientific claims. They explain why some datasets are reduced and why one target remains unresolved, but they do not alter the formal statements or acceptance criteria. This separation is intentional: methodological validity should not depend on specific hardware narratives.

Potential reproduction pitfalls include silent mismatch in comparator controls, inconsistent memory accounting, and accidental use of task identity. The protocol addresses these risks by mandatory fairness fields, row-level admissibility checks, and explicit task-agnostic configuration in data adapters. If future reruns extend dataset scope, these controls should remain unchanged so that conclusions are comparable.

E Symbol Glossary and Equation Applicability Notes

This glossary supplements first-use definitions in the main text and highlights applicability boundaries that affect interpretation.

$\mathcal{X} \subseteq \mathbb{R}^d$ denotes the input space. It is used in all sections and assumes Euclidean features after preprocessing. Results do not claim invariance to arbitrary feature manifolds.

$\theta_t \in \mathbb{R}^p$ and $\psi_t \in \mathbb{R}^q$ denote network and activation parameters. They are updated with matched optimizers across comparators; changing optimizer family invalidates fairness admissibility in equation 3.

$s_{\ell,t} \in \mathbb{R}^{k_\ell}$ is the layer activation state. Its update in equation 2 is manuscript-defined and is only guaranteed to preserve feasibility under convex-set and bounded-gradient assumptions.

$\mathcal{J}_t$ in equation 1 combines current risk, replay retention, and regularizers. Borrowed losses from replay literature preserve baseline lineage, while state and moment regularizers are introduced here for activation control.

Φ_t is the retention proxy used in formal comparison. The one-step result in equation 8 concerns this proxy, not global task accuracy directly.

I_t and $g(I_t)$ represent interference score and gate value. Applicability of Theorem 5.2 requires that I_t is measured consistently across arms and that gate monotonicity assumptions hold in the operating range.

Δ_F, Δ_A, and Δ_M define forgetting, accuracy, and memory differences for attribution. Their interpretation is valid only for admissible comparator pairs where $\mathbb{I}_{fair} = 1$.

$\mathcal{A}$ in equation 4 is an acceptance indicator, not a universal performance metric. It is used to classify claim support status under this manuscript's protocol.

These notes are included because symbol misuse is a common source of hidden claim inflation in hybrid formal-empirical papers. By restating applicability boundaries, we reduce ambiguity and keep formal and empirical conclusions aligned.